

# Central Limit Theorems for Transition Probabilities of Controlled Markov Chains

**Ziwei Su**

ZIWEI.SU@NORTHWESTERN.EDU

DEPARTMENT OF INDUSTRIAL ENGINEERING AND MANAGEMENT SCIENCES  
NORTHWESTERN UNIVERSITY  
EVANSTON, IL 60208, USA

**Imon Banerjee**

IMON.BANERJEE@NORTHWESTERN.EDU

DEPARTMENT OF INDUSTRIAL ENGINEERING AND MANAGEMENT SCIENCES  
NORTHWESTERN UNIVERSITY  
EVANSTON, IL 60208, USA

**Diego Klabjan**

D-KLABJAN@NORTHWESTERN.EDU

DEPARTMENT OF INDUSTRIAL ENGINEERING AND MANAGEMENT SCIENCES  
NORTHWESTERN UNIVERSITY  
EVANSTON, IL 60208, USA

## Abstract

We develop a central limit theorem (CLT) for a non-parametric estimator of the transition matrices in controlled Markov chains (CMCs) with finite state-action spaces. Our results establish precise conditions on the logging policy under which the estimator is asymptotically normal, and reveal settings in which no CLT can exist. We then build on it to derive CLTs for the value, Q-, and advantage functions of any stationary stochastic policy, including the optimal policy recovered from the estimated model. Goodness-of-fit tests are derived as a corollary, which enable to test whether the logged data is stochastic. These results provide new statistical tools for offline policy evaluation and optimal policy recovery, and enable hypothesis tests for transition probabilities.

**Keywords:** Markov chains, controlled Markov chains, mixing processes, non-parametric estimation, CLT

## 1 Introduction

A discrete-time stochastic process  $\{(X_i, a_i)\}_{i \geq 0}$  is called a **controlled Markov chain** (Borkar, 1991) if the next “state”  $X_{i+1}$  depends only on the current state  $X_i$  and the current (not necessarily Markovian) “control”  $a_i$ , where actions  $a_i$  depend only on the information available up to time  $i$ . Conditioned on  $a_i$ , the state sequence  $\{X_i\}_{i=0}^n$  follows a Markov transition kernel (see Definition 1 for the formal definition). We study the asymptotic distribution of the transition probabilities of finite state-action non-Markovian controlled Markov chains.

In general, controlled Markov chains with Markovian controls (where  $a_i$  depends only on  $X_i$ ) can be used to model time-homogeneous data (like i.i.d., Markovian), time-inhomogeneous data (like i.n.i.d., time-inhomogeneous Markovian (Dolgopyat and Sarig, 2023; Merlevède et al., 2022), Markov decision process (Hernández-Lerma et al., 1991)), etc.

Such models have wide applicability in reinforcement learning (Sutton and Barto, 2018), system stabilization (Yu et al., 2023), or system identification (Ljung, 1999; Mania et al., 2020).

Beyond that, certain categories of adversarial Markov games (Wang et al., 2023), reward machines (Icarte et al., 2018), and minimum entropy explorations (Mutti et al., 2022) are known to induce Markovian state transitions with non-Markovian controls. Therefore, there is an acute need to sharply characterize the transition dynamics of controlled Markovian systems in the presence of non-Markovian controls.

A particularly relevant downstream task in modern machine learning is offline (batch) reinforcement learning (RL). Offline RL often begins with a fixed dataset of trajectories collected by some unknown logging policy (behavior policy), with no further control over data collection. A fundamental challenge in this setting is to accurately estimate the Markov transition dynamics of the environment from these logged data. The estimated transition probabilities are then used to evaluate policies or to find an optimal policy for the given dynamics (see Section 4 for more details).

There have been some recent developments towards understanding statistical properties of the marginal distributions of time-inhomogeneous Markov chains (see Dolgopyat and Sarig (2023) for details). However, despite the widespread prevalence of using the count-based estimator for transition probabilities in model-based RL (Mannor and Tsitsiklis, 2005; Li et al., 2022; Zhu et al., 2024), the statistical properties of this estimator are poorly understood. When the state space and the action space are finite, the simplest and most natural approach is to use the non-parametric count-based empirical estimator (model) of the transition probabilities (see Equation (2.3)). We show that this estimator is essentially the maximum likelihood estimate of the transition probabilities (see Proposition 1).

One recent work (Banerjee et al., 2025a) establishes probably approximately correct (PAC) (Valiant, 1984) guarantees on the estimation error. Their analysis underpins the theoretical nature of PAC learning which does not always align with practice. In particular, it leaves open the important question of the exact asymptotic behavior of the estimators. This gap in theory limits our ability to evaluate estimation error, construct confidence intervals, or perform rigorous hypothesis tests for transition probabilities.

This paper addresses this gap by developing the first asymptotic normality theory for count-based transition estimators in controlled Markov chains with finite states and actions. We answer the following question:

*Under what conditions does there exist a scaling such that the scaled estimation error of the empirical transition estimates obeys a CLT, and when does an asymptotic distribution fail to exist?*

We show that the answer depends crucially on properties of the logging policy (such as sufficient coverage of all state-action pairs). In fact, under some mild recurrence and mixing conditions, we prove that the empirical transition probabilities converge in distribution to a Gaussian under self-normalized scaling equaling the square root of the visitation count of each state-action pair. Our analysis reveals that the estimated transition probabilities under this scaling have a multinomial behavior in the limit and are asymptotically independent across state-action pairs (see Theorem 1).

Our results generalize classical CLTs for transition probabilities of ergodic Markov chains (Billingsley, 1961) to non-ergodic, non-stationary stochastic processes, being a first of its kind. Furthermore, we also identify scenarios in which a CLT cannot be established for any estimator of the transition probabilities.

To demonstrate the applicability of our results in downstream tasks such as offline policy evaluation (OPE) and optimal policy recovery (OPR), we then derive asymptotic guarantees for the value ( $V$ ), action-value ( $Q$ ), or advantage ( $A$ ) functions for any stationary target policy by solving the Bellman equation with the plug-in transition estimator. In particular, by analyzing the value of the optimal policy calculated from the estimated transition probabilities, we establish a CLT for the value function of the optimal policy under the true transition kernel.

Our results enable, for example, confidence intervals on the value of the optimal policy and high-probability guarantees for recovering the true optimal policy, providing a principled way for OPE and OPR using asymptotic statistics.

**Contributions.** The paper develops an asymptotic theory for non-parametric estimation of transition probabilities in finite controlled Markov chains (CMCs) under general, possibly non-stationary and history-dependent logging policies. Our main contributions are as follows.

1. **A central limit theorem for count-based transition estimators in CMCs.** We establish the first CLT for the classical empirical estimator of controlled transition kernels (Theorem 1). The result holds under mild assumptions on the return times of state–action pairs and weak  $\eta$ -mixing of the joint state–action process. The covariance structure is multinomial and asymptotically independent across state–action pairs. Based on the CLT, we derive a chi-square goodness-of-fit test for transition probabilities (Proposition 4).
2. **A characterization of when a CLT cannot hold.** We show that if the logging policy fails to visit certain state–action pairs infinitely often asymptotically, then no estimator can satisfy any CLT under any diverging scaling (Proposition 3). This result draws clear regions of testable and untestable dynamics in offline RL data.
3. **CLTs for value, Q-, and advantage functions under arbitrary stationary target policies.** Using the transition-kernel CLT, we obtain joint asymptotic normality for estimates of value, Q-, and advantage functions (Theorem 2), as well as the value of the optimal policy computed from the estimated model (Theorem 3). These results provide a unified asymptotic framework for offline policy evaluation and optimal policy recovery.

Beyond high-level contributions outlined above, the paper introduces two core technical innovations that may be of independent interest.

1. **Relaxation of uniform bounded return times to sublinear growth.** Prior work assumes finite expected return times for every state–action pair. This includes both regularity assumptions for Markov chains (Meyn and Tweedie, 2012, Chapter 11) and its controlled counterpart (Banerjee et al., 2025a). In contrast, our main theorem allows return times to grow sublinearly with the number of returns. This is enabled

by Proposition 2, which gives a novel martingale-based lower bound on the expected number of visits to each state-control pair, even when expected return times grow.

2. **A policy-dependent recurrence classification for CMCs.** Classical recurrence classification in Markov chains does not extend to controlled processes. We identify the CMC analogue:

- a positive-recurrent-like regime guaranteeing CLTs (Proposition 2),
- a transient-like regime where no CLT is possible (Proposition 3),
- and an intermediate “null-recurrent” regime that remains theoretically delicate.

This provides, to the best of our knowledge, the first testability boundary for asymptotic inference in non-stationary CMCs.

**Outline.** The remainder of the paper is organized as follows. Section 1.1 concludes the introduction with a review of the literature. Section 2 introduces the notions of hitting and return times, mixing, and ergodic occupation measure, and formally introduces our assumptions. Section 3 introduces the CLT for transition probabilities, a goodness-of-fit test for transition probabilities, and scenarios where a CLT does not hold. Section 4 applies our main results to derive CLTs for the value,  $Q$ -, advantage functions and the optimal policy value.

### 1.1 Literature Review

We divide the literature review into two parts. In the first part, we place our work in the context of the existing literature on non-parametric estimation for stochastic processes. In the second part, we place our work in the context of reinforcement learning.

**Non-parametric Estimation.** Non-parametric estimation and limit theory for Markov chains are well documented in the literature. Billingsley (1961) develops empirical transition estimators and asymptotic guarantees for finite ergodic chains, while Yakowitz (1979) extends these ideas to infinite or continuous state spaces. For time-homogeneous chains, Geyer (1998) proves laws of large numbers (LLNs) and Jones (2004) establishes corresponding CLTs under mixing conditions. More recent work shifts attention to minimax sample complexity: Wolfer and Kontorovich (2019, 2021) derives optimal rates for ergodic chains, and Banerjee et al. (2025a) does so for discrete-time, finite state-action controlled Markov chains. Beyond transition-kernel estimation, classical results by Dobrushin (1956a,b); Rosenblatt-Roth (1963, 1964) provide LLNs and CLTs for normalized sample means in time-inhomogeneous settings, but leaves open the question of estimating the transition matrix itself.

Despite this progress, statistical inference for time-inhomogeneous controlled Markov chains—particularly when controls are stochastic and the state–action process is non-Markovian—remains unexplored. The present paper fills this gap by establishing the first CLT for the count-based, non-parametric estimator of the transition kernel in such controlled settings, together with its implications for model-based offline reinforcement learning.

**Reinforcement Learning.** In model-based offline (batch) reinforcement learning (Levine et al., 2020; Kidambi et al., 2020; Yu et al., 2020), it is important to learn or evaluate policies

solely from logged trajectories without additional exploration. Applications range from clinical decision making—where logged physician actions serve as controls (Shortreed et al., 2011; Liu et al., 2020; Yu et al., 2021; Chen et al., 2021)—to service-operations settings such as patient flow and inventory routing (Armony et al., 2015).

Parallel works on OPE include importance sampling (Precup et al., 2000), bootstrap (Hanna et al., 2017; Hao et al., 2021), doubly-robust (Jiang and Li, 2016; Thomas and Brunskill, 2016) and concentration-inequality-based (Thomas et al., 2015) estimators that provide high-confidence bounds on a target policy’s value, while research on OPR (Antos et al., 2008; Munos and Szepesvári, 2008; Chen and Jiang, 2019; Zanette, 2021; Shi et al., 2022) examines when a policy optimized on the learned model is near-optimal. Most existing guarantees are finite-sample or PAC in nature and assume either sufficient state–action coverage or Markovian behavior policies. From a minimax sample-complexity perspective, Li et al. (2022); Yan et al. (2022) establish sharp optimal rates for discounted and finite-horizon settings with stationary logging policies, and Banerjee et al. (2025a) extends this analysis to discrete-time, finite state–action controlled Markov chains with stochastic, history-dependent logging.

Recent work by Zhu et al. (2024) studies statistical uncertainty quantification in reinforcement learning and establishes CLTs for value and Q-functions of the optimal policy. In particular, their Theorem 1 builds on classical Markov chain CLTs (Trevezas and Linnios, 2009) by viewing each state–action pair as a single composite state and thus assuming stationary, irreducible logging data. While they also employ count-based estimators for transitions and empirical estimators for rewards, the asymptotic normality of these estimators is taken as an assumption rather than derived, and their framework does not accommodate non-stationary logging policies. In contrast, we derive from first principles a CLT for the count-based estimator of controlled transition probabilities themselves under general logging policies, including fully non-stationary, stochastic, and history-dependent policies. By establishing transition-level asymptotics without stationarity, our results provide a fundamentally different inference foundation for OPE and OPR.

## 2 Notations, Background and Assumptions

### 2.1 Notations

We assume that the state process  $\{X_i\}_{i \geq 0}$  takes values in a finite state space  $\mathcal{S}$  with  $|\mathcal{S}| = d$ , and the action process  $\{a_i\}_{i \geq 0}$  takes values in a finite action space  $\mathcal{A}$  with  $|\mathcal{A}| = k$ . Throughout this paper, all random variables are defined on a filtered probability space  $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ , where  $\mathbb{F} := \{\mathcal{F}_j\}_{j \geq 0}$ , and  $\mathcal{F}_j$  is a filtration with  $\mathcal{F}_j \subset \mathcal{F}$  defined as below. Let the history from time  $p$  to  $j$  be denoted by  $\mathcal{H}_p^j := \{(X_i, a_i)\}_{i=p}^j$ , and define the  $\sigma$ -algebra generated by  $\mathcal{H}_0^j$  as  $\mathcal{F}_j$ . For readability and to avoid over-notation, we use  $\mathcal{F}_j$  in measure-theoretic contexts as in Equation (2.4) and  $\mathcal{H}_0^j$  when writing explicit conditioning in transition probabilities as in Equation (2.1).

Following Borkar (1991), we introduce the following definition of a controlled Markov chain.

**Definition 1** A controlled Markov chain (CMC) is an  $\mathbb{F}$ -adapted pair process  $\{(X_i, a_i)\}_{i \geq 0}$  whose transition probabilities are given by

$$M_{s,t}^{(l)} := \mathbb{P}(X_{i+1} = s \mid X_i = t, a_i = l), \quad \text{for all time points } i \quad (2.1)$$

and which satisfies the Markov property

$$\mathbb{P}(X_{i+1} = s_{i+1} \mid X_i = s_i, a_i = l_i) = \mathbb{P}(X_{i+1} = s_{i+1} \mid \mathcal{H}_0^i = \bar{h}_0^i),$$

where  $\bar{h}_0^i := \{X_0 = s_0, a_0 = l_0, \dots, X_i = s_i, a_i = l_i\}$  is the sample history up to time  $i$ . A Markov decision process (MDP) is a CMC with a reward process  $\{r_i\}_{i \geq 0}$ .

The action sequence  $\{a_i\}$  is adapted to  $\{\mathcal{F}_i\}$  for all  $i \geq 0$  and is allowed to be non-Markovian and time-inhomogeneous. For example, when  $\{a_i\}$  is deterministic,  $\{X_i\}$  forms a time-inhomogeneous Markov chain whose transition matrix at time  $i$  is determined by  $a_i$ . Example 1 illustrates such a chain. Conversely, when  $a_i$  depends only on  $X_i$ , the pair process  $\{(X_i, a_i)\}$  forms a time-homogeneous Markov chain.

For each state  $s \in \mathcal{S}$ , let  $M_s$  represent the matrix

$$\begin{bmatrix} M_{s,1}^{(1)}, \dots, M_{s,1}^{(k)} \\ M_{s,2}^{(1)}, \dots, M_{s,2}^{(k)} \\ \vdots \\ M_{s,d-1}^{(1)}, \dots, M_{s,d-1}^{(k)} \\ M_{s,d}^{(1)}, \dots, M_{s,d}^{(k)} \end{bmatrix}. \quad (2.2)$$

As  $d$  and  $k$  are finite, the collection of such matrices is finite.

Let  $N_s^{(l)}$  be the number of visits to the (state, action) pair  $(s, l)$ , and  $N_{s,t}^{(l)}$  be the number of transitions from state  $s$  to state  $t$  under  $l$ . Formally, with  $\mathbb{1}[\cdot]$  as indicator function,

$$N_s^{(l)} := \sum_{i=0}^{n-1} \mathbb{1}[X_i = s, a_i = l], \quad N_{s,t}^{(l)} := \sum_{i=0}^{n-1} \mathbb{1}[X_i = s, X_{i+1} = t, a_i = l].$$

These quantities depend on the size of the offline data  $n$ . We omit this dependency when there is no scope of confusion, otherwise we write  $N_s^{(l)}(n)$  and  $N_{s,t}^{(l)}(n)$ .

Our count-based estimator denoted  $\hat{M}_{s,t}^{(l)}$  of the transition probability from state  $s$  to  $t$  conditioned on action  $l$ , is defined as

$$\hat{M}_{s,t}^{(l)} := \frac{N_{s,t}^{(l)}}{N_s^{(l)}}. \quad (2.3)$$

It follows from first principles that the random variable  $\hat{M}_{s,t}^{(l)}$  is the maximum likelihood estimator of the transition probability  $M_{s,t}^{(l)}$ . The following proposition formalizes this and we provide a proof in Appendix E.1.

**Proposition 1** Random variable  $\hat{M}_{s,t}^{(l)}$  is the maximum likelihood estimator of the transition probability  $M_{s,t}^{(l)}$ .

Understanding the asymptotic distribution of maximum likelihood estimators is a classic topic in statistics, with a long history in the analysis of i.i.d. and stationary data (Cramér, 1946; Billingsley, 1961). Our goal here is to establish the asymptotic normality of the scaled estimation error

$$b_n \left( \hat{M}_{s,t}^{(l)}(n) - M_{s,t}^{(l)} \right),$$

where  $b_n$  is a (possibly random) scaling factor.

In our setting, however, the challenge is considerably greater: the data is non-stationary, and this dependence structure makes determining the correct scaling in any CLT argument particularly delicate. Our analysis reveals that the correct scaling factor is random and, in fact, equal to the visitation count  $N_s^{(l)}(n)$ .

