Model-based Bootstrap of Controlled Markov Chains

Ziwei Su
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
ziwei.su@northwestern.edu

Imon Banerjee
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
imon.banerjee@northwestern.edu

Diego Klabjan
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
d-klabjan@northwestern.edu

Abstract

We propose and analyze a model-based bootstrap for transition kernels in finite controlled Markov chains (CMCs) with possibly nonstationary or history-dependent control policies, a setting that arises naturally in offline reinforcement learning (RL) when the behavior policy generating the data is unknown. We establish distributional consistency of the bootstrap transition estimator in both a single long-chain regime and the episodic offline RL regime. The key technical tools are a novel bootstrap law of large numbers (LLN) for the visitation counts and a novel use of the martingale central limit theorem (CLT) for the bootstrap transition increments. We extend bootstrap distributional consistency to the downstream targets of offline policy evaluation (OPE) and optimal policy recovery (OPR) via the delta method by verifying Hadamard differentiability of the Bellman operators, yielding asymptotically valid confidence intervals for value and Q-functions. Experiments on the RiverSwim problem show that the proposed bootstrap confidence intervals (CIs), especially the percentile CIs, outperform the episodic bootstrap and plug-in CLT CIs, and are often close to nominal (50%, 90%, 95%) coverage, while the baselines are poorly calibrated at small sample sizes and short episode lengths.

1 Introduction

The bootstrap—first proposed by Efron [1979]—quickly became a cornerstone of modern statistical inference. Even when analytic variance formulas are available, bootstrap methods may deliver sharper finite-sample confidence intervals (CIs) with improved coverage; see Hall [2013, p. 14]. Despite these advantages, existing bootstrap approaches face two important limitations on reinforcement learning setting: (i) bootstrap methods for Markov transition matrices remain comparatively sparse, and (ii) existing approaches are not designed to handle non-stationarity that arises naturally in reinforcement learning. This limitation is particularly relevant in offline reinforcement learning (RL), where the control policy (behavior policy) that generates the offline data may be nonstationary and history-

Preprint.

dependent. To address this gap, we propose a new *model-based bootstrap* method for transition kernels in finite controlled Markov chains (CMCs) with possibly nonstationary or history-dependent control policies.

To set the stage, we introduce some notation. Assume that the state process $\{X_i\}_{i\geq 0}$ takes values in a finite state space $\mathcal{S}$ with $|\mathcal{S}| = S$, and the action process $\{A_i\}_{i\geq 0}$ takes values in a finite action space $\mathcal{A}$ with $|\mathcal{A}| = A$. A CMC is a paired state-action process $\big\{(X_i, A_i)\big\}_{i\geq 0}$; conditioned on A_i, the state sequence $\{X_i\}$ follows a Markov transition kernel M, where

$$M_{s,t}^{(a)} := \mathbb{P}\left(X_{i+1} = t \ \mid \ X_i = s, A_i = a\right)$$

for all time points i, $s, t \in \mathcal{S}$, and $a \in \mathcal{A}$ [Borkar, 1991]. We observe an offline dataset $\mathcal{D}_n = \big\{(X_i^{(k)}, A_i^{(k)}, X_{i+1}^{(k)}) : k = 1, \dots, K,\ i = 0, \dots, T-1\big\}$ of K episodes generated by an unknown behavior policy π_b, each of length T, with $n = KT$. Note that K is allowed to be 1 and the policy π_b may be nonstationary and history-dependent.

We compute from $\mathcal{D}_n$ the counts $N_{s,t}^{(a)} := \sum_{i,k} \mathbb{1}\big\{X_i^{(k)} = s,\ A_i^{(k)} = a,\ X_{i+1}^{(k)} = t\big\}$, $N_s^{(a)} := \sum_t N_{s,t}^{(a)}$, and $N_s := \sum_a N_s^{(a)}$. We use the counts to compute the count-based empirical transition estimator $\hat{M}_{s,t}^{(a)} := N_{s,t}^{(a)}/N_s^{(a)}$ and the empirical behavior policy $\hat{\pi}_b(a \mid s) := N_s^{(a)}/N_s$.