Establishing such a CLT requires both sufficient coverage of the relevant state–action pairs and suitably weak temporal dependence. To formalize these conditions, we next introduce the necessary background concepts and assumptions.

## 2.2 Assumptions and Implications

**Hitting and Return Times.** For standard (uncontrolled) Markov chains, the notion of *recurrence* is rigorously characterized using hitting and return times. In particular, a Markov chain can be classified as either positive recurrent, null recurrent, or transient—an exhaustive and mutually exclusive trichotomy. To the best of our knowledge, no canonical analogue of this classification exists for CMCs. We therefore define analogous notions through the return time of the state-action pairs. We recursively define the ‘time of return’ for every state-action pair  $(s, l)$  as follows.

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

$$\tau_{s,l}^{(1)} \in \min \{i > 0 : X_i = s, a_i = l\}.$$

*When  $i \geq 2$ , the  $i$ -th time to return (or return time) of the state-action pair  $(s, l)$  is recursively defined as*

$$\tau_{s,l}^{(i)} \in \min \left\{ k > 0 : X_{\sum_{j=1}^{i-1} \tau_{s,l}^{(j)} + k} = s, a_{\sum_{j=1}^{i-1} \tau_{s,l}^{(j)} + k} = l \right\}.$$

*In these definitions, the sets on the right-hand side are singletons.*

The following assumption plays a role analogous to *positive recurrence* in standard Markov chains, ensuring that every state-action pair  $(s, l)$  is visited sufficiently often.

**Assumption 1** *For all state-action pairs  $(s, l) \in \mathcal{S} \times \mathcal{A}$ , there exists  $0 < \alpha < 1$  such that*

$$\mathbb{E} \left[ \tau_{s,l}^{(i)} \mid \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] \leq T_i \quad a.s., \quad T_i = O(i^\alpha). \quad (2.4)$$

Unlike previous assumptions in the literature (like that of regularity in Markov chains—see Theorem 11.0.1 or Equation (11.1) in Meyn and Tweedie (2012) or Assumption 2 in Banerjee et al. (2025a))—which require  $T_i$  in (2.4) to be uniformly bounded for all  $i$  Assumption 1 allows the expected return time to grow sublinearly with  $i$ , which admits CLT

for non-stationary CMCs in which some state–action pairs are visited increasingly rarely over time, yet still infinitely often with probability one (jointly established by Proposition 2 and Lemma 2). This relaxation allows the (possibly non-stationary and history-dependent) logging policy to concentrate sampling on only a subset of state–action pairs rather than enforcing uniform exploration of the entire state–action space. Example 1 illustrates such a non-stationary CMC, and Appendix A.1 discusses this example in detail.

**Example 1 (Inhomogeneous Markov chain)** *A controlled Markov chain is said to be an inhomogeneous Markov chain if there exists a sequence of actions  $l_0, l_1, \dots \in \mathcal{A}$  such that for any time  $i$ , state  $s$ , sample history  $\mathcal{H}_0^{i-1} \in (\mathcal{S} \times \mathcal{A})^i$ , we have*

$$\mathbb{P}(a_i = l_i \mid \mathcal{H}_0^{i-1} = \mathcal{h}_0^{i-1}, X_i = s) = 1.$$

*Suppose that each action  $l \in \mathcal{A}$  appears at least once in any window of fixed length  $T_i = O(i^\alpha)$  between successive returns to any state–action pair  $(s, l)$ , and that all transition probabilities are strictly positive (i.e.,  $M_{s,t}^{(l)} \geq M_{\min} > 0$  for all  $s, t, l$ ). We show in Appendix A.1 that, for this process, the probability of not returning to  $(s, l)$  within  $k$  steps conditional on the past is bounded by*

$$\mathbb{P}\left(\tau_{s,l}^{(i)} > k \mid \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}}\right) \leq (1 - M_{\min})^{\lfloor k/T_i \rfloor - 1},$$

which implies

$$\mathbb{E}\left[\tau_{s,l}^{(i)} \mid \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}}\right] = \sum_{k=1}^{\infty} \mathbb{P}\left(\tau_{s,l}^{(i)} > k \mid \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}}\right) = O(i^\alpha).$$

Hence Assumption 1 holds. The full details are formalized in Proposition 5 in Appendix A.1.

The following proposition shows that Assumption 1 guarantees a minimum visitation frequency for all state–action pairs.

**Proposition 2** *For controlled Markov chain that satisfies Assumption 1, we have*

$$\mathbb{E}\left[N_s^{(l)}(n)\right] = \Omega\left(n^{\frac{1}{1+\alpha}}\right).$$

The proof of this proposition can be found in Appendix E.2.

**Mixing Coefficients.** Following Kontorovich et al. (2008), we define the weak and uniform mixing coefficients to quantify temporal dependence in the paired process of states and actions. Intuitively, these coefficients measure how much the future of the process can change if we alter the state–action pair at a single point of the past.

For any  $j \leq n$ ,  $j, n \in \mathbb{N}$ , sample history  $\mathcal{H}_0^{i-1} \in (\mathcal{S} \times \mathcal{A})^i$ , let  $\mathcal{T} \subseteq (\mathcal{S} \times \mathcal{A})^{n-j+1}$ ,  $s_1, s_2 \in \mathcal{S}$ , and  $l_1, l_2 \in \mathcal{A}$ . For any  $\mathcal{h}_0^{i-1}$  and  $i < j$ , we define

$$\eta_{i,j}(\mathcal{T}, s_1, s_2, l_1, l_2, \mathcal{h}_0^{i-1}, n) := \left| \mathbb{P}\left((X_j, a_j, \dots, X_n, a_n) \in \mathcal{T} \mid X_i = s_1, a_i = l_1, \mathcal{H}_0^{i-1} = \mathcal{h}_0^{i-1}\right) - \mathbb{P}\left((X_j, a_j, \dots, X_n, a_n) \in \mathcal{T} \mid X_i = s_2, a_i = l_2, \mathcal{H}_0^{i-1} = \mathcal{h}_0^{i-1}\right) \right|.$$

Then the *weak mixing* ( $\bar{\eta}$ -mixing) coefficient  $\bar{\eta}_{i,j}(n)$  is defined as

$$\bar{\eta}_{i,j} := \sup_{\substack{\mathcal{T}, s_1, s_2, l_1, l_2, h_0^{i-1}, \\ \mathbb{P}(X_i=s_1, a_i=l_1, \mathcal{H}_0^{i-1}=h_0^{i-1})>0, \\ \mathbb{P}(X_i=s_2, a_i=l_2, \mathcal{H}_0^{i-1}=h_0^{i-1})>0}} \eta_{i,j}(\mathcal{T}, s_1, s_2, l_1, l_2, h_0^{i-1}, n). \quad (2.5)$$

We impose the following boundedness condition on the cumulative decay of  $\bar{\eta}$ -mixing coefficients of the paired process  $\{(X_i, a_i)\}$ .

**Assumption 2** *There exists a constant  $C_\Delta > 1$  independent of  $n$  such that,*

$$\|\Delta_n\| := \max_{1 \leq i \leq n} (1 + \bar{\eta}_{i,i+1}(n) + \bar{\eta}_{i,i+2}(n) + \dots + \bar{\eta}_{i,n}(n)) \leq C_\Delta.$$

This assumption controls the cumulative temporal dependence of the paired process and, in particular, it helps bound the deviation of  $N_s^{(l)}$  from its expectation  $\mathbb{E}[N_s^{(l)}]$ , a key step in establishing CLT.

However, verifying this condition directly is typically difficult; it requires bounding joint dependencies between all states and actions over the entire trajectory, which may be analytically intractable.

To make the condition more interpretable and easier to verify, we next show how it can be reduced to separate assumptions on the mixing of the state process and the action process. This decomposition isolates two distinct sources of dependence, allowing each to be analyzed independently.

**Reduction of Mixing Coefficients.** We begin by defining the  $\gamma$ -mixing coefficients  $\gamma_{p,j,i}$  for actions as the following total variation distance

$$\begin{aligned} \gamma_{p,j,i} := & \sup_{\substack{s_p, h_{i+j}^{p-1}, h_0^i \\ \mathbb{P}(X_p=s_p, \mathcal{H}_{i+j}^{p-1}=h_{i+j}^{p-1}, \mathcal{H}_0^i=h_0^i)>0}} \left\| \mathbb{P}\left(a_p \mid X_p = s_p, \mathcal{H}_{i+j}^{p-1} = h_{i+j}^{p-1}, \mathcal{H}_0^i = h_0^i\right) \right. \\ & \left. - \mathbb{P}\left(a_p \mid X_p = s_p, \mathcal{H}_{i+j}^{p-1} = h_{i+j}^{p-1}\right) \right\|_{TV}. \end{aligned}$$

The total variation norm here captures the worst-case sensitivity of the future action distribution to distant past information, thereby quantifying the degree of temporal dependence induced by the logging policy.

We now impose a summability condition on these  $\gamma$ -mixing coefficients. This condition requires that, as the time lag increases, the effect of earlier histories on future actions decays sufficiently fast.

**Assumption 3 (Mixing of Actions)** *There exists a constant  $C \geq 0$  such that*

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

In Markovian settings, when the sequence of actions  $a_p$  depends only on  $X_p$ ,  $\gamma_{p,j,i} = 0$  for all  $p, j, i$ . In such a case, Assumption 3 is satisfied with  $C = 0$ . This extends to the case where  $a_p$  depends on  $k \geq 2$  many past time points. If  $a_p$  depends only on  $X_{p-k+1}, a_{p-k+1}, \dots, X_{p-1}, a_{p-1}, X_p$ , then Assumption 3 is satisfied with  $C = (k-1)(k-2)$ .

With the definitions of  $\gamma$ -mixing coefficients for actions in hand, we introduce counterparts for the state process, which capture how the conditional distribution of future states depends on the current state–action pair. This extends the classical Dobrushin coefficients (Dobrushin, 1956a,b; Mukhamedov, 2013) for inhomogenous Markov chains to the realm of controlled Markov chains.

For all integers  $j \geq i$ , we define the mixing coefficient

$$\bar{\theta}_{i,j} := \sup_{\substack{s_1, s_2 \in \mathcal{S}, l_1, l_2 \in \mathcal{A}, (s_1, l_1) \neq (s_2, l_2) \\ \mathbb{P}(X_i = s_1, a_i = l_1) > 0, \\ \mathbb{P}(X_i = s_2, a_i = l_2) > 0}} \|\mathbb{P}(X_j | X_i = s_1, a_i = l_1) - \mathbb{P}(X_j | X_i = s_2, a_i = l_2)\|_{TV}. \quad (2.6)$$

Intuitively,  $\bar{\theta}_{i,j}$  measures the rate at which the *state process* “forgets” its starting point when no condition on controls are given.

We now impose a summability condition on these  $\bar{\theta}$ -mixing coefficients. This condition ensures that the influence of an initial state–action pair on future states decays sufficiently fast.

**Assumption 4 (Mixing of States)** *There exists a constant  $C_\theta \geq 0$  such that*

$$\sup_{i \geq 1} \sum_{j=i+1}^{\infty} \bar{\theta}_{i,j} \leq C_\theta.$$

Example 2 illustrates a non-stationary CMC that satisfies Assumptions 3 and 4, and Appendix A.2 discusses this example in detail.

**Example 2 (Non-stationary Markov controls)** *A controlled Markov chain is said to have non-stationary Markov controls if for any time  $i$ , state  $s$ , action  $l$ , and sample history  $\mathcal{H}_0^{i-1}$ , we have*

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

*We observe that the law of the action sequence can depend on the time  $i$ . Let  $P_{s,l}^{(i)} := \mathbb{P}(a_i = l | X_i = s)$ . We can write the transition probability of the state-action pair as*

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

*It is straightforward to see that the state-action pair is a time-inhomogeneous Markov chain with transition probabilities given by  $M_{s,t}^{(l)} P_{s,l'}^{(i)}$ .*

As  $a_i$  depends only on the current state  $X_i$  (and not on deeper history), it has no long-range dependence. Thus,

$$\begin{aligned} \gamma_{p,j,i} &= \sup_{s_p, h_{i+j}^{p-1}, h_0^i} \left\| \mathbb{P} \left( a_p \mid X_p = s_p, \mathcal{H}_{i+j}^{p-1} = \bar{h}_{i+j}^{p-1}, \mathcal{H}_0^i = \bar{h}_0^i \right) - \mathbb{P} \left( a_p \mid X_p = s_p, \mathcal{H}_{i+j}^{p-1} = \bar{h}_{i+j}^{p-1} \right) \right\|_{TV} \\ &\equiv 0. \end{aligned}$$

Therefore, Assumption 3 holds for all controlled Markov chains with non-stationary controls with  $C = 0$ .

Suppose that all transition probabilities are strictly positive (i.e.,  $M_{s,t}^{(l)} \geq M_{\min} > 0$  for all  $s, t, l$ ). Then, by Banerjee et al. (2025a, Lemma 12), Assumption 4 holds with  $C_\theta = 1/(dM_{\min})$ .

Neither of Assumptions 3 and 4 imply the other, as the following counter-examples illustrate.

1. Let  $(X_i, a_i)$  be an inhomogeneous Markov chain for which  $\sup_{1 \leq i \leq \infty} \sum_{j=i+1}^{\infty} \bar{\theta}_{i,j} = \infty$ . However, since the actions are deterministic, every inhomogeneous Markov chain satisfies Assumption 3. We prove this fact formally in Proposition 5. Therefore, this chain satisfies Assumption 3 but not Assumption 4.
2. For the second counter-example consider a controlled Markov chain  $(X_i, a_i)$  where the  $a_i$ 's do not satisfy Assumption 3. Let  $X_i$  be independent draws from a uniform distribution over  $\mathcal{S}$ . It is easily seen that  $\bar{\theta}_{i,j} = 0$  for this example. Therefore, this chain satisfies Assumption 4 but not 3.

Finally, we observe that the previous assumptions on the states and actions imply the summability of the weak mixing coefficients. Due to Banerjee et al. (2025a, Lemma 3), for any controlled Markov chain that satisfies Assumptions 3 and 4,  $\|\Delta_n\| \leq C + C_\theta + 1$ .

**Ergodic Occupation Measure.** Next, we define the *ergodic occupation measure* in the context of CMCs.

**Definition 3 (Ergodic Occupation Measure)** For every  $(s, l) \in \mathcal{S} \times \mathcal{A}$ , we define

$$p_s^{(l)} := \lim_{n \rightarrow \infty} \frac{\sum_{i=0}^{n-1} \mathbb{P}(X_i = s, a_i = l)}{n},$$

whenever the limit exists. The limit  $p_s^{(l)}$ , if it exists for all  $(s, l) \in \mathcal{S} \times \mathcal{A}$ , is called the *ergodic occupation measure of the CMC*.

Observe that for stationary CMCs,  $p_s^{(l)} = \mathbb{P}(X_0 = s, a_0 = l)$ . Some usual processes, like episodic CMCs, are not stationary, but an ergodic occupation measure exists. We make the following assumption essential for the CLTs for value, Q-, and advantage functions.

**Assumption 5** The CMC has the ergodic occupation measure, and  $\min_{s,l} p_s^{(l)} = T > 0$ .

The previous assumption can be relaxed with appropriate assumptions on the return time of  $(X_i, a_i)$  which would translate into an assumption about the logging policy.

### 3 CLT for Controlled Markov Chains

In this section, we establish CLTs for count-based empirical estimates of transition probabilities in CMCs. We begin with the “properly scaled” CLT, which characterizes the asymptotic distribution of transition estimates under state-action pair-specific normalization, and provide a sketch of its proof, highlighting a new sampling construction that generalizes the classical approach of Billingsley (1961). We then introduce an auxiliary,  $\sqrt{n}$ -scaled CLT that serves as the basis for the CLTs of value,  $Q$ -, and advantage functions in Section 4, followed by a discussion of regimes where no CLT exists. The section concludes with a goodness-of-fit test that illustrates the practical implications of the established asymptotic results.

#### 3.1 Properly Scaled CLT

We begin with introducing additional notations. Let  $\odot$  denote the Schur or Hadamard product of two compatible vectors and let  $\otimes$  denote the Kronecker product. Let  $\text{Vec}(A)$  be the vectorization of the matrix  $A$  given by stacking the columns.

Let  $\mathbf{N}$  be the  $dk \times d$  dimensional matrix given by

$$\begin{bmatrix} \left( \sqrt{N_1^{(l)}} \right)_{l=1}^k \otimes \mathbf{1}_d^\top \\ \vdots \\ \left( \sqrt{N_d^{(l)}} \right)_{l=1}^k \otimes \mathbf{1}_d^\top \end{bmatrix}, \text{ where } \mathbf{1}_d^\top := (1, \dots, 1)^\top \in \mathbb{R}^d.$$

The  $((s-1)k + l, j)$ -th entry of  $\mathbf{N}$  equals  $\sqrt{N_s^{(l)}}$  for all  $1 \leq j \leq d$ .

Let  $\mathbf{M}$  be the  $dk \times d$  dimensional matrix given by  $[M_1, M_2, \dots, M_d]^\top$ , where  $M_s$  is given in Equation (2.2). We denote its count-based estimated counterpart by  $\hat{\mathbf{M}}$ . Let  $\xi := \left( \text{Vec}(\hat{\mathbf{M}}^\top) - \text{Vec}(\mathbf{M}^\top) \right) \odot \text{Vec}(\mathbf{N}^\top)$ . Elementwise, the  $(s-1)dk + (l-1)d + t$ -th element of  $\xi$  is  $\sqrt{N_s^{(l)}} \left( \hat{M}_{s,t}^{(l)} - M_{s,t}^{(l)} \right)$ . We now state the CLT for count-based empirical estimates of transition probabilities in CMCs.

**Theorem 1** *Under Assumptions 1 and 2,*

$$\xi \xrightarrow{d} \mathcal{N}(0, \Lambda),$$

where  $\Lambda$  is a covariance matrix. The elements of  $\Lambda$  are given by  $\lambda_{slt, s'l't'}$  which denotes the covariance between the  $(s-1)dk + (l-1)d + t$ -th and the  $(s'-1)dk + (l'-1)d + t'$ -th elements of  $\xi$ . Value  $\lambda_{slt, s'l't'}$  has the expression

$$\lambda_{slt, s'l't'} = \mathbb{1}[(s, l) = (s', l')] \left( \mathbb{1}[t = t'] M_{s,t}^{(l)} - M_{s,t}^{(l)} M_{s',t'}^{(l')} \right). \quad (3.1)$$

We observe that the covariance matrix  $\Lambda$  depends only on the transition probabilities  $\{M_{s,t}^{(l)}\}$  and does not depend on the existence of an ergodic occupation measure. Consequently,

the asymptotic covariance structure is invariant to the logging policy, provided the conditions of Theorem 1 hold.

Theorem 1 generalizes the classical CLT for ergodic Markov chains (Billingsley, 1961) to the setting of controlled and potentially non-stationary dynamics. Unlike standard Markov chain CLTs that rely on ergodicity or stationarity, our result accommodates both non-Markovian and time-inhomogeneous dependencies in the action sequence and transition dynamics. The covariance matrix  $\Lambda$  mirrors that of multinomial trials, capturing the asymptotic covariance of transition counts within each state–action pair and the asymptotic independence across distinct pairs. In practice,  $\Lambda$  can be consistently estimated by replacing  $M_{s,t}^{(l)}$  with their empirical counterparts  $\hat{M}_{s,t}^{(l)}$  in Equation (3.1).

The main challenge in proving Theorem 1 is the long-range dependence between the state and action sequences in a CMC induced by history-dependent, non-stationary policies. This dependence breaks the Markov property of the marginal state process and invalidates classical proof techniques based on regenerative arguments (Doebelin, 1938; Dobrushin, 1956a,b; Nummelin, 1978). To overcome this difficulty, we introduce an auxiliary sampling scheme that replicates the joint evolution of  $(X_i, a_i)$  while decoupling temporal dependence across samples. This construction extends the seminal sampling scheme approach of Billingsley (1961) to the controlled and non-stationary setting. It provides the key technical innovation that enables the use of multinomial-type CLT arguments in the controlled regime.

We next provide a proof sketch of Theorem 1. The complete proof is deferred to Appendix C.

### 3.2 Proof Sketch of Theorem 1

**Proof** The proof adapts the classical CLT for ergodic, finite Markov chains (Billingsley, 1961) to the controlled setting by constructing an auxiliary sampling scheme that decouples temporal dependence from the empirical transition counts. We outline the key steps.

**Step 1 (Consistency).** We first establish (see Lemma 2 in Appendix B.1) that

$$\frac{N_s^{(l)}}{\mathbb{E} \left[ N_s^{(l)} \right]} \xrightarrow{a.s.} 1.$$

**Step 2 (Sampling Scheme).** For each action  $l \in \mathcal{A}$ , we construct an infinite array of i.i.d. random variables that are also independent of the observed data  $\{(X_0, a_0), \dots, (X_n, a_n)\}$ :

$$\mathbb{X}^{(l)} = \begin{bmatrix} X_{1,1}^{(l)} & X_{1,2}^{(l)} & \dots & X_{1,\tau}^{(l)} & \dots \\ X_{2,1}^{(l)} & X_{2,2}^{(l)} & \dots & X_{2,\tau}^{(l)} & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{d,1}^{(l)} & X_{d,2}^{(l)} & \dots & X_{d,\tau}^{(l)} & \dots \end{bmatrix}.$$

Each element  $X_{s,\tau}^{(l)}$  represents a possible next state when action  $l$  is taken from state  $s$ , and follows the transition law

$$\mathbb{P} \left( X_{s,\tau}^{(l)} = t \right) = M_{s,t}^{(l)}, \quad (s, t, \tau) \in \mathcal{S} \times \mathcal{S} \times \mathbb{N}.$$

In addition, for every time  $i \geq 1$  and history  $(h_0^{i-1}, s_i) = ((s_0, l_0), \dots, (s_{i-1}, l_{i-1}), s_i) \in (\mathcal{S} \times \mathcal{A})^i \times \mathcal{S}$ , we define an independent random variable

$$\alpha_i^{h_0^{i-1}, s_i} \in \mathcal{A},$$

with conditional distribution

$$\mathbb{P} \left( \alpha_i^{h_0^{i-1}, s_i} = l \right) = \mathbb{P} (a_i = l \mid X_i = s_i, \mathcal{H}_0^{i-1} = h_0^{i-1}).$$

The sampling proceeds recursively as follows. We initialize  $(\tilde{X}_0, \tilde{a}_0) \stackrel{d}{=} (X_0, a_0)$ . At each step  $i \geq 0$ ,

$$\tilde{X}_{i+1} = X_{\tilde{X}_i, \tilde{N}_{\tilde{X}_i}^{(i, \tilde{a}_i)} + 1}^{(\tilde{a}_i)}, \quad \tilde{a}_{i+1} = \alpha_{i+1}^{(\tilde{X}_0, \tilde{a}_0, \dots, \tilde{X}_{i+1})},$$

where  $\tilde{N}_s^{(i, l)} = \sum_{j \leq i} \mathbb{1}[\tilde{X}_j = s, \tilde{a}_j = l]$  counts the number of visits to  $(s, l)$  up to time  $i$ . Finally, let  $\tilde{N}_s^{(l)} = \sum_{i=1}^n \mathbb{1}[\tilde{X}_i = s, \tilde{a}_i = l]$  denote the total number of visits under the constructed process.

Intuitively, each array  $\mathbb{X}^{(l)}$  provides an independent “stream” of next-state samples for action  $l$ , while the variables  $\alpha_i$  govern the (possibly non-Markovian) control mechanism. The resulting sequence  $(\tilde{X}_i, \tilde{a}_i)$  thus preserves the same joint distribution as the original process  $(X_i, a_i)$  but decouples temporal dependence across samples. Formally, we show in Proposition 7 in Appendix B.2 that  $(\tilde{X}_0, \tilde{a}_0, \dots, \tilde{X}_n, \tilde{a}_n) \stackrel{d}{=} (X_0, a_0, \dots, X_n, a_n)$ , yielding a coupling that retains the CMC law while enabling the application of multinomial CLTs.

**Step 3 (Approximate Multinomial Representation).** For each  $(s, l)$ , we define

$$\tilde{N}_{s,t}^{(l)} = \frac{\mathbb{E}[N_s^{(l)}]}{\sum_{\tau=1} \mathbb{1}[\tilde{X}_{s,\tau}^{(l)} = t]}, \quad \tilde{\xi}_{s,l,t} = \frac{\tilde{N}_{s,t}^{(l)} - \mathbb{E}[N_s^{(l)}]}{\sqrt{\mathbb{E}[N_s^{(l)}]}} M_{s,t}^{(l)}.$$

Conditional on  $N_s^{(l)}$ , the vector  $(\tilde{N}_{s,1}^{(l)}, \dots, \tilde{N}_{s,d}^{(l)})$  follows a multinomial distribution with number of trials  $\mathbb{E}[N_s^{(l)}]$  and event probabilities  $(M_{s,1}^{(l)}, \dots, M_{s,d}^{(l)})$ , implying that

$$\tilde{\xi} \xrightarrow{d} \mathcal{N}(0, \Lambda),$$

where  $\Lambda$  is given by Equation (3.1).

**Step 4 (Equivalence and Convergence).** In this step (Lemma 3 in Appendix B.3) we show that the difference between the empirical and auxiliary statistics vanishes in probability:

$$\tilde{\xi}_{s,l,t} - \xi_{s,l,t} \xrightarrow{p} 0.$$

Combining these steps yields  $\xi \xrightarrow{d} \mathcal{N}(0, \Lambda)$ , establishing Theorem 1. ■

### 3.3 Improperly Scaled CLT

While Theorem 1 establishes asymptotic normality under state-action pair-specific normalization by  $\sqrt{N_s^{(l)}}$ , certain downstream analyses—such as aggregating transition effects or studying plug-in functionals—require a common global scaling. To this end, we introduce an “improperly scaled” CLT, obtained by normalizing all transition estimates by the same factor (typically  $\sqrt{n}$ ). Although this scaling is no longer correct for individual state–action pairs, it yields a unified limit that is analytically convenient and forms the basis for the CLTs of value, Q-, and advantage functions in Section 4.

For a semi-ergodic controlled Markov chain with no absorbing states, we define  $\rho$  to be the matrix populated by the inverse of the square roots of the  $p_s^{(l)}$  probabilities,

$$\rho = \begin{bmatrix} \left(1/\sqrt{p_1^{(l)}}\right)_{l=1}^k \otimes \mathbf{1}_d^\top \\ \vdots \\ \left(1/\sqrt{p_d^{(l)}}\right)_{l=1}^k \otimes \mathbf{1}_d^\top \end{bmatrix},$$

where

$$\left(1/\sqrt{p_s^{(l)}}\right)_{l=1}^k := \begin{bmatrix} 1/\sqrt{p_s^{(1)}} \\ \vdots \\ 1/\sqrt{p_s^{(k)}} \end{bmatrix} \in \mathbb{R}^k.$$

Using Theorem 1 and Lemma 2, the proof of the following corollary is immediate.

**Corollary 1** *Let  $\bar{\xi} := \sqrt{n} \left( \text{Vec}(\hat{\mathbf{M}}^\top) - \text{Vec}(\mathbf{M}^\top) \right)$  be the “improperly scaled” vector of the differences between  $\hat{\mathbf{M}}$  and  $\mathbf{M}$ . Consider the setting of Theorem 1 and suppose that Assumptions 1, 2 and 5 hold. Then,*

$$\bar{\xi} \xrightarrow{d} \mathcal{N}(0, \bar{\Lambda})$$

where  $\bar{\Lambda}$  is the improperly scaled covariance matrix. The elements of  $\bar{\Lambda}$  are given by  $\bar{\lambda}_{slt,s'l't'}^{(is)}$  which denotes the covariance between the  $(s-1)dk + (l-1)d + t$ -th and the  $(s'-1)dk + (l'-1)d + t'$ -th elements of  $\bar{\xi}$ . Value  $\bar{\lambda}_{slt,s'l't'}$  has the expression

$$\bar{\lambda}_{slt,s'l't'} = \frac{\mathbb{1}[(s, l) = (s', l')] \left( \mathbb{1}[t = t'] M_{s,t}^{(l)} - M_{s,t}^{(l)} M_{s',t'}^{(l')} \right)}{\sqrt{p_s^{(l)} p_{s'}^{(l')}}}. \quad (3.2)$$

While the statement is more general than the one in Theorem 1, due to generous scaling, we require Assumption 5.

### 3.4 Non-Existence of CLT

The validity of the preceding CLTs relies critically on the assumption that every state–action pair is visited infinitely often with a sufficiently regular frequency. When this condition fails—such as under transient or weakly exploratory control policies—the empirical transition estimates may not satisfy any CLT irrespective of scalings. This subsection formalizes these failure cases and delineates the boundary between recurrent regimes admitting asymptotic normality and those where no normalization yields a tight limiting distribution.

The following proposition entails occasions when there exists no CLT.

**Proposition 3** *If there exists a state-action pair  $(s, l) \in \mathcal{S} \times \mathcal{A}$  such that  $\mathbb{P}(a_i = l | X_i = s)$  is independent of  $d$  for all  $i \geq 0$ , and*

$$\sum_{i=0}^{\infty} \mathbb{P}(a_i = l | X_i = s) < \infty,$$

then

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{P}(X_i = s, a_i = l) < \infty,$$

and there exists a CMC with transition kernel  $M$  such that for any possible sequence of estimators  $\{\hat{M}_{s,t}^{(l)}\}_{n \geq 1}$  and for any positive sequence  $\{b_n\}$  with  $b_n \rightarrow \infty$ ,

$$\left\{ b_n \sup_{s,l,t} |\hat{M}_{s,t}^{(l)}(n) - M_{s,t}^{(l)}| \right\}_{n \geq 1} \quad (3.3)$$

is not tight i.e., for every  $\varepsilon > 0$ ,

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left( b_n \sup_{s,l,t} |\hat{M}_{s,t}^{(l)}(n) - M_{s,t}^{(l)}| > \varepsilon \right) > 0.$$

The proof of this proposition can be found in Appendix E.3.

We remark that in contrast to Proposition 2, it can be verified using the Borel-Cantelli lemma that  $\lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{P}(X_i = s, a_i = l) < \infty$  implies

$$\sum_{i=1}^{\infty} \mathbb{P} \left( N_s^{(l)}(\infty) \geq i - 1 \right) / T_i < \infty$$

where  $N_s^{(l)}(\infty) := \lim_{n \rightarrow \infty} N_s^{(l)}(n)$ .

Unlike in the uncontrolled setting, null-recurrence properties of CMCs are entwined with non-stationarity of the control sequence. Some controls may ensure persistent exploration of the entire state-action space, while others may confine the trajectory to a restricted subset. Consequently, the sharp positive-recurrent/null-recurrent/transient classification in Markov chains is replaced by a policy-dependent continuum in CMCs. This leads to a “grey area” in defining null-recurrence, where the trajectory may drift widely under non-stationary controls while still revisiting a fixed state–action pair every so often. We classify null-recurrence as the razor’s edge between the regimes of Propositions 2 and 3. The statistical properties of such a controlled Markov chain are difficult to characterize by the current analysis. However, since exploration problems are not necessarily recurrent (Mutti et al., 2022), this remains an important open direction for future studies.

### 3.5 Goodness-of-Fit Test for the Transition Probabilities

Beyond asymptotic characterization, the properly scaled CLT insinuates a method for direct statistical inference on transition probabilities through a goodness-of-fit (G.O.F.) test that assesses whether an estimated transition kernel is consistent with a hypothesized model. The test statistic leverages the covariance structure  $\Lambda$  from Theorem 1, leading to a chi-square-type limit under the null hypothesis and providing a practical diagnostic for model parameters. One natural example is testing whether the observed data arise from a fully controlled MDP—where the next state depends on both state and action—or from a contextual bandit model, in which transitions are independent of the previous state. This can be thought of as a CMC counterpart of the Scheffé’s multiple comparison test (Scheffé, 1953).

Formally, let  $\{M_{s,t}^{(l)}\}_{(s,l,t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}}$  denote a fixed, hypothesized transition kernel. We consider the null hypothesis

$$H_0 : \quad \mathbb{P}(X_{i+1} = t \mid X_i = s, a_i = l) = M_{s,t}^{(l)} \quad \text{for all } (s, l, t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}.$$

Under  $H_0$ , Theorem 1 implies that the scaled estimation errors are asymptotically Gaussian, which in turn yields chi-square test statistics. The following proposition states the resulting asymptotic distribution.

**Proposition 4** *Given any  $(s, l) \in \mathcal{S} \times \mathcal{A}$ , under the null hypothesis  $H_0$  and Assumptions 1 and 2,*

$$\sum_{t \in \mathcal{S}} \frac{\left(N_{s,t}^{(l)} - N_s^{(l)} M_{s,t}^{(l)}\right)^2}{N_s^{(l)} M_{s,t}^{(l)}} \xrightarrow{d} \chi_{d_{(s,l)} - 1}^2,$$

where  $\chi_{d_{(s,l)} - 1}^2$  is the chi-square distribution with  $d_{(s,l)} - 1$  degrees of freedom, and  $d_{(s,l)} = \sum_{t \in \mathcal{S}} \mathbb{1}[M_{s,t}^{(l)} > 0]$ .

Furthermore, the  $dk$  chi-square statistics  $\{\chi_{d_{(s,l)} - 1}^2\}_{(s,l) \in \mathcal{S} \times \mathcal{A}}$  are asymptotically independent, and

$$\sum_{(s,l) \in \mathcal{S} \times \mathcal{A}, t \in \mathcal{S}} \frac{\left(N_{s,t}^{(l)} - N_s^{(l)} M_{s,t}^{(l)}\right)^2}{N_s^{(l)} M_{s,t}^{(l)}} \xrightarrow{d} \chi_{\sum_{(s,l) \in \mathcal{S} \times \mathcal{A}} d_{(s,l)} - dk}^2. \quad (3.4)$$

Equation (3.4) hence provides a G.O.F. measure for assessing the validity of assumed transition probabilities  $\{M_{s,t}^{(l)}\}$ . The proof of this proposition can be found in Appendix E.4.

## 4 CLTs for the Value, Q-, and Advantage Functions, and the Optimal Policy Value

We now demonstrate how Theorem 1 and Corollary 1 extend naturally to downstream tasks in offline RL, such as *offline policy evaluation* (OPE) and *offline policy recovery* (OPR). In OPE, one seeks to estimate the value of a fixed target policy from logged trajectories

generated under an unknown behavior policy, whereas in OPR, one aims to identify the optimal policy that maximizes the expected return given an estimated transition model. Both problems rely on accurate estimation of the value, Q-, and advantage functions, for which we now establish joint asymptotic normality results.

Let  $\Delta(\mathcal{A})$  denote the probability simplex over the action space  $\mathcal{A}$ , and let  $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$  be a stationary target policy. In offline RL, the (known) target policy is usually different from the (unknown) logging policy under which we obtain the logged data  $\{(X_0, a_0), \dots, (X_n, a_n)\}$ .

For each state  $s$ , denote  $\pi_s = [\pi(s, 1), \dots, \pi(s, k)]$  and define  $\Pi = \text{diag}(\pi_1, \dots, \pi_d)$  as a  $d \times dk$  block-diagonal matrix. Let  $\tilde{g} := (\tilde{g}(x, a) : (x, a) \in \mathcal{S} \times \mathcal{A})$  denote the per-state-action reward vector, and define the per-state expected reward

$$g(x) = \sum_{a \in \mathcal{A}} \pi(x, a) \tilde{g}(x, a), \quad g = (g(x) : x \in \mathcal{S}) \in \mathbb{R}^d.$$

For a discount factor  $0 < \alpha < 1$ , the value function is defined by

$$V_\Pi = (I - \alpha \Pi \mathbf{M})^{-1} g,$$

and its plug-in estimate  $\hat{V}_\Pi = (I - \alpha \hat{\Pi} \mathbf{M})^{-1} g$  follows from the Bellman equation (Bertsekas, 2011).