Model-based Bootstrap: We now describe the proposed model-based bootstrap method. Denote the block matrices stacking the entries $M_{s,t}^{(a)}$ and $\hat{M}_{s,t}^{(a)}$ by $\mathbf{M}$ and $\hat{\mathbf{M}}$, respectively. The bootstrap CMC $\{(X_i^*, A_i^*)\}_{i\geq 0}$ has the empirical transition kernel $\hat{M}$ and the empirical behavior policy $\hat{\pi}_b$. We generate the bootstrap dataset $\mathcal{D}_n^* = \big\{(X_i^{*(k)}, A_i^{*(k)}, X_{i+1}^{*(k)}) : k = 1, \dots, K,\ i = 0, \dots, T - 1\big\}$ from the bootstrap CMC with the same episodic structure as $\mathcal{D}_n$, and compute the bootstrap transition estimator $\hat{M}_{s,t}^{*(a)} := N_{s,t}^{*(a)}/N_s^{*(a)}$ from $\mathcal{D}_n^*$, where $N_{s,t}^{*(a)} := \sum_{i,k} \mathbb{1}\big\{X_i^{*(k)} = s,\ A_i^{*(k)} = a,\ X_{i+1}^{*(k)} = t\big\}$ and $N_s^{*(a)} := \sum_t N_{s,t}^{*(a)}$. We denote the block transition matrix stacking the entries $\hat{M}_{s,t}^{*(a)}$ by $\hat{\mathbf{M}}^*$. We repeat the bootstrap B times independently to obtain B bootstrap replicates $\hat{\mathbf{M}}^{*(1)}, \dots, \hat{\mathbf{M}}^{*(B)}$. The total bootstrap computational cost is $O(Bn) = O(BKT)$, linear in dataset size n and number of bootstrap replicates B.

We describe two downstream applications of offline policy evaluation (OPE) and optimal policy recovery (OPR) in offline infinite-horizon discounted RL. In OPE, for a target policy π, we use $\hat{\mathbf{M}}$ in the Bellman equations to obtain plug-in estimators $\hat{V}_\pi$ and $\hat{Q}_\pi$ of the value function V_π and action-value function Q_π, respectively. We then substitute $\hat{\mathbf{M}}^{*(j)}$ in place of $\hat{\mathbf{M}}$ to obtain bootstrap counterparts $\hat{V}_\pi^{*(j)}$ and $\hat{Q}_\pi^{*(j)}$. In OPR, we solve the optimal Bellman equations under $\hat{\mathbf{M}}$ to obtain the plug-in estimators $\hat{V}_\star$ and $\hat{Q}_\star$ of the optimal value function $V_\star$ and optimal action-value function $Q_\star$, respectively. We then substitute $\hat{\mathbf{M}}^{*(j)}$ in place of $\hat{\mathbf{M}}$ to obtain bootstrap counterparts $\hat{V}_\star^{*(j)}$ and $\hat{Q}_\star^{*(j)}$. We compute $100(1-\alpha)\%$ CIs for V_π, Q_π, or $V_\star$, $Q_\star$ from the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of their B bootstrap replicates.

The focus of this paper is to establish distributional consistency of $\hat{\mathbf{M}}^*$ in two settings: the single-chain regime ($K = 1$, $T \to \infty$) in Theorem 1, and the episodic offline-RL regime (T fixed, $K \to \infty$) in Proposition 1. In both settings, conditional on $\mathcal{D}_n$, $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}})$ converges in distribution to the same Gaussian limit as $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}} - \mathbf{M})$. Proposition 2 extends this to OPE and OPR targets via the delta method. Su et al. [2026] provides explicit formulas for various distributional limits. We validate the finite-sample performance of our proposed model-based bootstrap CIs on the RiverSwim problem, where the proposed model-based bootstrap percentile CIs achieve near-nominal coverage while baselines systematically undercover.

We now briefly summarize the key contributions of this paper.

- **Bootstrap distributional consistency.** To the best of our knowledge, we establish the first distributional consistency result for the proposed model-based bootstrap in finite CMCs with possibly nonstationary or history-dependent control policies. Theorem 1 covers the single-chain regime ($K = 1$, $T \to \infty$) and Proposition 1 covers the episodic offline-RL regime (T fixed, $K \to \infty$). In both regimes, we establish that conditional on $\mathcal{D}_n$, $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}})$

converges in distribution to the same Gaussian law as $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}} - \mathbf{M})$. The proof hinges on a novel bootstrap law of large numbers (LLN) $N_s^{*(a)}/n \to p_s^{(a)}$ a.s. and a novel use of the martingale CLT for the bootstrap transition increments. (See Section 3.2.)

- **Bootstrap validity for OPE and OPR.** In Proposition 2, we are the first one to extend bootstrap distributional consistency to the downstream RL targets V_π, Q_π, $V_\star$ and $Q_\star$ with non-stationary or history-dependent behavior policies. For OPE, we establish that conditional on $\mathcal{D}_n$, $\sqrt{n}(\hat{V}_\pi^* - \hat{V}_\pi)$ and $\sqrt{n}(\hat{Q}_\pi^* - \hat{Q}_\pi)$ converge in distribution to the same Gaussian limits as $\sqrt{n}(\hat{V}_\pi - V_\pi)$ and $\sqrt{n}(\hat{Q}_\pi - Q_\pi)$, respectively. For OPR, we establish that conditional on $\mathcal{D}_n$, $\sqrt{n}(\hat{V}_\star^* - \hat{V}_\star)$ converges in distribution to the same Gaussian limit as $\sqrt{n}(\hat{V}_\star - V_\star)$. The proof exploits the delta method by verifying Hadamard differentiability (see Appendix D.2) of the Bellman operators with respect to the transition block matrix. (See Section 3.3.)
- **Finite-sample coverage on RiverSwim.** Proposition 2 guarantees the asymptotic validity of the proposed model-based bootstrap CIs for OPE and OPR. In the setting of Zhu et al. [2024], we validate the model-based bootstrap CIs empirically on the RiverSwim problem [Strehl and Littman, 2004]. In a coverage study with $B = 1{,}000$ bootstrap replicates and $N_{\mathrm{reps}} = 1{,}000$ Monte Carlo replications, the model-based bootstrap CIs, especially the percentile CI, substantially outperform the episodic bootstrap and plug-in CLT CIs across the OPE and OPR tasks considered. The percentile CI attains near-nominal coverage at nominal 50% (about 0.48–0.54), 90% (about 0.87–0.94), and 95% (about 0.92–0.97) in the better-sampled regimes, notably for OPE and for OPR with $T \in \{50, 100\}$ at $n \geq 500$. (See Section 4.)

The rest of the paper is organized as follows. Section 2 reviews background and related work. Section 3 formalizes the problem setup and states the main theoretical results. Section 4 presents the numerical experiments. Section 5 concludes the paper.

2 Related Work

The practice of using CLTs for statistical inference, such as for Markov chains and controlled Markov chains [Billingsley, 1961, Zhu et al., 2024, Su et al., 2026], is well established in the literature. These plug-in CIs, however, require covariance estimation, and the resulting errors degrade finite-sample CI coverage. The bootstrap addresses this: the intuitive appeal, computational simplicity, and better performance in complex models or nonstandard error structure of the bootstrap [Hall and Padmanabhan, 1992] lead to its widespread use in modern statistical inference [Modayil and Kuipers, 2004, Han and Liu, 2016, Nakkiran et al., 2020]. The applications of the bootstrap span sparse principal component analysis [Babamoradi et al., 2013, Rahoma et al., 2021], time-series forecasting [Ruiz and Pascual, 2002], econometric and financial inference [Horowitz, 2019, Gonçalves et al., 2023], and clustering and classification [Jedra et al., 2023, Jain and Moreau, 1987].

Directly relevant to this paper, bootstrap methods are widely applied in Markov chains [Kulperger and Rao, 1989, Athreya and Fuh, 1992]. Closest in spirit to this work, Horowitz [2003] introduces a model-based bootstrap for stationary Markov chains by resampling from an estimated transition kernel under geometric mixing and a Cramér-type moment condition, while Bertail and Clémençon [2006] develop a regenerative block bootstrap for Harris recurrent chains via Nummelin splitting, requiring explicit identification of a minorization condition and an accessible small set. However, these methods do not provide guarantees for the finite CMC setting with possibly nonstationary or history-dependent control policies.