Similarly, let  $\tilde{r} := (\tilde{r}(x, a, y) : (x, a, y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S})$  denote the immediate reward on transition from  $x$  to  $y$  under action  $a$ , and define

$$r(x, a) = \sum_{y \in \mathcal{S}} M_{x,y}^{(a)} \tilde{r}(x, a, y), \quad r = (r(x, a) : (x, a) \in \mathcal{S} \times \mathcal{A}) \in \mathbb{R}^{dk}.$$

Then the Q-function satisfies (Agarwal et al., 2019)

$$Q_\Pi = (I - \alpha \mathbf{M} \Pi)^{-1} r, \quad \hat{Q}_\Pi = (I - \alpha \hat{\mathbf{M}} \Pi)^{-1} r.$$

Finally, the advantage function  $A_\Pi = Q_\Pi - K V_\Pi$ , with  $K = \text{diag}(\mathbf{1}_k, \dots, \mathbf{1}_k)$ , admits plug-in estimate  $\hat{A}_\Pi = \hat{Q}_\Pi - K \hat{V}_\Pi$ .

For compactness, we define the following matrices corresponding to the three functions:

$$\begin{aligned} \Sigma_V^\Pi &= \alpha^2 \left[ (I - \alpha \Pi \mathbf{M})^{-1} \Pi \otimes V_\Pi^\top \right] \bar{\Lambda} \left[ \Pi^\top (I - \alpha \mathbf{M}^\top \Pi^\top)^{-1} \otimes V_\Pi \right], \\ \Sigma_Q^\Pi &= \alpha^2 \left[ (I - \alpha \mathbf{M} \Pi)^{-1} \otimes Q_\Pi^\top \Pi^\top \right] \bar{\Lambda} \left[ (I - \alpha \Pi^\top \mathbf{M}^\top)^{-1} \otimes \Pi Q_\Pi \right], \\ \Sigma_A^\Pi &= \alpha^2 \left[ (I - \alpha \mathbf{M} \Pi)^{-1} \otimes Q_\Pi^\top \Pi^\top - K \left( (I - \alpha \Pi \mathbf{M})^{-1} \Pi \otimes V_\Pi^\top \right) \right] \bar{\Lambda} \\ &\quad \cdot \left[ (I - \alpha \Pi^\top \mathbf{M}^\top)^{-1} \otimes \Pi Q_\Pi - (\Pi^\top (I - \alpha \mathbf{M}^\top \Pi^\top)^{-1} \otimes V_\Pi) K^\top \right], \end{aligned}$$

where  $\bar{\Lambda}$  is the improperly scaled covariance matrix in Corollary 1.

The following theorem shows that  $\Sigma_V^{(\Pi)}$ ,  $\Sigma_Q^{(\Pi)}$  and  $\Sigma_A^{(\Pi)}$  are the asymptotic covariance matrices for the scaled estimation errors  $\sqrt{n} (\hat{V}_\Pi - V_\Pi)$ ,  $\sqrt{n} (\hat{Q}_\Pi - Q_\Pi)$ ,  $\sqrt{n} (\hat{A}_\Pi - A_\Pi)$ , respectively.

**Theorem 2 (CLTs for  $V_\Pi$ ,  $Q_\Pi$ , and  $A_\Pi$ )** *Let  $\{(X_0, a_0), \dots, (X_n, a_n)\}$  be a sample from a controlled Markov chain under a logging policy. Suppose that Assumptions 1, 2, and 5 hold for the logging policy, and let  $\pi$  be a stationary target policy. Then, as  $n \rightarrow \infty$ ,*

$$\begin{aligned}\sqrt{n}(\hat{V}_\Pi - V_\Pi) &\xrightarrow{d} \mathcal{N}(0, \Sigma_V^\Pi), \\ \sqrt{n}(\hat{Q}_\Pi - Q_\Pi) &\xrightarrow{d} \mathcal{N}(0, \Sigma_Q^\Pi), \\ \sqrt{n}(\hat{A}_\Pi - A_\Pi) &\xrightarrow{d} \mathcal{N}(0, \Sigma_A^\Pi),\end{aligned}$$

and each estimator is consistent:  $\hat{V}_\Pi \xrightarrow{p} V_\Pi$ ,  $\hat{Q}_\Pi \xrightarrow{p} Q_\Pi$ , and  $\hat{A}_\Pi \xrightarrow{p} A_\Pi$ .

The proof of this theorem can be found in Appendix D.1. Observe that Theorem 2 derives the asymptotic distribution for the estimated  $V/Q/A$  functions. In Theorem 3, we produce the asymptotic distribution of the optimal value function. Similarly, we derive the CLT for the  $Q$ -function.

**Corollary 2 (Properly scaled CLT for  $Q_\Pi$ )** *Suppose that Assumptions 1, 2, and 5 hold for the logging policy, and let  $\pi$  be a stationary target policy, then the properly scaled  $Q$ -estimation error satisfies*

$$(\hat{Q}_\Pi - Q_\Pi) \odot \text{Vec}(\mathbf{N}^\top) \xrightarrow{d} \mathcal{N}(0, \Lambda_Q^\Pi),$$

where the  $(sl, s'l')$ -th element of  $\Lambda_Q^\Pi$  equals  $\lambda_{sl, s'l'}^{Q_\Pi} = \sqrt{p_s^{(l)} p_{s'}^{(l')}} \sigma_{sl, s'l'}^{Q_\Pi}$ , with  $\sigma_{sl, s'l'}^{Q_\Pi}$  denoting the  $(sl, s'l')$ -th element of  $\Sigma_Q^\Pi$ .

The proof of this corollary can be found in Appendix D.2.

Although the  $Q$ -estimation error is scaled by the visitation counts, this “correct” scaling does not remove dependence on the ergodic occupation measure in the asymptotic covariance. The reason is structural rather than technical. The count-based scaling normalizes the local estimation noise of each transition probability, rendering the errors asymptotically multinomial. However, the  $Q$ -function is a global functional of the transition kernel: estimation errors from different state–action pairs are propagated and aggregated through repeated applications of the Bellman operator. As a result, the asymptotic variance depends on the long-run visitation frequency of state-action pairs under the logging policy.

The optimal policy can be found, for instance, by policy iteration (Bertsekas, 2011, Chapter 1) or policy gradient (Sutton and Barto, 2018, Chapter 13). For any policy  $\pi$  let  $\Pi$  be its block diagonal matrix and let  $\Pi_{opt}$  and  $\hat{\Pi}_{opt}$  be the block-diagonal matrices associated with the optimal policies  $\pi_{opt}$  and  $\hat{\pi}_{opt}$ , each maximizing the expected reward under the true and estimated transition matrices  $\mathbf{M}$  and  $\hat{\mathbf{M}}$ , respectively. The following theorem characterizes the asymptotic behavior of plug-in estimator of the value function for the estimated optimal policy.

**Theorem 3** *Let us assume that for any  $\varepsilon > 0$ , and policy  $\pi'$  such that  $\|\Pi_{opt} - \Pi'\|_\infty > \varepsilon$ , the value functions corresponding to  $\pi_{opt}$  and  $\pi'$  satisfy  $\inf_{s \in \mathcal{S}} \{V_{\Pi_{opt}}(s) - V_{\Pi'}(s)\} > 0$ . Then  $\hat{\Pi}_{opt} \xrightarrow{p} \Pi_{opt}$  and,*

$$\sqrt{n}(\hat{V}_{\hat{\Pi}_{opt}} - V_{\Pi_{opt}}) \xrightarrow{d} \mathcal{N}(0, \Sigma_V^{(\Pi_{opt})}).$$

The proof of this theorem can be found in Appendix D.3.

**Implications for OPE and OPR.** Taken together, Theorems 2 and 3 yield an asymptotic framework for model-based offline reinforcement learning. We now formalize how the established CLTs can be used to construct asymptotically valid confidence intervals for OPE and OPR.

Without loss of generality, we consider inference for the value function. Given logged data  $\{(X_i, a_i)\}_{i=0}^{n-1}$  generated under a possibly history-dependent logging policy, and a fixed stationary target policy  $\pi$ , the goal of OPE is to infer the value function  $V_\Pi$ . By Theorem 2, the scaled estimation error of  $V_\Pi$  is jointly asymptotically normal:

$$\sqrt{n}(\hat{V}_\Pi - V_\Pi) \xrightarrow{d} \mathcal{N}(0, \Sigma_V^\Pi).$$

Next, let  $\hat{\Sigma}_V^\Pi$  denote the plug-in estimator of  $\Sigma_V^\Pi$ . Then

$$\hat{\Sigma}_V^\Pi = \alpha^2 \left[ (I - \alpha \Pi \hat{\mathbf{M}})^{-1} \Pi \otimes \hat{V}_\Pi^\top \right] \hat{\Lambda} \left[ \Pi^\top (I - \alpha \hat{\mathbf{M}}^\top \Pi^\top)^{-1} \otimes \hat{V}_\Pi \right],$$

where the elements of  $\hat{\Lambda}$  are given by

$$\hat{\lambda}_{slt, s'l't'} = \frac{\mathbb{1}[(s, l) = (s', l')] \left( \mathbb{1}[t = t'] \hat{M}_{s,t}^{(l)} - \hat{M}_{s,t}^{(l)} \hat{M}_{s',t'}^{(l')} \right)}{\sqrt{\hat{p}_s^{(l)} \hat{p}_{s'}^{(l')}}}, \quad \hat{p}_s^{(l)} = \frac{N_s^{(l)}}{n}.$$

For a target level  $1 - \delta$ , the confidence set

$$\text{CS}_{V_\Pi}^\delta = \left\{ v \in \mathbb{R}^d : n(\hat{V}_\Pi - v)^\top (\hat{\Sigma}_V^\Pi)^{-1} (\hat{V}_\Pi - v) \leq \chi_{1-\delta}^2(d) \right\}$$

is an asymptotically valid  $100(1 - \delta)\%$  confidence set for  $V_\Pi$ , where  $\chi_{1-\delta}^2(d)$  is the  $(1 - \delta)$ -quantile of the chi-squared distribution with  $d = |S|$  degrees of freedom.

Let  $\sigma_s^2$  denote the marginal variance of the  $s$ -th scaled estimation error  $\sqrt{n}(\hat{V}_\Pi(s) - V_\Pi(s))$ , the  $(s, s)$ -th entry of  $\Sigma_V^\Pi$  and let  $\hat{\sigma}_s^2$  denote its plug-in estimator, the  $(s, s)$ -th entry of  $\hat{\Sigma}_V^\Pi$ . Then an approximate two-sided  $100(1 - \delta)\%$  confidence interval for  $V_\Pi(s)$  is

$$\hat{V}_\Pi(s) \pm z_{1-\delta/2} \sqrt{\frac{\hat{\sigma}_s^2}{n}}, \quad s \in \mathcal{S},$$

where  $z_{1-\delta/2}$  is the  $(1 - \delta/2)$ -quantile of the standard normal distribution. The derivation of the confidence sets and intervals for  $Q_\Pi$  and  $A_\Pi$  are analogous.

Both the ellipsoidal confidence sets and the marginal confidence intervals above are first-order asymptotically valid. A systematic investigation of their higher-order accuracy and finite-sample performance—via Edgeworth expansions and potential improvements using model-based bootstrap—makes an interesting problem for future work.

## 5 Conclusion

We develop the first asymptotic theory for *non-parametric* estimation in finite CMCs. Our main result (Theorem 1) establishes a central limit theorem for the classical count-based estimator of state-action transition probabilities. This non-trivially extends prior results to

accommodate non-Markovian and non-stationary data, which renders classical regenerative techniques inapplicable. Our proof introduces a new auxiliary sampling scheme that decouples temporal dependence while preserving the finite-state controlled dynamics, thereby extending the sampling construction of Billingsley (1961) to the controlled and time-inhomogeneous setting.

Building on this foundation, we show that plug-in estimators of the value,  $Q$ -, and advantage functions under arbitrary stationary target policies—as well as the value of the estimated optimal policy—also satisfy CLTs. All results hold under general, possibly history-dependent logging policies, providing a unified large-sample framework for model-based offline policy evaluation (OPE) and optimal policy recovery (OPR) in reinforcement learning.

Several directions for future work are possible. Analyzing the higher-order accuracy and finite-sample performance of the confidence sets and intervals for OPE and OPR derived from our asymptotic results would provide a deeper understanding of their practical reliability in offline reinforcement learning. Refining the testability criteria to close the gap between the current sufficiency results and necessary impossibility statements would delineate precisely when asymptotic inference is attainable. Finally, extending the theory to infinite (Kontorovich et al., 2008) or continuous state–action spaces, as suggested by the adaptive density estimation work in Markov settings (Sart, 2014), and to continuous-time controlled processes, would markedly broaden the scope of rigorous statistical inference of controlled Markov chains and its applications in reinforcement learning.

## Appendix A. Examples of Controlled Markov Chains

### A.1 Inhomogeneous Markov Chains

We consider an inhomogeneous Markov chain as the first example. A controlled Markov chain is said to be an inhomogeneous Markov chain if there exists a sequence of actions  $l_0, l_1, \dots \in \mathcal{A}$  such that for any time  $i$ , state  $s$ , sample history  $\mathcal{H}_0^{i-1}$ , we have

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

We make the following assumption on the logging policy.

#### Assumption 6

1. Let  $T_i = O(i^\alpha)$ ,  $\alpha \in (0, 1)$  be a sequence of known integers. Then, we assume that

$$\sum_{p=j}^{T_i+j} \mathbb{1}[a_p = l] \geq 1, \quad (\text{A.1})$$

for all  $\tau_{s,l}^{(i-1)} + 1 \leq j \leq \tau_{s,l}^{(i)} - T_i$  and every  $l \in \mathcal{A}$ .

2. There exists  $M_{min}$  and  $M_{max}$  such that for all  $s, t \in \mathcal{S}$  and  $l \in \mathcal{A}$ ,

$$0 < M_{min} \leq M_{s,t}^{(l)} \leq M_{max} < 1. \quad (\text{A.2})$$

**Proposition 5** Let  $\{(X_0, a_0), \dots, (X_n, a_n)\}$  be a sample from an inhomogeneous Markov chain satisfying Assumption 6. Then Theorem 1 holds with

$$C = 0, \quad C_\theta = e/(e-1).$$

**Proof** We begin our proof by verifying Assumption 1. For any  $k > i$ , (A.2) implies that

$$\begin{aligned} \mathbb{P}\left(\tau_{s,l}^{(i)} > k | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}}\right) &= \mathbb{P}\left(\left\{X_j \neq s \cup a_j \neq l\right\} \forall j \in \{i+1, \dots, k+i\} | X_i = s, a_i = l\right) \\ &\leq (1 - M_{min})^{\lfloor \frac{k}{T_i} \rfloor} \\ &\leq (1 - M_{min})^{\frac{k}{T_i} - 1}. \end{aligned}$$

Thus,

$$\mathbb{E}[\tau_{s,l}^{(i)} | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}}] = \sum_{k=1}^{\infty} \mathbb{P}\left(\tau_{s,l}^{(i)} > k | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}}\right) \leq \frac{(1 - M_{min})^{\frac{1}{T_i} - 1}}{\left(1 - (1 - M_{min})^{\frac{1}{T_i}}\right)}. \quad (\text{A.3})$$

As  $T_i = O(i^\alpha)$ , there exists a constant  $c > 0$  such that  $T_i \leq ci^\alpha$  for sufficiently large  $i$ .

Let  $q := 1 - M_{min}$ ,  $\tilde{T}_i := \log q / \log(ci^\alpha / ci^\alpha - \log q)$ . Then

$$\begin{aligned} \tilde{T}_i &= \frac{\log q}{\log\left(\frac{ci^\alpha}{ci^\alpha - \log q}\right)} \\ &= -\frac{\log q}{\log\left(1 - \frac{\log q}{ci^\alpha}\right)} \end{aligned} \quad (\text{A.4})$$

By the Taylor expansion of  $\log(1+x)$ ,  $\log(1+x) = x - o(x)$  as  $x \rightarrow 0^+$ . Then for sufficiently large  $i$ , the right-hand side of (A.4) becomes

$$\begin{aligned}\tilde{T}_i &= -\frac{\log q}{-\frac{\log q}{ci^\alpha} - o(1/i^\alpha)} \\ &= \frac{-ci^\alpha \log q}{-\log q - o(1)} \\ &\geq \frac{-ci^\alpha \log q}{-\log q} \\ &= ci^\alpha = T_i.\end{aligned}$$

Thus,  $-1/\tilde{T}_i \geq -1/T_i$ , and  $q^{-1/\tilde{T}_i} \leq q^{-1/T_i}$ . Therefore,

$$\frac{q^{\frac{1}{\tilde{T}_i}-1}}{1 - q^{\frac{1}{\tilde{T}_i}}} = \frac{q^{-1}}{q^{-\frac{1}{\tilde{T}_i}} - 1} \leq \frac{q^{-1}}{q^{-\frac{1}{T_i}} - 1} = \frac{q^{\frac{1}{T_i}-1}}{1 - q^{\frac{1}{T_i}}}.$$

Substituting this value into the right hand side of (A.3), we get

$$\begin{aligned}\mathbb{E}[\tau_{s,l}^{(i)}] &\leq \frac{q^{\frac{1}{\tilde{T}_i}-1}}{1 - q^{\frac{1}{\tilde{T}_i}}} \\ &= \frac{q^{\frac{\log(\frac{ci^\alpha}{ci^\alpha - \log q})}{\log q} - 1}}{1 - q^{\frac{\log(\frac{ci^\alpha}{ci^\alpha - \log q})}{\log q}}} \\ &= \frac{\frac{ci^\alpha}{ci^\alpha - \log q}}{q - \frac{qci^\alpha}{ci^\alpha - \log q}} \\ &= \frac{c}{-q \log q} i^\alpha.\end{aligned}$$

Hence, Assumption 1 holds.