RL is another domain where bootstrap methods are widely applied [Krishnamurthy, 2016, Faradonbeh et al., 2019, Ramprasad et al., 2023, Banerjee, 2023, Banerjee et al., 2025, Banerjee and Chakrabarty, 2025]. For OPE, Kostrikov and Nachum [2020] apply Efron's bootstrap to empirical transition tuples treated as i.i.d. draws from a stationary occupancy distribution. This requires a stationary behavior policy and ignores within-episode dependence. Hanna et al. [2017] propose model-based bootstrap methods that learn transition dynamics from offline data, simulate trajectories under the target policy using the learned model, and bootstrap trajectory-level cumulative rewards to construct high-confidence lower bounds on a policy value. This approach does not establish distributional consistency of the bootstrap estimator. Hao et al. [2021] establishes bootstrap validity for a finite-horizon episodic bootstrap via fitted Q-evaluation, under the assumption that a large number of independent i.i.d. episodes are available. This assumption does not cover the finite CMC setting with possibly nonstationary or history-dependent control policies, and the episodic bootstrap can be poorly calibrated when the number of episodes is small, which brings us to the following open problems.

Open problems. We address two open problems in the theory of *bootstrap for finite CMCs:* **(i)** Is it possible to develop a model-based bootstrap for finite CMCs that does not require stationarity or Markovianity of the control policy? **(ii)** Can we then establish distributional consistency of the bootstrap transition estimator, then lift this result to common RL tasks like OPE and OPR?

We move on to formally state our main results.

3 Main Theoretical Results

The goal of this section is to establish our main results: distributional consistency of the model-based bootstrap in finite CMCs for the bootstrap transition estimator $\hat{M}^*$ and, via the delta method, for the bootstrap plug-in value and Q-function estimators.

In order to formalize our results, we introduce some additional notation. All random variables are defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, where $\mathbb{F} := \{\mathcal{F}_j\}_{j\geq 0}$, $\mathcal{F}_j := \sigma(H_0^j)$ is a filtration with $\mathcal{F}_j \subset \mathcal{F}$, and $H_p^j := \{(X_i, A_i)\}_{i=p}^{j}$ denotes the history from time p to j. The transition kernel M satisfies the Markov property $M_{s_i,s_{i+1}}^{(a_i)} = \mathbb{P}\left(X_{i+1} = s_{i+1} \mid X_i = s_i, A_i = a_i\right) = \mathbb{P}\left(X_{i+1} = s_{i+1} \mid H_0^i = h_0^i\right)$, where $h_0^i := \{X_0 = s_0, A_0 = a_0, \ldots, X_i = s_i, A_i = a_i\}$ is the sample history up to time i. For each $s \in \mathcal{S}$, let $M_s := [M_{s,t}^{(a)}]_{t\in\mathcal{S},\, a\in\mathcal{A}} \in \mathbb{R}^{S\times A}$, and let $\mathbf{M} := [M_1, \ldots, M_S]^\top \in \mathbb{R}^{SA\times S}$ be the block transition matrix stacking $M_{s,t}^{(a)}$ entries. The empirical block transition matrix is $\hat{\mathbf{M}} := [\hat{M}_1, \ldots, \hat{M}_S]^\top \in \mathbb{R}^{SA\times S}$, where $\hat{M}_s := [\hat{M}_{s,t}^{(a)}]_{t\in\mathcal{S},\, a\in\mathcal{A}}$.

We now turn to the formal statements of our main results. We begin by introducing some standard assumptions.

3.1 Assumptions

The three assumptions below control, respectively, the frequency of return visits to each state–action pair, the rate of mixing of the paired state–action process $\{(X_i, A_i)\}_{i\geq 0}$, and the positivity of the limiting state–action visitation frequencies. Together they underpin the CLT for the empirical transition estimator established by Su et al. [2026] and ensure that the bootstrap chain inherits the same asymptotic properties.