Now we verify Assumptions 3 and 4. As the controls are deterministic, the total variation distance

$$\begin{aligned}\gamma_{p,j,i} &= \sup_{s_p, \bar{h}_{i+j}^{p-1}, \bar{h}_0^i} \left\| \mathbb{P}\left(a_p \mid X_p = s_p, \mathcal{H}_{i+j}^{p-1} = \bar{h}_{i+j}^{p-1}, \mathcal{H}_0^i = \bar{h}_0^i\right) - \mathbb{P}\left(a_p \mid X_p = s_p, \mathcal{H}_{i+j}^{p-1} = \bar{h}_{i+j}^{p-1}\right) \right\|_{TV} \\ &\equiv 0.\end{aligned}$$

Therefore, Assumption 3 holds for all inhomogeneous Markov chains with  $C = 0$ . Observe from the definition of  $\bar{\theta}_{i,j}$  in (2.6) that,

$$\bar{\theta}_{i,j} = \sup_{l, l', s_1, s_2} \left\| \mathbb{P}(X_j \mid X_i = s_1, a_i = l) - \mathbb{P}(X_j \mid X_i = s_2, a_i = l') \right\|_{TV}.$$

From the definition of an inhomogenous Markov chains it follows that

$$\begin{aligned} & \sup_{s_1, s_2} \left\| \mathbb{P}(X_j | X_i = s_1, a_i = l_1) - \mathbb{P}(X_j | X_i = s_2, a_i = l_2) \right\|_{TV} \\ &= \sup_{s_1, s_2} \left\| \mathbb{P}(X_j | X_i = s_1, (a_{j-1}, \dots, a_i) = (l_{j-1}, \dots, l_i)) \right. \\ & \quad \left. - \mathbb{P}(X_j | X_i = s_2, (a_{j-1}, \dots, a_i) = (l_{j-1}, \dots, l_i)) \right\|_{TV} \end{aligned}$$

where the last line follows due the controls being deterministic.

Since all transition matrices are positive, an application of Wolfowitz (1963, Theorem 1) implies that there exists an integer  $C$  for which,

$$\bar{\theta}_{i,j} \leq e^{-C(j-i)}.$$

Since  $e^{-C(j-i)} < e^{-(j-i)}$  for any integer  $C$ , it consequently implies that,

$$\sup_{i \geq 1} \sum_{j > i} \bar{\theta}_{i,j} \leq \frac{e^{-1}}{1 - e^{-1}} \leq \frac{1}{1 - e^{-1}} = \frac{e}{e - 1}.$$

Therefore, Assumption 4 holds with  $C_\theta = e/e-1$ . Then, by Banerjee et al. (2025a, Lemma 3), Assumption 2 holds. As both Assumptions 1 and 2 hold, Theorem 1 holds. This completes the proof.  $\blacksquare$

## A.2 Controlled Markov Chains with Non-Stationary Markov Controls

We consider a controlled Markov chain with non-stationary control process as the second example. A controlled Markov chain is said to have non-stationary Markov controls if for any time  $i$ , state  $s$ , action  $l$ , and sample history  $\bar{h}_0^{i-1}$ ,

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

We observe that an action can depend on the time  $i$ . Let  $P_{s,l}^{(i)} := \mathbb{P}(a_i = l | X_i = s)$ . We can write the transition probability of the state-action pair as

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

It is straightforward to see that the state-action pair is a time inhomogeneous Markov chain with transition probabilities given by  $M_{s,t}^{(l)} P_{s,l'}^{(i)}$ . The goal is to perform statistical inference on 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 actions.

**Definition 4** We define  $\tau_{s,l}^{(i,*,j)}$  to be the time between the  $(j-1)$ -th and  $j$ -th visit to action  $l$  after visiting state-action pair  $s, l$  for the  $i$ -th time. For ease of notation, let  $\sum_{k=1}^i \tau_{s,l}^{(k)} + \sum_{k=1}^{j-1} \tau_{s,l}^{(i,*,k)} = \tau_*$ . We can then define  $\tau_{s,l}^{(i,*,j)}$  recursively as  $\tau_{s,l}^{(i,*,j)} := \min\{n > 0 : a_{\tau_*+n} = l, a_j \neq l, \forall \tau_* < j < \tau_* + n\}$ .

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

**Assumption 7**

1. For all state-action pairs  $(s, l) \in \mathcal{S} \times \mathcal{A}$ , there exists  $\alpha \in (0, 1)$  such that

$$\mathbb{E}[\tau_{s,l}^{(i,*,j)} | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)} + \sum_{p=1}^{j-1} \tau_{s,l}^{(i,*,p)}}] \leq T_i^* \text{ almost surely, with } T_i^* = O(i^\alpha).$$

2. There exists  $M_{min}$  and  $M_{max}$  such that for all  $s, t \in \mathcal{S}$  and  $l \in \mathcal{A}$ ,

$$0 < M_{min} \leq M_{s,t}^{(l)} \leq M_{max} < 1. \quad (\text{A.5})$$

The next lemma proves that under this assumption  $\{(X_i, a_i)\}$  satisfies Assumption 1.

**Lemma 1** Under the conditions of Assumption 7, for all  $(i, s, l) \in \mathbb{N} \times \mathcal{S} \times \mathcal{A}$ , it holds almost surely that

$$\mathbb{E}[\tau_{s,l}^{(i)} | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}}] \leq \frac{T_i^* M_{max}}{(1 - \max\{M_{max}, 1 - M_{min}\})^2}. \quad (\text{A.6})$$

**Proof** We begin by observing that

$$\tau_{s,l}^{(i+1)} = \begin{cases} \tau_{s,l}^{(i,*,1)} & \text{if } X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \tau_{s,l}^{(i,*,1)}} = s, \\ \sum_{k=1}^j \tau_{s,l}^{(i,*,k)} & \text{if } X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s, \quad \forall 1 \leq m \leq j \\ & \text{and } X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^j \tau_{s,l}^{(i,*,k)}} = s. \end{cases}$$

Therefore,

$$\begin{aligned} \tau_{s,l}^{(i+1)} &= \tau_{s,l}^{(i,*,1)} \mathbb{1} \left[ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \tau_{s,l}^{(i,*,1)}} = s \right] \\ &+ \sum_{j=1}^{\infty} \sum_{k=1}^j \tau_{s,l}^{(i,*,k)} \mathbb{1} \left[ \begin{array}{l} X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s, \quad \forall m \in \{1, \dots, j\}, \\ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^j \tau_{s,l}^{(i,*,k)}} = s \end{array} \right], \end{aligned}$$

which in turn implies that

$$\begin{aligned} &\mathbb{E}[\tau_{s,l}^{(i+1)} | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}}] \\ &= \mathbb{E} \left[ \tau_{s,l}^{(i,*,1)} \mathbb{1} \left[ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \tau_{s,l}^{(i,*,1)}} = s \right] \mid \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] \\ &+ \sum_{j=1}^{\infty} \sum_{k=1}^j \mathbb{E} \left[ \tau_{s,l}^{(i,*,k)} \mathbb{1} \left[ \begin{array}{l} X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s, \quad \forall m \in \{1, \dots, j\}, \\ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^j \tau_{s,l}^{(i,*,k)}} = s \end{array} \right] \mid \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] \\ &= \text{Term 1} + \text{Term 2}. \quad (\text{A.7}) \end{aligned}$$

We now analyze the terms one by one.

**Term 1:** The law of conditional expectation implies that

$$\begin{aligned}
 & \mathbb{E} \left[ \tau_{s,l}^{(i,*,1)} \mathbb{1} \left[ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \tau_{s,l}^{(i,*,1)}} = s \right] \middle| \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] \\
 &= \mathbb{E} \left[ \mathbb{E} \left[ \tau_{s,l}^{(i,*,1)} \mathbb{1} \left[ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \tau_{s,l}^{(i,*,1)}} = s \right] \middle| \tau_{s,l}^{(i,*,1)} \right] \middle| \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] \\
 &= \mathbb{E} \left[ \tau_{s,l}^{(i,*,1)} \mathbb{P} \left( X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \tau_{s,l}^{(i,*,1)}} = s \middle| \tau_{s,l}^{(i,*,1)} \right) \middle| \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right]. \tag{A.8}
 \end{aligned}$$

Recall from (A.5) that

$$\max_{s,t,l} M_{s,t}^{(l)} = M_{max}, \quad \text{and} \quad \min_{s,t,l} M_{s,t}^{(l)} = M_{min}$$

for two numbers  $0 < M_{min}, M_{max} < 1$ . Then for any time  $p$ , state  $s$ , and history  $\mathfrak{h}_0^{p-1}$ ,

$$M_{min} \leq \mathbb{P} \left( X_p = s \mid \mathcal{H}_0^{p-1} = \mathfrak{h}_0^{p-1} \right) \leq M_{max}, \quad \text{and} \tag{A.9}$$

$$\mathbb{P} \left( X_p \neq s \mid \mathcal{H}_0^{p-1} = \mathfrak{h}_0^{p-1} \right) \leq M_{opt} := \max \{ M_{max}, 1 - M_{min} \}. \tag{A.10}$$

It follows from (A.9) that  $\mathbb{P} \left( X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \tau_{s,l}^{(i,*,1)}} = s \middle| \tau_{s,l}^{(i,*,1)} \right) \leq M_{max}$ . Substituting this value in the right hand side of (A.8), we get the following upper bound to Term 1:

$$\mathbb{E} \left[ \tau_{s,l}^{(i,*,1)} \mathbb{1} \left[ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \tau_{s,l}^{(i,*,1)}} = s \right] \middle| \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] \leq \mathbb{E} \left[ \tau_{s,l}^{(i,*,1)} M_{max} \middle| \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] \leq T_i^* M_{max},$$

where the last inequality follows from the tower property.

**Term 2:** We define for ease of notation

$$\mathbb{E}^*[\cdot] = \mathbb{E}[\cdot \middle| \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}}],$$

and proceed similarly as before to get

$$\begin{aligned}
 & \mathbb{E}^* \left[ \tau_{s,l}^{(i,*,k)} \mathbb{1} \left[ \begin{array}{l} X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s, \quad \forall m \in \{1, \dots, j\}, \\ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^j \tau_{s,l}^{(i,*,k)}} = s \end{array} \right] \right] \\
 &= \mathbb{E}^* \left[ \tau_{s,l}^{(i,*,k)} \mathbb{P} \left( \begin{array}{l} X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s, \quad \forall m \in \{1, \dots, j\}, \\ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^j \tau_{s,l}^{(i,*,k)}} = s \end{array} \middle| \tau_{s,l}^{(i,*,k)}, k = 1, \dots, j \right) \right]. \tag{A.11}
 \end{aligned}$$

The Bayes' theorem gives

$$\begin{aligned}
 & \mathbb{P} \left( \begin{array}{l} X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s, \quad \forall m \in \{1, \dots, j\}, \\ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^j \tau_{s,l}^{(i,*,k)}} = s \end{array} \middle| \tau_{s,l}^{(i,*,k)}, k = 1, \dots, j \right) \\
 &= \mathbb{P} \left( \begin{array}{l} X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^j \tau_{s,l}^{(i,*,k)}} = s \middle| \tau_{s,l}^{(i,*,k)}, k = 1, \dots, j, \\ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s, \quad \forall m \in \{1, \dots, j\} \end{array} \right) \\
 & \quad \times \mathbb{P} \left( X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s \quad \forall m \in \{1, \dots, j\} \middle| \tau_{s,l}^{(i,*,k)}, k = 1, \dots, j \right) \\
 &= \mathbb{P} \left( \begin{array}{l} X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^j \tau_{s,l}^{(i,*,k)}} = s \middle| \tau_{s,l}^{(i,*,k)}, k = 1, \dots, j, \\ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s, \quad \forall m \in \{1, \dots, j\} \end{array} \right) \\
 & \quad \times \prod_{m=1}^{j-1} \mathbb{P} \left( X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s \middle| \tau_{s,l}^{(i,*,k)}, k = 1, \dots, j \right). \tag{A.12}
 \end{aligned}$$

Using (A.9) and (A.10) to bound the first and the second term from (A.12) we get

$$\mathbb{P} \left( \begin{array}{l} X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s, \quad \forall m \in \{1, \dots, j\} \text{ and} \\ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^j \tau_{s,l}^{(i,*,k)}} = s \end{array} \middle| \tau_{s,l}^{(i,*,k)}, k = 1, \dots, j \right) \leq M_{max} M_{opt}^{j-1}.$$

Substituting this value in the right hand side of (A.11), we get

$$\begin{aligned}
 & \mathbb{E}^* \left[ \tau_{s,l}^{(i,*,k)} \mathbb{1} \left[ X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^m \tau_{s,l}^{(i,*,k)}} \neq s, \quad \forall m \in \{1, \dots, j\} \text{ and } X_{\sum_{p=1}^i \tau_{s,l}^{(p)} + \sum_{k=1}^j \tau_{s,l}^{(i,*,k)}} = s \right] \right] \\
 & \leq \mathbb{E}^* [\tau_{s,l}^{(i,*,k)}] M_{max} M_{opt}^{j-1} \\
 & \leq T_{\star}^i M_{max} M_{opt}^{j-1},
 \end{aligned}$$

where the last inequality follows from the tower property. Substituting the obtained upper bounds of Term 1 and Term 2 back into (A.7), we get

$$\begin{aligned}
 \mathbb{E} \left[ \tau_{s,l}^{(i)} \middle| \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] & \leq \sum_{j=1}^{\infty} j T_{\star}^i M_{max} M_{opt}^{j-1} \\
 & = \frac{T_{\star}^i M_{max}}{1 - M_{opt}} \sum_{j=1}^{\infty} j (1 - M_{opt}) M_{opt}^{j-1} \\
 & = \frac{T_{\star}^i M_{max}}{(1 - M_{opt})^2}.
 \end{aligned}$$

■

We can now state our main result about the CLT of a controlled Markov chain with a non-stationary Markov controls.

**Proposition 6** *Let  $\{(X_0, a_0), \dots, (X_n, a_n)\}$  be a sample from a controlled Markov chain with non-stationary Markovian controls satisfying Assumption 7, then Theorem 1 holds with*

$$C = 0, \quad C_\theta = 1/(dM_{\min}).$$

**Proof** As Assumption 7 holds, Lemma 1 holds. Then Assumption 1 holds. We just need to verify Assumptions 3 and 4.

As the controls are non-stationary Markov, the total variation distance

$$\begin{aligned} \gamma_{p,j,i} &= \sup_{s_p, h_{i+j}^{p-1}, h_0^i} \left\| \mathbb{P} \left( a_p \mid X_p = s_p, \mathcal{H}_{i+j}^{p-1} = h_{i+j}^{p-1}, \mathcal{H}_0^i = h_0^i \right) - \mathbb{P} \left( a_p \mid X_p = s_p, \mathcal{H}_{i+j}^{p-1} = h_{i+j}^{p-1} \right) \right\|_{TV} \\ &\equiv 0. \end{aligned}$$

Therefore, Assumption 3 holds for all controlled markov chains with non-stationary controls with  $C = 0$ .

Recall from (A.5),  $M_{\min} = \min_{s,t,l} M_{s,t}^{(l)} > 0$ . Then, by Banerjee et al. (2025a, Lemma 12), Assumption 4 holds with  $C_\theta = 1/(dM_{\min})$ .  $\blacksquare$

## Appendix B. Statements and Proofs of the Propositions and Lemmas Used in Theorem 1

### B.1 Statement and Proof of Lemma 2

**Lemma 2** *Let  $\{(X_0, a_0), \dots, (X_n, a_n)\}$  be a sample from a controlled Markov chain. Under Assumptions 1 and 2,*

$$\frac{N_s^{(l)}}{\mathbb{E}[N_s^{(l)}]} \xrightarrow{a.s.} 1.$$

*Further suppose that Assumption 5 holds, then  $N_s^{(l)}$  is 1-st order almost surely consistent for  $np_s^{(l)}$ . In other words,*

$$\frac{N_s^{(l)}}{n} \xrightarrow{a.s.} p_s^{(l)}.$$

**Proof** Using Banerjee et al. (2025a, Lemma 6), for any  $t > 0$ , we have:

$$\mathbb{P}(|N_s^{(l)} - \mathbb{E}[N_s^{(l)}]| > t) \leq 2 \exp\left(-\frac{t^2}{2n\|\Delta_n\|^2}\right), \quad (\text{B.1})$$

where  $\|\Delta_n\| = \max_{1 \leq i \leq n} (1 + \bar{\eta}_{i,i+1} + \bar{\eta}_{i,i+2} + \dots + \bar{\eta}_{i,n})$ , and  $\bar{\eta}_{i,j}$  are the coefficients. Setting  $t = \varepsilon \mathbb{E}[N_s^{(l)}]$  we get

$$\mathbb{P}\left(\left|\frac{N_s^{(l)}}{\mathbb{E}[N_s^{(l)}]} - 1\right| > \varepsilon\right) \leq 2 \exp\left(-\frac{\mathbb{E}[N_s^{(l)}]^2 \varepsilon^2}{2n\|\Delta_n\|^2}\right). \quad (\text{B.2})$$

By Assumption 2,  $\sup_n \|\Delta_n\| < \infty$ . As Assumption 1 holds, it follows from Proposition 2 that  $\mathbb{E}[N_s^{(l)}] \geq cn^{1/1+\alpha}$  for sufficiently large  $n$ , where  $c > 0$  is a constant, and we get  $\mathbb{P}\left(\left|N_s^{(l)}/\mathbb{E}[N_s^{(l)}] - 1\right| > \varepsilon\right) \leq 2 \exp\left(-c^2\varepsilon^2n^{(1-\alpha)/(1+\alpha)}/2\|\Delta_n\|^2\right)$ .

As  $\alpha \in (0, 1)$ , the series  $\sum_{n=1}^{\infty} 2 \exp\left(-c^2\varepsilon^2n^{(1-\alpha)/(1+\alpha)}/2\|\Delta_n\|^2\right)$  converges. Observing that  $\varepsilon$  is arbitrary, using the Borel-Cantelli lemma, we get  $\left|N_s^{(l)}/\mathbb{E}[N_s^{(l)}] - 1\right| \xrightarrow{a.s.} 0$ , hence  $N_s^{(l)}/\mathbb{E}[N_s^{(l)}] \xrightarrow{a.s.} 1$ .