For every $(s, a) \in \mathcal{S}\times\mathcal{A}$, we recursively define the hitting and return times to state–action pair (s, a) as follows.

Definition 1 (Hitting and Return Times)**.** *The first hitting time to (s, a) is $\tau_{s,a}^{(1)} := \min\{i > 0 : X_i = s,\ A_i = a\}$. For $i \geq 2$, the i-th return time is*

$$\tau_{s,a}^{(i)} := \min\Big\{k > 0 : X_{\sum_{j=1}^{i-1}\tau_{s,a}^{(j)}+k} = s,\ A_{\sum_{j=1}^{i-1}\tau_{s,a}^{(j)}+k} = a\Big\}.$$

Assumption 1 requires the conditional expected return time to grow at most sublinearly in the visit index.

Assumption 1 (Sublinear Return-Time Growth)**.** *For any $(s, a) \in \mathcal{S}\times\mathcal{A}$, there exists $\beta \in (0, 1)$ such that $\mathbb{E}\Big[\tau_{s,a}^{(i)} \,\Big|\, \mathcal{F}_{\sum_{p=1}^{i-1}\tau_{s,a}^{(p)}}\Big] \leq T_i$ a.s., $T_i = O(i^\beta)$.*

Our next assumption is on the rate of mixing. We define the weak $\bar{\eta}$-mixing coefficient for the CMC $\{(X_i, A_i)\}$ as follows.

Definition 2 (Weak $\bar{\eta}$-Mixing Coefficient)**.** *For $i < j \leq n$, $\mathcal{T} \subseteq (\mathcal{S}\times\mathcal{A})^{n-j+1}$, $s_1, s_2 \in \mathcal{S}$, and $a_1, a_2 \in \mathcal{A}$, let*

$$\eta_{i,j}(\mathcal{T}, s_1, s_2, a_1, a_2, h_0^{i-1}, n) := \Big|\mathbb{P}\big((X_j, A_j, \ldots, X_n, A_n) \in \mathcal{T} \mid X_i = s_1, A_i = a_1, H_0^{i-1} = h_0^{i-1}\big) - \mathbb{P}\big((X_j, A_j, \ldots, X_n, A_n) \in \mathcal{T} \mid X_i = s_2, A_i = a_2, H_0^{i-1} = h_0^{i-1}\big)\Big|.$$

The weak $\bar{\eta}$-mixing coefficient for the CMC is

$$\bar{\eta}_{i,j}(n) := \sup_{\substack{\mathcal{T}, s_1, s_2, a_1, a_2, h_0^{i-1}, \\ \mathbb{P}\left(X_i = s_1, A_i = a_1, H_0^{i-1} = h_0^{i-1}\right) > 0, \\ \mathbb{P}\left(X_i = s_2, A_i = a_2, H_0^{i-1} = h_0^{i-1}\right) > 0}} \eta_{i,j}(\mathcal{T}, s_1, s_2, a_1, a_2, h_0^{i-1}, n).$$

We omit the dependence on n in $\bar{\eta}_{i,j}$ by convention. The coefficient $\bar{\eta}_{i,j}$ quantifies the worst-case influence of the state–action pair at time i on the trajectory distribution from time j onward. Assumption 2 imposes a uniform bound on the cumulative decay of $\bar{\eta}$-mixing coefficients. Sufficient conditions for Assumption 2 in terms of separate mixing coefficients for the state and action processes are given in Appendix A.

Assumption 2 (Weak Mixing)**.** *There exists a finite constant C_Δ such that $\|\Delta_n\| := \max_{1 \le i \le n}(1 + \bar{\eta}_{i,i+1} + \cdots + \bar{\eta}_{i,n}) \le C_\Delta$ for all n.*

We next define the ergodic occupation measure of the CMC as the Cesàro limit of the marginal state–action distributions.

Definition 3 (Ergodic Occupation Measure)**.** *The ergodic occupation measure of the CMC is $p_s^{(a)} := \lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{P}(X_i = s,\ A_i = a)$, whenever the limit exists for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.*

Assumption 3 requires the ergodic occupation measure to exist and be uniformly bounded away from zero.

Assumption 3 (Positive Ergodic Occupation Measure)**.** *The CMC admits an ergodic occupation measure, and $\min_{s,a} p_s^{(a)} \ge p_0 > 0$.*

3.2 Bootstrap Distributional Consistency

We now state our main results. Theorem 1 establishes bootstrap distributional consistency for the single long-chain setting; the episodic extension follows as Proposition 1.

Theorem 1 (Single Chain Bootstrap Distributional Consistency)**.** *Let $\bar{\xi} := \sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}})$ be the vector of differences between $\hat{\mathbf{M}}^*$ and $\hat{\mathbf{M}}$. Suppose $\mathcal{D}_n = \{(X_0, A_0, X_1), \ldots, (X_{n-1}, A_{n-1}, X_n)\}$ is a single trajectory of length n satisfying Assumptions 1–3, and there exists $(s_0, a_0) \in \mathcal{S} \times \mathcal{A}$ with $M_{s_0,s_0}^{(a_0)} > 0$. Then, for $\mathbb{P}$-almost every sequence of datasets $(\mathcal{D}_n)_{n\ge1}$, conditional on $\mathcal{D}_n$,*

$$\bar{\xi} = \sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}}) \ \xrightarrow{\mathrm{d}}\ \mathcal{N}(0, \bar{\Lambda}), \textit{ as } n \to \infty.$$

The elements of $\bar{\Lambda}$ are given by $\bar{\Lambda}_{sat,\,s'a't'}$, which denotes the covariance between the $(s-1)S \cdot A + (a-1)S + t$-th and the $(s'-1)S \cdot A + (a'-1)S + t'$-th elements of $\bar{\xi}$. Value $\bar{\Lambda}_{sat,\,s'a't'}$ has the expression $\bar{\Lambda}_{sat,\,s'a't'} := \mathbb{1}\{(s,a) = (s',a')\}\frac{1}{p_s^{(a)}}\left(\mathbb{1}\{t = t'\}\,M_{s,t}^{(a)} - M_{s,t}^{(a)} M_{s,t'}^{(a)}\right)$. Matrix $\bar{\Lambda}$ is the asymptotic covariance of $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}} - \mathbf{M})$ given in Su et al. [2026, Corollary 1].

Proof sketch. The complete proof is given in Appendix C; the supporting lemmas are collected in Appendix B. We sketch the three main steps here.

Step 1 (Bootstrap LLN). The pivotal intermediate result, established formally as Lemma 7 in Appendix B.5, is $N_s^{*(a)}/n \xrightarrow{\text{a.s.}} p_s^{(a)}$. We write $\frac{N_s^{*(a)}}{n} = \frac{N_s^{*(a)}}{\mathbb{E}[N_s^{*(a)} \mid \mathcal{D}_n]} \cdot \frac{\mathbb{E}[N_s^{*(a)} \mid \mathcal{D}_n]}{n}$. We show that the first term converges to 1 a.s. This requires $\|\Delta_n^*\|$ to be eventually bounded, which we establish in Lemma 5 in Appendix B.3. The second term converges to $p_s^{(a)}$ a.s. by Lemma 6 in Appendix B.4.

Step 2 (Decomposition). For each (s,a,t), we write $\sqrt{n}(\hat{M}_{s,t}^{*(a)} - \hat{M}_{s,t}^{(a)}) = S_n^{*,sat} + R_n^{*,sat}$, where

$$S_n^{*,sat} := \frac{1}{\sqrt{n}\,p_s^{(a)}} \sum_{i=0}^{n-1}\left(\mathbb{1}\{X_i^* = s, A_i^* = a, X_{i+1}^* = t\} - \hat{M}_{s,t}^{(a)}\mathbb{1}\{X_i^* = s, A_i^* = a\}\right)$$

is the leading martingale sum, and

$$R_n^{*,sat} := \left(\frac{n}{N_s^{*(a)}} - \frac{1}{p_s^{(a)}}\right)\frac{1}{\sqrt{n}} \sum_{i=0}^{n-1}\left(\mathbb{1}\{X_i^* = s, A_i^* = a, X_{i+1}^* = t\} - \hat{M}_{s,t}^{(a)}\mathbb{1}\{X_i^* = s, A_i^* = a\}\right).$$

By Step 1, $n/N_s^{*(a)} - 1/p_s^{(a)} \xrightarrow{\text{a.s.}} 0$. We show that the sum in $R_n^{*,sat}$ has conditional variance $O(n)$, and divided by $\sqrt{n}$ it is $O_{\mathbb{P}}(1)$. By the Slutsky's theorem, $R_n^{*,sat} = o_{\mathbb{P}}(1)$, and it suffices to analyze $S_n^{*,sat}$.