Using the Slutsky's theorem, we have  $N_s^{(l)}/n \xrightarrow{a.s.} p_s^{(l)}$  under Assumption 5. This completes the proof.  $\blacksquare$

## B.2 Statement and Proof of Proposition 7

**Proposition 7**  $(X_0, a_0, \dots, X_n, a_n)$  is identically distributed to  $(\tilde{X}_0, \tilde{a}_0, \dots, \tilde{X}_n, \tilde{a}_n)$ .

**Proof** We prove this fact by induction. Obviously,  $(X_0, a_0) \stackrel{d}{=} (\tilde{X}_0, \tilde{a}_0)$ . Now, for some  $i \geq 1$ , let  $(X_0, a_0, \dots, X_i, a_i)$  be identically distributed to  $(\tilde{X}_0, \tilde{a}_0, \dots, \tilde{X}_i, \tilde{a}_i)$ . Then, for  $i + 1$ , we note that,

$$\begin{aligned} & \mathbb{P}\left(\tilde{X}_{i+1} = s_{i+1}, \tilde{a}_{i+1} = l_{i+1}, \dots, \tilde{X}_0 = s_0, \tilde{a}_0 = l_0\right) \\ &= \mathbb{P}\left(\tilde{a}_{i+1} = l_{i+1} | \tilde{X}_{i+1} = s_{i+1}, \dots, \tilde{X}_0 = s_0\right) \\ & \quad \times \mathbb{P}\left(\tilde{X}_{i+1} = s_{i+1} | \tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{a}_0 = l_0, \tilde{X}_0 = s_0\right) \\ & \quad \times \mathbb{P}(\tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{X}_0 = s_0, \tilde{a}_0 = l_0) \\ &= \mathbb{P}\left(\alpha_i^{(\tilde{X}_0, \tilde{a}_0, \dots, \tilde{X}_{i+1})} = l_{i+1} | \tilde{X}_{i+1} = s_{i+1}, \dots, \tilde{X}_0 = s_0\right) \\ & \quad \times \mathbb{P}\left(X_{\tilde{X}_i, \tilde{N}_{\tilde{X}_i}^{(i, \tilde{a}_i)} + 1}^{(\tilde{a}_i)} = s_{i+1} | \tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{a}_0 = l_0, \tilde{X}_0 = s_0\right) \\ & \quad \times \mathbb{P}(\tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{X}_0 = s_0, \tilde{a}_0 = l_0) \\ &= \mathbb{P}\left(\alpha_i^{(s_0, l_0, \dots, s_{i+1})} = l_{i+1} | \tilde{X}_{i+1} = s_{i+1}, \dots, \tilde{X}_0 = s_0\right) \\ & \quad \times \mathbb{P}\left(X_{s_i, \tilde{N}_{s_i}^{(i, l_i)} + 1}^{(l_i)} = s_{i+1} | \tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{a}_0 = l_0, \tilde{X}_0 = s_0\right) \\ & \quad \times \mathbb{P}(\tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{X}_0 = s_0, \tilde{a}_0 = l_0), \end{aligned}$$

where the equalities follow by substituting in the corresponding value of each quantity. Observe that under the given conditional, such that  $\tilde{N}_{s_i}^{(i, l_i)} + 1$  is some fixed integer  $n$ . Then,

$$\begin{aligned} \mathbb{P}\left(\alpha_i^{(s_0, l_0, \dots, s_{i+1})} = l_{i+1} | \tilde{X}_{i+1} = s_{i+1}, \dots, \tilde{X}_0 = s_0\right) &= \mathbb{P}\left(\alpha_i^{(s_0, a_0, \dots, s_{i+1})} = l_{i+1}\right) \\ &= P_l^{(s_i, l_{i-1}, \dots, l_0, s_0)} \end{aligned}$$

and

$$\begin{aligned}
 & \mathbb{P} \left( X_{s_i, \tilde{N}_{s_i}^{(i, l_i)} + 1}^{(l_i)} = s_{i+1} \mid \tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{a}_0 = l_0, \tilde{X}_0 = s_0 \right) \\
 & \quad \times \mathbb{P}(\tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{X}_0 = s_0, \tilde{a}_0 = l_0) \\
 & = \mathbb{P} \left( X_{s_i, n}^{(l_i)} = s_{i+1} \mid \tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{a}_0 = l_0, \tilde{X}_0 = s_0 \right) \\
 & \quad \times \mathbb{P}(\tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{X}_0 = s_0, \tilde{a}_0 = l_0) \\
 & = \mathbb{P} \left( X_{s_i, n}^{(l_i)} = s_{i+1} \right) \mathbb{P}(\tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{X}_0 = s_0, \tilde{a}_0 = l_0) \\
 & = M_{s, t}^{(l_i)} \mathbb{P}(\tilde{X}_i = s_i, \tilde{a}_i = l_i, \dots, \tilde{X}_0 = s_0, \tilde{a}_0 = l_0) \\
 & = M_{s, t}^{(l_i)} \mathbb{P}(X_i = s_i, a_i = l_i, \dots, X_0 = s_0, a_0 = l_0),
 \end{aligned}$$

where the last equality follows by induction hypothesis. We finally get,

$$\begin{aligned}
 & \mathbb{P} \left( \tilde{X}_{i+1} = s_{i+1}, \tilde{a}_{i+1} = l_{i+1}, \dots, \tilde{X}_0 = s_0, \tilde{a}_0 = l_0 \right) \\
 & = P_l^{(s_i, l_{i-1}, \dots, l_0, s_0)} M_{s, t}^{(l_i)} \mathbb{P}(X_i = s_i, a_i = l_i, \dots, X_0 = s_0, a_0 = l_0)
 \end{aligned}$$

which is the same as

$$\begin{aligned}
 & \mathbb{P} \left( \tilde{X}_{i+1} = s_{i+1}, \tilde{a}_{i+1} = l_{i+1}, \dots, \tilde{X}_0 = s_0, \tilde{a}_0 = l_0 \right) \\
 & = \mathbb{P}(X_{i+1} = s_{i+1}, a_{i+1} = l_{i+1}, \dots, X_0 = s_0, a_0 = l_0).
 \end{aligned}$$

This completes the proof. ■

Observe that Proposition 7 does not require any assumption on the transition matrices  $M$ 's or the distribution of the actions  $a_i$ 's. However, it does require the finiteness structure of the CMC. It is unclear whether such an equivalent sampling schema exists for continuous state control processes and existing literature seems to rely on technically challenging adaptive estimation procedures (see Theorem 1 in Banerjee et al. (2025b)). However, to the best of our knowledge, while adaptive estimation works well while deriving risk bounds (Birgé, 2006; Baraud and Birgé, 2009, 2018, etc.) and the techniques cannot be generalized to derive (functional) CLT's for the estimated transition densities.

### B.3 Statement and Proof of Lemma 3

**Lemma 3** For each  $(s, l, t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ ,

$$\tilde{\xi}_{s, l, t} - \xi_{s, l, t} = \frac{\left( \tilde{N}_{s, t}^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor M_{s, t}^{(l)} \right)}{\sqrt{\mathbb{E}[N_s^{(l)}]}} - \frac{\left( N_{s, t}^{(l)} - N_s^{(l)} M_{s, t}^{(l)} \right)}{\sqrt{N_s^{(l)}}} \xrightarrow{p} 0.$$

**Proof** To show this, let  $I^{(i)} := \mathbb{1}[X_{s, i}^{(l)} = t] - M_{s, t}^{(l)}$  and define  $S_n := \sum_{i=1}^n I^{(i)}$ . We have

$$\frac{\left( \tilde{N}_{s, t}^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor M_{s, t}^{(l)} \right)}{\sqrt{n^{\frac{1}{1+\alpha}}}} - \frac{\left( N_{s, t}^{(l)} - N_s^{(l)} M_{s, t}^{(l)} \right)}{\sqrt{n^{\frac{1}{1+\alpha}}}} = \frac{S_{\lfloor \mathbb{E}[N_s^{(l)}] \rfloor} - S_{N_s^{(l)}}}{\sqrt{n^{\frac{1}{1+\alpha}}}}.$$

Using Lemma 2, for all  $\varepsilon > 0$  and  $n$  sufficiently large,  $\mathbb{P}\left(\left|N_s^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor\right| > \sqrt{n^{1/\alpha}}\varepsilon\right) < \varepsilon$ . Therefore,

$$\begin{aligned} & \mathbb{P}\left(\left|S_{\lfloor \mathbb{E}[N_s^{(l)}] \rfloor} - S_{N_s^{(l)}}\right| > \sqrt{n^{1/\alpha}}\varepsilon\right) \\ & \leq \mathbb{P}\left(\left|N_s^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor\right| > \sqrt{n^{1/\alpha}}\varepsilon\right) \\ & \quad + \mathbb{P}\left(\left|N_s^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor\right| \leq \sqrt{n^{1/\alpha}}\varepsilon, \left|S_{\lfloor \mathbb{E}[N_s^{(l)}] \rfloor} - S_{N_s^{(l)}}\right| > \sqrt{n^{1/\alpha}}\varepsilon\right) \\ & \leq \varepsilon + \mathbb{P}\left(\left|N_s^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor\right| \leq \sqrt{n^{1/\alpha}}\varepsilon, \left|S_{\lfloor \mathbb{E}[N_s^{(l)}] \rfloor} - S_{N_s^{(l)}}\right| > \sqrt{n^{1/\alpha}}\varepsilon\right). \end{aligned}$$

Observe that  $\mathbb{1}[X_{s,i}^{(a)} = t]$  and  $M_{s,t}^{(a)}$  are between 0 and 1. Hence  $|I^{(i)}| \leq 1$ . Then

$$\mathbb{P}\left(\left|N_s^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor\right| \leq \sqrt{n^{1/\alpha}}\varepsilon, \left|S_{\lfloor \mathbb{E}[N_s^{(l)}] \rfloor} - S_{N_s^{(l)}}\right| > \sqrt{n^{1/\alpha}}\varepsilon\right) = 0,$$

and

$$\mathbb{P}\left(\left|S_{\lfloor \mathbb{E}[N_s^{(l)}] \rfloor} - S_{N_s^{(l)}}\right| > \sqrt{n^{1/\alpha}}\varepsilon\right) \leq \varepsilon.$$

Since  $\varepsilon$  is arbitrary, it now follows that

$$\frac{S_{\lfloor \mathbb{E}[N_s^{(l)}] \rfloor} - S_{N_s^{(l)}}}{\sqrt{n^{1/\alpha}}} \xrightarrow{p} 0.$$

Next, observe that

$$\begin{aligned} & \frac{\left(\tilde{N}_{s,t}^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor M_{s,t}^{(l)}\right)}{\sqrt{\mathbb{E}[N_s^{(l)}]}} - \frac{\left(N_{s,t}^{(l)} - N_s^{(l)} M_{s,t}^{(l)}\right)}{\sqrt{N_s^{(l)}}} \\ & = \sqrt{\frac{n^{1/\alpha}}{\mathbb{E}[N_s^{(l)}]}} \left( \frac{\left(\tilde{N}_{s,t}^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor M_{s,t}^{(l)}\right)}{\sqrt{n^{1/\alpha}}} - \frac{\left(N_{s,t}^{(l)} - N_s^{(l)} M_{s,t}^{(l)}\right)}{\sqrt{n^{1/\alpha}} \left(\sqrt{N_s^{(l)}}/\mathbb{E}[N_s^{(l)}]\right)} \right). \end{aligned}$$

By Proposition 2, for sufficiently large  $n$ ,  $\mathbb{E}[N_s^{(l)}] \geq 1/cn^{1/\alpha}$ , where  $c > 0$  is a constant, and  $\alpha \in (0, 1)$ . By Lemma 2,  $N_s^{(l)}/\mathbb{E}[N_s^{(l)}] \xrightarrow{a.s.} 1$ . Therefore,  $\sqrt{n^{1/\alpha}/\mathbb{E}[N_s^{(l)}]} \leq \sqrt{c}$  and

$$\begin{aligned} & \frac{\left(\tilde{N}_{s,t}^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor M_{s,t}^{(l)}\right)}{\sqrt{\mathbb{E}[N_s^{(l)}]}} - \frac{\left(N_{s,t}^{(l)} - N_s^{(l)} M_{s,t}^{(l)}\right)}{\sqrt{N_s^{(l)}}} \\ & \leq \sqrt{c} \left( \frac{\left(\tilde{N}_{s,t}^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor M_{s,t}^{(l)}\right)}{\sqrt{n^{1/\alpha}}} - \frac{\left(N_{s,t}^{(l)} - N_s^{(l)} M_{s,t}^{(l)}\right)}{\sqrt{n^{1/\alpha}} \left(\sqrt{N_s^{(l)}}/\mathbb{E}[N_s^{(l)}]\right)} \right) \xrightarrow{p} 0. \end{aligned}$$

This completes the proof. ■

### Appendix C. Proof of Theorem 1

**Proof** We begin by presenting the following generalization of the sampling scheme given in Billingsley (1961) to the case of controlled Markov chains. For each  $l \in \mathcal{A}$ , create the following infinite array of i.i.d random variables which are also *independent* of the data  $\{(X_0, a_0), \dots, (X_n, a_n)\}$ :

$$\mathbb{X}^{(l)} : \begin{bmatrix} X_{1,1}^{(l)} & X_{1,2}^{(l)} & \dots & X_{1,\tau}^{(l)} & \dots \\ X_{2,1}^{(l)} & X_{2,2}^{(l)} & \dots & X_{2,\tau}^{(l)} & \dots \\ \dots & \dots & \dots & \dots & \dots \\ X_{d,1}^{(l)} & X_{d,2}^{(l)} & \dots & X_{d,\tau}^{(l)} & \dots \end{bmatrix},$$

where,  $\forall (s, t, \tau) \in \{1, \dots, d\} \times \{1, \dots, d\} \times \mathbb{N}$ , the random variables  $X_{s,\tau}^{(l)}$  follow the probability mass function given by  $\mathbb{P}(X_{s,\tau}^{(l)} = t) = M_{s,t}^{(l)}$ . Moreover, for every time  $i \geq 1$ , and  $((s_0, l_0), \dots, (s_{i-1}, l_{i-1}), s_i) \in (\mathcal{S} \times \mathcal{A})^i \times \mathcal{S}$ , let  $\alpha_i^{((s_0, l_0), \dots, (s_{i-1}, l_{i-1}), s_i)}$  be independent random variables with support  $\mathcal{A}$  and probability mass function given by

$$\begin{aligned} & P_l^{((s_0, l_0), \dots, (s_{i-1}, l_{i-1}), s_i)} \\ & := \mathbb{P}(\alpha_i^{((s_0, l_0), \dots, (s_{i-1}, l_{i-1}), s_i)} = l) \\ & = \mathbb{P}(a_i = l | X_i = s_i, \mathcal{H}_0^{i-1} = s_0, l_0, \dots, s_{i-1}, l_{i-1}). \end{aligned}$$

The sampling scheme runs as follows. Set  $(\tilde{X}_0, \tilde{a}_0) \stackrel{d}{=} (X_0, a_0)$  and recursively sample  $\tilde{X}_{i+1} = X_{\tilde{X}_i, \tilde{N}_{\tilde{X}_i}^{(i, \tilde{a}_i)} + 1}^{(\tilde{a}_i)}$  from the array  $\mathbb{X}^{(\tilde{a}_i)}$  and  $\tilde{a}_{i+1} = \alpha_{i+1}^{(\tilde{X}_0, \tilde{a}_0, \dots, \tilde{X}_{i+1})}$ , where each  $i \geq 0$ .

Define  $\tilde{N}_s^{(i, l)} := \sum_{j \leq i} \mathbb{1}[\tilde{X}_j = s, \tilde{a}_j = l]$  and  $\tilde{N}_s^{(l)} := \sum_{i=1}^n \mathbb{1}[\tilde{X}_i = s, \tilde{a}_i = l]$ , where  $\tilde{N}_s^{(n, l)} = \tilde{N}_s^{(l)}$ .

This completes the sampling scheme.

Proposition 7 implies that  $\tilde{N}_s^{(l)} \stackrel{d}{=} N_s^{(l)}$ , hence  $\mathbb{E}[\tilde{N}_s^{(l)}] = \mathbb{E}[N_s^{(l)}]$ . Let  $\tilde{N}_{s,t}^{(l)}$  be a random variable defined as

$$\tilde{N}_{s,t}^{(l)} := \sum_{\tau=1}^{\lfloor \mathbb{E}[\tilde{N}_s^{(l)}] \rfloor} \mathbb{1}[\tilde{X}_{i,\tau}^{(l)} = t]. \quad (\text{C.1})$$

Let the  $kd^2$  dimensional vector  $(\tilde{\xi}_{s,l,t})$  be defined element-wise as

$$\tilde{\xi}_{s,l,t} := \frac{\left( \tilde{N}_{s,t}^{(l)} - \lfloor \mathbb{E}[\tilde{N}_s^{(l)}] \rfloor M_{s,t}^{(l)} \right)}{\sqrt{\mathbb{E}[\tilde{N}_s^{(l)}]}} = \frac{\left( \tilde{N}_{s,t}^{(l)} - \lfloor \mathbb{E}[N_s^{(l)}] \rfloor M_{s,t}^{(l)} \right)}{\sqrt{\mathbb{E}[N_s^{(l)}]}}.$$

By definition,  $(\tilde{N}_{s,1}^{(l)}, \dots, \tilde{N}_{s,d}^{(l)})$  is the frequency count of  $\{X_{s,1}^{(l)}, \dots, X_{s, \lfloor \mathbb{E}[N_s^{(l)}] \rfloor}^{(l)}\}$ . From the independence of  $\{\mathbb{X}^{(l)}\}$  and the central limit theorem for multinomial trials, it follows that  $\tilde{\xi}$

is distributed asymptotically normally with covariance matrix given by (3.1) and

$$\tilde{\xi} \xrightarrow{d} \mathcal{N}(0, \Lambda).$$

Lemma 3 implies that  $\tilde{\xi} \xrightarrow{p} \xi$ . Therefore,  $\xi \xrightarrow{d} \mathcal{N}(0, \Lambda)$ . This completes the proof.  $\blacksquare$

## Appendix D. Proofs of CLTs for the Value, Q-, and Advantage Functions, and the Optimal Policy Value

### D.1 Proof of Theorem 2

**Proof** First, we show the asymptotic normality of  $\sqrt{n}(\hat{V}_{\Pi} - V_{\Pi})$ . Observe that

$$\hat{V}_{\Pi} - V_{\Pi} = \left( (I - \alpha\Pi\hat{\mathbf{M}})^{-1} - (I - \alpha\Pi\mathbf{M})^{-1} \right) g,$$

where the right-hand side can be rewritten as

$$\begin{aligned} \text{RHS} &= \alpha (I - \alpha\Pi\hat{\mathbf{M}})^{-1} \Pi (\hat{\mathbf{M}} - \mathbf{M}) (I - \alpha\Pi\mathbf{M})^{-1} g \\ &= \alpha (I - \alpha\Pi\hat{\mathbf{M}})^{-1} \Pi (\hat{\mathbf{M}} - \mathbf{M}) V_{\Pi}, \end{aligned}$$

using the matrix property  $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ . Taking the transpose and vectorizing both sides, we get

$$\begin{aligned} (\hat{V}_{\Pi} - V_{\Pi})^{\top} &= \alpha V_{\Pi}^{\top} (\hat{\mathbf{M}}^{\top} - \mathbf{M}^{\top}) \Pi^{\top} (I - \alpha\hat{\mathbf{M}}^{\top} \Pi^{\top})^{-1} \\ \text{Vec} \left( (\hat{V}_{\Pi} - V_{\Pi})^{\top} \right) &= \alpha \left[ (I - \alpha\Pi\hat{\mathbf{M}})^{-1} \Pi \otimes V_{\Pi}^{\top} \right] \text{Vec} (\hat{\mathbf{M}}^{\top} - \mathbf{M}^{\top}) \\ \hat{V}_{\Pi} - V_{\Pi} &= \alpha \left[ (I - \alpha\Pi\hat{\mathbf{M}})^{-1} \Pi \otimes V_{\Pi}^{\top} \right] \text{Vec} (\hat{\mathbf{M}}^{\top} - \mathbf{M}^{\top}). \end{aligned}$$

Now, multiplying both sides by  $\sqrt{n}$  and using Corollary 1, the asymptotic normality of  $\sqrt{n}(\hat{V}_{\Pi} - V_{\Pi})$  follows from the Slutsky's theorem.