Step 3 (Conditional Martingale CLT). We show that the sum $S_n^{*,sat}$ is a martingale difference array under joint filtration $\mathcal{F}_{n,i} := \sigma(\mathcal{D}_n, X_0^*, A_0^*, \dots, X_i^*, A_i^*)$ and $\mathbb{P}$. Then, we show that the Lindeberg condition holds for all large enough n, the predictable quadratic variation converges a.s. to $\bar{\Lambda}_{sat,sat}$ via Lemma 7, and $\hat{M}_{s,t}^{(a)} \to M_{s,t}^{(a)}$ a.s. We then apply the martingale CLT [Hall and Heyde, 1980, Corollary 3.1] conditioned on $\mathcal{D}_n$, yielding $S_n^{*,sat} \xrightarrow{\text{d}} \mathcal{N}(0, \bar{\Lambda}_{sat,sat})$. The multivariate statement follows by the Cramér–Wold theorem [Cramér and Wold, 1936]. □

We now move to the episodic setting. Offline RL datasets are almost always episode-structured, arising from an environment that resets between episodes under a behavior policy that may be nonstationary or history-dependent across episodes. Theorem 1 does not directly cover this structure, as the episodes are separated by resets and do not form a single uninterrupted CMC.

Here, we observe an offline dataset $\mathcal{D}_n = \{(X_i^{(k)}, A_i^{(k)}, X_{i+1}^{(k)}) : k = 1, \dots, K,\ i = 0, \dots, T-1\}$ of K episodes of fixed length T with $n = KT$. To apply Theorem 1, we embed the episodic data into a single chain by inserting, between episode k and episode $k+1$, a pseudo-transition under a dummy reset action $a_\dagger \notin \mathcal{A}$ that maps $X_T^{(k)}$ to the observed initial state $X_0^{(k+1)}$ of the next episode. The extended action space is $\widetilde{\mathcal{A}} = \mathcal{A} \cup \{a_\dagger\}$, and we use $\widetilde{M}$ to denote the resulting transition kernel on $\mathcal{S} \times \widetilde{\mathcal{A}}$. We note that no structural restriction is imposed linking $X_0^{(k+1)}$ to $X_T^{(k)}$.

Formally, let $\tilde{\mathcal{D}} := \{(\widetilde{X}_i, \widetilde{A}_i, \widetilde{X}_{i+1})\}_{i=0}^{n'-1}$, with $n' = K(T+1) - 1$, be the concatenated single-chain dataset in the format of Theorem 1, where $\widetilde{X}_i \in \mathcal{S}$ and $\widetilde{A}_i \in \widetilde{\mathcal{A}}$ are given by

$$
\begin{aligned}
\widetilde{X}_{(k-1)(T+1)+j} &:= X_j^{(k)}, \quad k = 1, \dots, K,\ j = 0, \dots, T; \\
\widetilde{A}_{(k-1)(T+1)+j} &:= A_j^{(k)}, \quad k = 1, \dots, K,\ j = 0, \dots, T-1; \\
\widetilde{A}_{k(T+1)-1} &:= a_\dagger, \quad k = 1, \dots, K-1.
\end{aligned}
$$

The paired process $\{(\widetilde{X}_i, \widetilde{A}_i)\}_{i \geq 0}$ forms a CMC on $\mathcal{S} \times \widetilde{\mathcal{A}}$ with transition kernel $\widetilde{M}$, where $\widetilde{M}_{s,t}^{(a)} = M_{s,t}^{(a)}$ for $a \in \mathcal{A}$, and we can take $\widetilde{M}^{(a_\dagger)}$ as any stochastic kernel on $\mathcal{S}$ such that $\widetilde{M}_{s,t}^{(a_\dagger)} > 0$ for all $s, t \in \mathcal{S}$, since $a_\dagger$-transitions are excluded from $\hat{\mathbf{M}}$.