Second, we show the asymptotic normality of  $\sqrt{n}(\hat{Q}_{\Pi} - Q_{\Pi})$ . Similar to the value function, observe that

$$\hat{Q}_{\Pi} - Q_{\Pi} = \left( (I - \alpha\hat{\mathbf{M}}\Pi)^{-1} - (I - \alpha\mathbf{M}\Pi)^{-1} \right) r,$$

whose right-hand side can be rewritten as

$$\begin{aligned} \text{RHS} &= \alpha (I - \alpha\hat{\mathbf{M}}\Pi)^{-1} (\hat{\mathbf{M}} - \mathbf{M}) \Pi (I - \alpha\mathbf{M}\Pi)^{-1} r \\ &= \alpha (I - \alpha\hat{\mathbf{M}}\Pi)^{-1} (\hat{\mathbf{M}} - \mathbf{M}) \Pi Q_{\Pi}. \end{aligned}$$

Taking the transpose and vectorizing both sides, we get

$$\begin{aligned} (\hat{Q}_\Pi - Q_\Pi)^\top &= \alpha Q_\Pi^\top \Pi^\top (\hat{\mathbf{M}}^\top - \mathbf{M}^\top) (I - \alpha \Pi^\top \hat{\mathbf{M}}^\top)^{-1} \\ \text{Vec} \left( (\hat{Q}_\Pi - Q_\Pi)^\top \right) &= \alpha \left[ (I - \alpha \hat{\mathbf{M}} \Pi)^{-1} \otimes Q_\Pi^\top \Pi^\top \right] \text{Vec} (\hat{\mathbf{M}}^\top - \mathbf{M}^\top) \\ \hat{Q}_\Pi - Q_\Pi &= \alpha \left[ (I - \alpha \hat{\mathbf{M}} \Pi)^{-1} \otimes Q_\Pi^\top \Pi^\top \right] \text{Vec} (\hat{\mathbf{M}}^\top - \mathbf{M}^\top). \end{aligned}$$

Now, multiplying both sides by  $\sqrt{n}$  and using Corollary 1, the asymptotic normality of  $\sqrt{n} (\hat{Q}_\Pi - Q_\Pi)$  follows from the Slutsky's theorem.

Finally, we show the asymptotic normality of  $\sqrt{n} (\hat{A}_\Pi - A_\Pi)$ . Recall that  $K = \text{diag}(\mathbf{1}_k, \dots, \mathbf{1}_k)$ , where  $\mathbf{1}_k$  is the ones vector of length  $k$ , and  $A_\Pi = Q_\Pi - KV_\Pi$ . Then

$$\begin{aligned} \hat{A}_\Pi - A_\Pi &= (\hat{Q}_\Pi - Q_\Pi) - K (\hat{V}_\Pi - V_\Pi) \\ &= \alpha \left[ (I - \alpha \hat{\mathbf{M}} \Pi)^{-1} \otimes Q_\Pi^\top \Pi^\top \right] \text{Vec} (\hat{\mathbf{M}}^\top - \mathbf{M}^\top) \\ &\quad - \alpha K \left[ (I - \alpha \Pi \hat{\mathbf{M}})^{-1} \Pi \otimes V_\Pi^\top \right] \text{Vec} (\hat{\mathbf{M}}^\top - \mathbf{M}^\top). \end{aligned}$$

Now, multiplying both sides by  $\sqrt{n}$  and using Corollary 1, the asymptotic normality of  $\sqrt{n} (\hat{A}_\Pi - A_\Pi)$  follows from the Slutsky's theorem.  $\blacksquare$

## D.2 Proof of Corollary 2

**Proof** Following an entirely analogous way to the proof of Theorem 2, we get

$$\hat{Q}_\Pi - Q_\Pi = B \text{Vec} (\hat{\mathbf{M}}^\top - \mathbf{M}^\top),$$

where  $B = \alpha \left[ (I - \alpha \hat{\mathbf{M}}\Pi)^{-1} \otimes Q_\Pi^\top \Pi^\top \right]$  is a  $dk \times d^2k$  matrix. Let  $b_{(s,l),(s,t)}^{(l)}$  denote the  $((s-1)k+l, (s-1)dk + (l-1)d+t)$  entry of  $B$ . Then

$$\begin{aligned} (\hat{Q}_\Pi - Q_\Pi) \odot \text{Vec}(\mathbf{N}^\top) &= B \text{Vec}(\hat{\mathbf{M}}^\top - \mathbf{M}^\top) \\ &= \begin{bmatrix} \sum_{(s,l) \in \mathcal{S} \times \mathcal{A}, t \in \mathcal{S}} \sqrt{N_1^{(1)}} b_{(1,1),(s,t)}^{(l)} (\hat{M}_{s,t}^{(l)} - M_{s,t}^{(l)}), \\ \dots, \sum_{(s,l) \in \mathcal{S} \times \mathcal{A}, t \in \mathcal{S}} \sqrt{N_d^{(k)}} b_{(d,k),(s,t)}^{(l)} (\hat{M}_{s,t}^{(l)} - M_{s,t}^{(l)}) \end{bmatrix}^\top \\ &= \begin{bmatrix} \sum_{(s,l) \in \mathcal{S} \times \mathcal{A}, t \in \mathcal{S}} \sqrt{\frac{N_1^{(1)}}{N_s^{(l)}}} b_{(1,1),(s,t)}^{(l)} \sqrt{N_s^{(l)}} (\hat{M}_{s,t}^{(l)} - M_{s,t}^{(l)}), \\ \dots, \sum_{(s,l) \in \mathcal{S} \times \mathcal{A}, t \in \mathcal{S}} \sqrt{\frac{N_d^{(k)}}{N_s^{(l)}}} b_{(d,k),(s,t)}^{(l)} \sqrt{N_s^{(l)}} (\hat{M}_{s,t}^{(l)} - M_{s,t}^{(l)}) \end{bmatrix}^\top. \end{aligned}$$

Lemma 2 ensures that  $\mathbb{E}[N_s^{(l)}]/N_s^{(l)} \xrightarrow{a.s.} 1$ , and Assumption 5 ensures that  $\mathbb{E}[N_s^{(l)}]/n \rightarrow p_s^{(l)} > 0$ , hence  $\sqrt{N_{s'}^{(l')}/N_s^{(l)}} \xrightarrow{a.s.} p_{s'}^{(l')}/p_s^{(l)}$  for any  $(s,l), (s',l') \in \mathcal{S} \times \mathcal{A}$ . The asymptotic normality of  $(\hat{Q}_\Pi - Q_\Pi) \odot \text{Vec}(\mathbf{N}^\top)$  then follows from Theorem 1.  $\blacksquare$

### D.3 Proof of Theorem 3

**Proof** Under the hypothesis of the theorem, as long as  $\sup_{\Pi \in \Delta(\mathcal{A})} \|\hat{V}_\Pi - V_\Pi\|_\infty \xrightarrow{p} 0$  it follows from the consistency of M-estimators (Van der Vaart, 2000, Theorem 5.7) that  $\hat{\Pi}_{opt} \xrightarrow{p} \Pi_{opt}$ . Fix any  $\Pi$ , and observe that

$$\begin{aligned} \|V_\Pi - \hat{V}_\Pi\|_\infty &= \left\| \left( (I - \alpha \Pi \hat{\mathbf{M}})^{-1} - (I - \alpha \Pi \mathbf{M})^{-1} \right) g \right\|_\infty \\ &\leq \left\| \left( (I - \alpha \Pi \hat{\mathbf{M}})^{-1} - (I - \alpha \Pi \mathbf{M})^{-1} \right) \right\|_\infty \|g\|_1 \\ &\leq (1 - \alpha)^{-2} \sqrt{dk} \left\| (I - \alpha \Pi \hat{\mathbf{M}}) - (I - \alpha \Pi \mathbf{M}) \right\|_\infty \|g\|_1 \\ &\leq \frac{\alpha}{(1 - \alpha)^2} \sqrt{dk} \left\| \Pi \hat{\mathbf{M}} - \Pi \mathbf{M} \right\|_\infty \|g\|_1 \\ &\leq \frac{\alpha}{(1 - \alpha)^2} \sqrt{dk} \|\Pi\|_{1,\infty} \left\| \hat{\mathbf{M}} - \mathbf{M} \right\|_\infty \|g\|_1 \\ &= \frac{\alpha}{(1 - \alpha)^2} \sqrt{dk} \left\| \hat{\mathbf{M}} - \mathbf{M} \right\|_\infty \|g\|_1 \end{aligned}$$

where  $\|\cdot\|_{1,\infty}$  is the row-wise 1 and column-wise  $\infty$  norm of a matrix. Observe that the rows of any policy sum to 1, therefore  $\|\Pi\|_{1,\infty} = 1$ . The right-hand side is independent of  $\Pi$  and

$\hat{\mathbf{M}} \xrightarrow{p} \mathbf{M}$ . Therefore, by taking the supremum with respect to  $\Pi$  on both sides and letting  $n$  tend to  $\infty$ , an application of Van der Vaart (2000, Theorem 5.7) implies that  $\hat{\Pi}_{opt} \xrightarrow{p} \Pi_{opt}$ .

We now prove the second part of our theorem. Since  $\hat{\Pi}_{opt}$  maximizes  $\hat{V}_{\hat{\Pi}_{opt}}$ , it follows that for any given state  $s$ ,  $\hat{V}_{\hat{\Pi}_{opt}}(s) - \hat{V}_{\Pi_{opt}}(s) > 0$ , and

$$\begin{aligned} \hat{V}_{\hat{\Pi}_{opt}} - V_{\Pi_{opt}} &= \hat{V}_{\hat{\Pi}_{opt}} - \hat{V}_{\Pi_{opt}} + \hat{V}_{\Pi_{opt}} - V_{\Pi_{opt}} \\ &\geq \hat{V}_{\Pi_{opt}} - V_{\Pi_{opt}}, \end{aligned}$$

where the vector inequalities are coordinate-wise. It similarly follows that

$$\hat{V}_{\hat{\Pi}_{opt}} - V_{\Pi_{opt}} \leq \hat{V}_{\hat{\Pi}_{opt}} - V_{\hat{\Pi}_{opt}}.$$

Therefore,

$$\begin{aligned} \hat{V}_{\Pi_{opt}} - V_{\Pi_{opt}} &\leq \hat{V}_{\hat{\Pi}_{opt}} - V_{\Pi_{opt}} \leq \hat{V}_{\hat{\Pi}_{opt}} - V_{\hat{\Pi}_{opt}} \text{ and} \\ \sqrt{n} \left( \hat{V}_{\Pi_{opt}} - V_{\Pi_{opt}} \right) &\leq \sqrt{n} \left( \hat{V}_{\hat{\Pi}_{opt}} - V_{\Pi_{opt}} \right) \leq \sqrt{n} \left( \hat{V}_{\hat{\Pi}_{opt}} - V_{\hat{\Pi}_{opt}} \right). \end{aligned}$$

Since  $\hat{\Pi}_{opt} \xrightarrow{p} \Pi_{opt}$  it now follows using the Slutsky's theorem and the Sandwich theorem that

$$\sqrt{n} \left( \hat{V}_{\hat{\Pi}_{opt}} - V_{\Pi_{opt}} \right) \xrightarrow{d} \mathcal{N} \left( 0, \Sigma_V^{(\Pi_{opt})} \right).$$

This completes the proof. ■

## Appendix E. Proofs of Other Propositions

### E.1 Proof of Proposition 1

**Proof** Given a sample of the controlled Markov chain  $\{(s_i, l_i)\}_{i=0}^n$ , the likelihood of the transition probabilities  $\mathbf{M}$  is

$$\mathcal{L}(\mathbf{M}) = \mathbb{P}(X_0 = s_0) \pi^{(0)}(s_0, l_0) \prod_{i=1}^n \left( M_{s_{i-1}, s_i}^{(l_{i-1})} \pi^{(i)}(s_i, l_i) \right),$$

where  $\pi^{(i)}$  is the conditional probability of  $\{X_i = s_i, a_i = l_i\}$  given  $\{\mathcal{H}_0^{i-1} = \mathcal{h}_0^{i-1}\}$ . Then, the log-likelihood is

$$l(\mathbf{M}) = \sum_{i=1}^n \log M_{s_{i-1}, s_i}^{(l_{i-1})} + \sum_{i=1}^n \log \pi^{(i)}(s_i, l_i) + \log \mathbb{P}(X_0 = s_0) + \log \pi^{(0)}(s_0, l_0).$$

Given the  $dk$  constraint equations

$$\sum_{t \in \mathcal{S}} M_{s,t}^{(l)} = 1,$$

and  $dk$  lagrange multipliers  $\lambda_{s,l}$ ,  $s = 1, \dots, d$ ,  $l = 1, \dots, k$ , the Lagrangian is

$$l(\mathbf{M}) - \sum_{s=1}^d \sum_{l=1}^k \lambda_{s,l} \left( \sum_{t \in \mathcal{S}} M_{s,t}^{(l)} - 1 \right).$$

Taking derivatives with respect to  $M_{s,t}^{(l)}$ , and setting them to be 0 at  $\hat{M}_{s,t}^{(l)}$ , we have

$$\begin{aligned} \frac{N_{s,t}^{(l)}}{\hat{M}_{s,t}^{(l)}} - \lambda_{s,l} &= 0, \\ \frac{N_{s,t}^{(l)}}{\lambda_{s,l}} &= \hat{M}_{s,t}^{(l)}. \end{aligned}$$

Thus,

$$\sum_{t \in \mathcal{S}} \frac{N_{s,t}^{(l)}}{\lambda_{s,l}} = 1, \quad \lambda_{s,l} = \sum_{t \in \mathcal{S}} N_{s,t}^{(l)} = N_s^{(l)}.$$

Therefore,  $\hat{M}_{s,t}^{(l)} = \frac{N_{s,t}^{(l)}}{N_s^{(l)}}$ , the count-based empirical estimator is the maximum likelihood estimator. ■

## E.2 Proof of Proposition 2

**Proof** Define the process  $Z_i$  as

$$Z_i = \sum_{p=1}^i \left( \frac{\tau_{s,l}^{(p)}}{T_p} - 1 \right), \quad Z_0 = 0.$$

Then

$$\begin{aligned} \mathbb{E} \left[ Z_i | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] &= \sum_{p=1}^i \mathbb{E} \left[ \frac{\tau_{s,l}^{(p)}}{T_p} - 1 | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] \\ &= \mathbb{E} \left[ \frac{\tau_{s,l}^{(i)}}{T_i} - 1 | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] + \sum_{p=1}^{i-1} \left( \frac{\tau_{s,l}^{(p)}}{T_p} - 1 \right) \\ &= \mathbb{E} \left[ \frac{\tau_{s,l}^{(i)}}{T_i} - 1 | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] + Z_{i-1}. \end{aligned}$$

As  $\mathbb{E}[\tau_{s,l}^{(i)} | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}}] \leq T_i$  almost surely,  $\mathbb{E} \left[ \frac{\tau_{s,l}^{(i)}}{T_i} - 1 | \mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}} \right] < 0$ , hence  $\{Z_i\}$  is a supermartingale with respect to  $\mathcal{F}_{\sum_{p=1}^{i-1} \tau_{s,l}^{(p)}}$ .

Let  $N = N_s^{(l)}(n) + 1$ . As  $N_s^{(l)}(n)$  is a valid stopping time,  $N$  is a valid stopping time, and  $\sum_{i=1}^N \tau_{s,l}^{(i)} > n$ . By Doob's Optional Stopping Theorem,

$$\mathbb{E}[Z_N] \leq \mathbb{E}[Z_0] = 0 \implies \mathbb{E} \left[ \sum_{i=1}^N \frac{\tau_{s,l}^{(i)}}{T_i} \right] \leq \mathbb{E}[N].$$

We may assume without loss of generality that  $\{T_i\}$  is non-decreasing. Indeed, the non-decreasing case is the only regime in which the stopping-time argument below is needed; if  $\{T_i\}$  is not eventually non-decreasing, the claim follows by simpler deterministic bounds.

Since  $\sum_{i=1}^N \tau_{s,l}^{(i)} > n$ , and  $\{T_i\}$  is a non-decreasing sequence, we have

$$\sum_{i=1}^N \frac{\tau_{s,l}^{(i)}}{T_i} \geq \frac{\sum_{i=1}^N \tau_{s,l}^{(i)}}{T_N} > \frac{n}{T_N} \implies \mathbb{E} \left[ \frac{n}{T_N} \right] \leq \mathbb{E}[N].$$

For the convex function  $\phi(x) = n/x$ , the Jensen's inequality gives

$$\mathbb{E} \left[ \frac{n}{T_N} \right] \geq \frac{n}{\mathbb{E}[T_N]}.$$

Therefore,

$$\mathbb{E}[N] \geq \mathbb{E} \left[ \frac{n}{T_N} \right] \geq \frac{n}{\mathbb{E}[T_N]}.$$

As  $T_i = O(i^\alpha)$ , we have  $T_N = O(N^\alpha)$ , hence  $\mathbb{E}[T_N] = O(\mathbb{E}[N^\alpha])$ . For the concave function  $\phi(x) = x^\alpha$  ( $\alpha < 1$ ), Jensen's inequality gives

$$\mathbb{E}[N^\alpha] \leq \mathbb{E}[N]^\alpha.$$

Hence,  $\mathbb{E}[T_N] = O(\mathbb{E}[N]^\alpha)$  and thus  $\mathbb{E}[N] = \Omega \left( \frac{n}{\mathbb{E}[N]^\alpha} \right)$ , which implies  $\mathbb{E}[N]^{1+\alpha} = \Omega(n)$  and in turn  $\mathbb{E}[N_s^{(l)}(n)] = \Omega \left( n^{\frac{1}{1+\alpha}} \right)$ . ■

### E.3 Proof of Proposition 3

**Proof** Let  $\mathcal{M}_{\mathcal{S},\mathcal{A}}$  be the class of all probability measures over state-action pairs for a CMC with initial distribution  $D_0$ . Any element  $\mathcal{P} \in \mathcal{M}_{\mathcal{S},\mathcal{A}}$  has an associated set of transition matrices  $\{M^{(1)}, \dots, M^{(k)}\}$  and a conditional distributions over the action  $\{a_i\}$  conditional on the history until time  $i$ . Then the *minimax risk* of an estimator  $\hat{M} = (\hat{M}^{(1)}, \dots, \hat{M}^{(k)})$  of  $M = (M^{(1)}, \dots, M^{(k)})$  is defined as

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

The proof now proceeds by creating a counterexample and then bounding from below its minimax risk. Suppose for some  $(s, l) \in \mathcal{S} \times \mathcal{A}$ ,

$$\mathbb{P}(a_i = l | X_i = s) = 1/(i+1)^2 \text{ for all } i \geq 0.$$

It is easy to verify that  $\sum_{i=1}^{\infty} \mathbb{P}(X_i = s, a_i = l) < \infty$ . By the Borel-Cantelli lemma, as

$$\sum_{i=1}^{\infty} \mathbb{P}(X_i = s, a_i = l) < \infty,$$

$N_s^{(l)}(\infty) < \infty$  almost surely. Therefore, there exists a constant  $N$  such that  $N_s^{(l)}(\infty) \leq N$  almost surely.

Suppose that  $|\mathcal{S}| = 2(N+1)$ . Given any  $\varepsilon \in (0, 1/2(N+1))$ , the transition probabilities from the state-action pair  $(s, l)$  is given by

$$\begin{aligned} M_s^{(l)}(\xi) &:= \left[ M_{s,1}^{(l)}(\xi), \dots, M_{s,2(N+1)}^{(l)}(\xi) \right] \\ &= \left[ \underbrace{2\varepsilon\xi_1, \dots, 2\varepsilon\xi_{N+1}}_{N+1 \text{ times}}, \underbrace{\frac{1}{N+1} - 2\varepsilon\xi_1, \dots, \frac{1}{N+1} - 2\varepsilon\xi_{N+1}}_{N+1 \text{ times}} \right], \end{aligned} \quad (\text{E.1})$$

where  $\xi = \{\xi_1, \dots, \xi_{N+1}\} \in \{0, 1\}^{N+1}$  are unknown parameters. Therefore, to correctly estimate the transition matrix  $M_s^{(l)}(\xi)$ , we only need to correctly estimate  $\xi$ .

By (E.1), we have  $\mathbb{P}(N_{s,1}^{(l)} = 0, N_{s,N+2}^{(l)} = 0) \geq (N/N+1)^N \geq 1/e > 1/3$ . Now we observe that

$$\sup_{l \in \mathcal{A}} \|\hat{M}^{(l)} - M^{(l)}\|_{\infty} \geq \left\| \hat{M}_s^{(l)} - M_s^{(l)}(\xi) \right\|_{\infty}.$$

Therefore,

$$\begin{aligned} \mathcal{R}_m &\geq \inf_{\hat{M}^{(1)}} \sup_{\xi^{(1)} \in \{0,1\}^{(d/3)}} \mathbb{P} \left( \left\| \hat{M}_s^{(l)} - M_s^{(l)}(\xi) \right\|_{\infty} > \varepsilon \right) \\ &\geq \inf_{\hat{M}^{(1)}} \sup_{\xi^{(1)} \in \{0,1\}^{(d/3)}} \mathbb{P} \left( \left\| \hat{M}_s^{(l)} - M_s^{(l)}(\xi) \right\|_{\infty} > \varepsilon \mid N_{s,1}^{(l)} = 0, N_{s,N+2}^{(l)} = 0 \right) \\ &\quad \times \mathbb{P} \left( N_{s,1}^{(l)} = 0, N_{s,N+2}^{(l)} = 0 \right) \\ &> \frac{1}{3} \inf_{\hat{M}^{(1)}} \sup_{\xi^{(1)} \in \{0,1\}^{(d/3)}} \mathbb{P} \left( \left\| \hat{M}_s^{(l)} - M_s^{(l)}(\xi) \right\|_{\infty} > \varepsilon \mid N_{s,1}^{(l)} = 0, N_{s,N+2}^{(l)} = 0 \right). \end{aligned}$$

We note that whenever  $\xi^{(1)} \neq \xi^{(2)} \in \{0, 1\}^{N+1}$ , we have  $\left\| M_s^{(l)}(\xi^{(1)}) - M_s^{(l)}(\xi^{(2)}) \right\|_{\infty} = 2\varepsilon$ . For any estimate  $\hat{M}_s^{(l)}$  of  $M_s^{(l)}(\xi)$ , define  $\xi^*$  to be the  $\xi$  such that  $\xi^* = \arg \min_{\xi} \left\| \hat{M}_s^{(l)} - M_s^{(l)}(\xi) \right\|_{\infty}$ . Then for  $\xi \neq \xi^*$  we have

$$\begin{aligned} 2\varepsilon &= \left\| M_s^{(l)}(\xi) - M_s^{(l)}(\xi^*) \right\|_{\infty} \leq \left\| M_s^{(l)}(\xi) - \hat{M}_s^{(l)} \right\|_{\infty} + \left\| \hat{M}_s^{(l)} - M_s^{(l)}(\xi^*) \right\|_{\infty} \\ &\leq 2 \left\| M_s^{(l)}(\xi) - \hat{M}_s^{(l)} \right\|_{\infty}. \end{aligned}$$

Therefore,  $\{\xi^* \neq \xi\} \subset \left\{ \left\| M_s^{(l)}(\xi) - \hat{M}_s^{(l)} \right\|_\infty \geq \varepsilon \right\}$  and  $\mathcal{R}_m$  can be further lower bounded by

$$\begin{aligned} \mathcal{R}_m &> \frac{1}{3} \inf_{\hat{M}_s^{(l)}} \max_{\xi \in \{0,1\}^{N+1}} \mathbb{P} \left( \xi^* \neq \xi \mid N_{s,1}^{(l)} = 0, N_{s,N+2}^{(l)} = 0 \right) \\ &= \frac{1}{3} \inf_{\hat{\xi}} \max_{\xi \in \{0,1\}^{N+1}} \mathbb{P} \left( \hat{\xi} \neq \xi \mid N_{s,1}^{(l)} = 0, N_{s,N+2}^{(l)} = 0 \right), \end{aligned}$$

where  $\hat{\xi}$  is any estimate of  $\xi^*$  such that  $\hat{\xi} : (X_0, a_0, \dots, X_n, a_n) \mapsto \{0, 1\}^{N+1}$ .

When  $N_{s,1}^{(l)} = 0$  and  $N_{s,N+2}^{(l)} = 0$ , the estimate  $\hat{\xi}_1$  is equivalent to choosing uniformly over  $\{0, 1\}$ . The probability of choosing incorrectly is  $1/2$ . We get as a consequence that,

$$\frac{1}{3} \inf_{\hat{\xi}} \max_{\xi \in \{0,1\}^{N+1}} \mathbb{P} \left( \hat{\xi} \neq \xi \mid N_{s,1}^{(l)} = 0, N_{s,N+2}^{(l)} = 0 \right) \geq \frac{1}{6}.$$

In conclusion,

$$\mathcal{R}_m > \frac{1}{6}.$$

Therefore, for every  $\varepsilon \in (0, 1/2(N+1))$ , there exists a controlled Markov chain such that it has no estimator  $\hat{M}^{(l)}$  such that  $\mathbb{P} \left( \sup_{l \in \mathcal{A}} \|\hat{M}^{(l)} - M^{(l)}\|_\infty > \varepsilon \right) \leq 1/6$ .

It follows that for any sequence  $b_n \rightarrow \infty$ , the event  $\left\{ b_n \sup_{s,l,t} \left| \hat{M}_{s,t}^{(l)} - M_{s,t}^{(l)} \right| \rightarrow \infty \right\}$  has probability at least  $1/6$ . The conclusion follows.  $\blacksquare$

#### E.4 Proof of Proposition 4

**Proof** As Assumptions 1 and 2 hold, Theorem 1 holds. Then for any  $(s, l) \in \mathcal{S} \times \mathcal{A}$ , we have the following Pearson's chi-square test statistics

$$\sum_{t \in \mathcal{S}} \frac{\left( N_{s,t}^{(l)} - N_s^{(l)} M_{s,t}^{(l)} \right)^2}{N_s^{(l)} M_{s,t}^{(l)}} \xrightarrow{d} \chi_{d_{(s,l)} - 1}^2,$$

where  $d_{(s,l)}$  is the number of states  $t$  such that  $M_{s,t}^{(l)} > 0$ . There are  $dk$  such chi-square statistics in total.

Theorem 1 implies that the  $dk$  chi-square statistics are asymptotically independent. Then

$$\sum_{(s,l) \in \mathcal{S} \times \mathcal{A}, t \in \mathcal{S}} \frac{\left( N_{s,t}^{(l)} - N_s^{(l)} M_{s,t}^{(l)} \right)^2}{N_s^{(l)} M_{s,t}^{(l)}} \xrightarrow{d} \chi_{\sum_{(s,l) \in \mathcal{S} \times \mathcal{A}} d_{(s,l)} - dk}^2.$$

$\blacksquare$

## References

- Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.
- András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.
- Mor Armony, Shlomo Israelit, Avishai Mandelbaum, Yariv N Marmor, Yulia Tseytlin, and Galit B Yom-Tov. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5(1):146–194, 2015.
- Imon Banerjee, Harsha Honnappa, and Vinayak Rao. Off-line estimation of controlled markov chains: Minimaxity and sample complexity. *Operations Research*, 2025a.
- Imon Banerjee, Vinayak Rao, and Harsha Honnappa. Adaptive estimation of the transition density of controlled markov chains. *arXiv preprint arXiv:2505.14458*, 2025b.
- Yannick Baraud and Lucien Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probability Theory and Related Fields*, 143(1):239–284, 2009.
- Yannick Baraud and Lucien Birgé. Rho-estimators revisited: General theory and applications. *The Annals of Statistics*, 46(6B):3767–3804, 2018.
- Dimitri P Bertsekas. *Dynamic programming and optimal control, 3rd edition, volume II*. Belmont, MA: Athena Scientific, 2011.
- Patrick Billingsley. Statistical methods in Markov chains. *The Annals of Mathematical Statistics*, pages 12–40, 1961.
- Lucien Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l’IHP Probabilités et statistiques*, volume 42, pages 273–325, 2006.
- Vivek S Borkar. *Topics in controlled Markov chains*. Taylor & Francis, 1991.
- Irene Y Chen, Shalmali Joshi, Marzyeh Ghassemi, and Rajesh Ranganath. Probabilistic machine learning for healthcare. *Annual Review of Biomedical Data Science*, 4:393–415, 2021.
- Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pages 1042–1051. PMLR, 2019.
- Harald Cramér. *Mathematical Methods of Statistics*, volume 9 of *Princeton Mathematical Series*. Princeton University Press, 1946.
- Roland L Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability & Its Applications*, 1(1):65–80, 1956a.

- Roland L Dobrushin. Central limit theorem for nonstationary Markov chains. II. *Theory of Probability & Its Applications*, 1(4):329–383, 1956b.
- Wolfgang Doeblin. Sur deux problèmes de m. kolmogoroff concernant les chaînes dénombrables. *Bulletin de la Société Mathématique de France*, 66:210–220, 1938.
- Dmitry Dolgopyat and Omri Sarig. *Local Limit Theorems for Inhomogeneous Markov Chains*, volume 2331 of *Lecture Notes in Mathematics*. Springer International Publishing, Cham, 2023.
- Charles J Geyer. Markov chain monte carlo lecture notes. *Course notes, Spring Quarter*, 80, 1998.
- Josiah Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvari, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR, 2021.
- Onésimo Hernández-Lerma, Raúl Montes-de Oca, and Rolando Cavazos-Cadena. Recurrence conditions for Markov decision processes with Borel state space: a survey. *Annals of Operations Research*, 28(1):29–46, 1991.
- Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, and Sheila McIlraith. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2107–2116. PMLR, 2018.
- Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 652–661, New York, New York, USA, 2016. PMLR.
- Galin L Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.
- Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOREL: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823, 2020.
- Leonid Aryeh Kontorovich, Kavita Ramanan, and others. Concentration inequalities for dependent random variables via the martingale method. *Annals of Probability*, 36(6): 2126–2158, 2008.
- Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

- Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.
- Siqi Liu, Kay Choong See, Kee Yuan Ngiam, Leo Anthony Celi, Xingzhi Sun, Mengling Feng, et al. Reinforcement learning for clinical decision support in critical care: comprehensive review. *Journal of Medical Internet research*, 22(7):e18477, 2020.
- Lennart Ljung. *System identification (2nd ed.): theory for the user*. Prentice Hall PTR, NJ, USA, 1999.
- Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.
- Shie Mannor and John N Tsitsiklis. On the empirical state-action frequencies in Markov decision processes under general policies. *Mathematics of Operations Research*, 30(3):545–561, 2005.
- Florence Merlevède, Magda Peligrad, and Costel Peligrad. On the local limit theorems for lower psi-mixing Markov chains. *Latin American Journal of Probability and Mathematical Statistics*, 19(1):1103, 2022.
- Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.
- Farrukh Mukhamedov. The Dobrushin ergodicity coefficient and the ergodicity of noncommutative Markov chains. *Journal of Mathematical Analysis and Applications*, 408(1):364–373, 2013.
- Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.
- Mirco Mutti, Riccardo De Santi, and Marcello Restelli. The Importance of Non-Markovianity in Maximum State Entropy Exploration. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16223–16239. PMLR, 2022.
- Esa Nummelin. A splitting technique for harris recurrent markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 43(4):309–318, 1978.
- Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *ICML*, volume 2000, pages 759–766. Citeseer, 2000.
- M Rosenblatt-Roth. Some theorems concerning the law of large numbers for non-homogeneous Markoff chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1(5):433–445, 1963.
- M Rosenblatt-Roth. Some theorems concerning the strong law of large numbers for non-homogeneous Markov chains. *The Annals of Mathematical Statistics*, 35(2):566–576, 1964.
- Mathieu Sart. Estimation of the transition density of a Markov chain. In *Annales de l’IHP Probabilités et statistiques*, volume 50, pages 1028–1068, 2014.

- Henry Scheffé. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1-2):87–110, 1953.
- Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. *arXiv preprint arXiv:2202.13890*, 2022.
- Susan M Shortreed, Eric Laber, Daniel J Lizotte, T Scott Stroup, Joelle Pineau, and Susan A Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine Learning*, 84(1):109–136, 2011.
- Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.
- Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International conference on machine learning*, pages 2139–2148. PMLR, 2016.
- Philip Thomas, Georgios Theodorou, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.
- Samis Trevezas and Nikolaos Limnios. Variance estimation in the central limit theorem for markov chains. *Journal of Statistical Planning and Inference*, 139(7):2242–2253, 2009.
- L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
- Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. On the foundation of distributionally robust reinforcement learning. *arXiv preprint arXiv:2311.09018*, 2023.
- Geoffrey Wolfer and Aryeh Kontorovich. Estimating the Mixing Time of Ergodic Markov Chains. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 3120–3159. PMLR, 2019.
- Geoffrey Wolfer and Aryeh Kontorovich. Statistical estimation of ergodic Markov chain kernel over discrete state space. *Bernoulli*, 27(1):532–553, 2021.
- Jacob Wolfowitz. Products of indecomposable, aperiodic, stochastic matrices. *Proceedings of the American Mathematical Society*, 14(5):733–737, 1963.
- Sidney Yakowitz. Nonparametric estimation of Markov transition functions. *The Annals of Statistics*, pages 671–679, 1979.
- Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. Model-based reinforcement learning is minimax-optimal for offline zero-sum Markov games. *arXiv preprint arXiv:2206.04044*, 2022.

- Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.
- Jing Yu, Varun Gupta, and Adam Wierman. Online Adversarial Stabilization of Unknown Linear Time-Varying Systems. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 8320–8327, 2023.
- Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.
- Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. In *International Conference on Machine Learning*, pages 12287–12297. PMLR, 2021.
- Yi Zhu, Jing Dong, and Henry Lam. Uncertainty quantification and exploration for reinforcement learning. *Operations Research*, 72(4):1689–1709, 2024.