The following proposition extends Theorem 1 to the episodic setting.

Proposition 1. *Given $\mathcal{D}_n = \{(X_i^{(k)}, A_i^{(k)}, X_{i+1}^{(k)})\}_{i \in \{0, \dots, T-1\}}$, suppose the CMC $\{(\widetilde{X}_i, \widetilde{A}_i)\}$ with transition kernel $\widetilde{M}$ satisfies Assumptions 1–3 on $\mathcal{S} \times \widetilde{\mathcal{A}}$, and there exists $(s_0, a_0) \in \mathcal{S} \times \mathcal{A}$ with $M_{s_0,s_0}^{(a_0)} > 0$. Then, for $\mathbb{P}$-almost every sequence of datasets $(\mathcal{D}_n)_{n \geq 1}$, conditional on $\mathcal{D}_n$,*

$$
\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}}) \xrightarrow{\text{d}} \mathcal{N}(0, \bar{\Lambda}), \text{ as } K \to \infty \text{ and } T \text{ fixed.}
$$

The conditions on $\{(\widetilde{X}_i, \widetilde{A}_i)\}$ are satisfied whenever the individual episodes satisfy Assumptions 1–3. More generally, they accommodate policy dependence across episodes, since they are imposed on the concatenated chain rather than on individual episodes. The proof of this proposition can be found in Appendix D.1.

Before moving to downstream RL applications, we note two important considerations in the model-based bootstrap construction. First, we impose the assumptions of Proposition 1 with respect to the original observed data-generating process, not with respect to a fitted reset mechanism. Thus, the model-based bootstrap does not require estimating or modeling the initial state (reset) distribution. We therefore initialize each bootstrap episode by setting $X_0^{*(k)} := X_0^{(k)}$, copying the observed start states directly from $\mathcal{D}_n$. This is important because, in the episodic setting, the reset mechanism may be nonstationary or history-dependent and is not identifiable from D_n alone. Proposition 1 confirms that this choice does not affect bootstrap distributional consistency.

Second, we estimate the behavior policy by the one-step empirical estimator $\hat{\pi}_b(a \mid s) = N_s^{(a)}/N_s$. Theorem 1 and Proposition 1 confirm this one-step estimate suffices even when π_b is nonstationary or history-dependent, as the asymptotic distributions depend on the ergodic occupation measure $p_s^{(a)}$ induced by π_b rather than its history-dependent structure.

We now move to the downstream RL applications of Theorem 1 and Proposition 1.

3.3 Applications to OPE and OPR

We consider downstream RL tasks (OPE and OPR) under the offline infinite-horizon discounted RL setting with a known reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. OPE concerns using $\mathcal{D}_n$ to estimate the value function V_π and Q-function Q_π for a fixed stationary target policy π. Let $\Pi = \mathrm{diag}(\pi_1, \ldots, \pi_S) \in \mathbb{R}^{S \times SA}$ denote the associated policy matrix, where $\pi_s := [\pi(s,1), \ldots, \pi(s,A)]$. Vectors V_π and Q_π are the unique solutions to the Bellman equations [Bertsekas, 2011, Agarwal et al., 2019]

$$V_\pi = (I - \gamma \Pi \mathbf{M})^{-1} g, \qquad Q_\pi = (I - \gamma \mathbf{M} \Pi)^{-1} r, \tag{3.1}$$

where $0 < \gamma < 1$ is the discount factor, $g(s) := \sum_a \pi(s,a) r(s,a)$, $g := (g(s) : s \in \mathcal{S})$, and $r := (r(s,a) : s \in \mathcal{S}, a \in \mathcal{A})$.

OPR concerns using $\mathcal{D}_n$ to estimate the optimal policy $\pi_\star$, optimal value function $V_\star$ and optimal Q-function $Q_\star$, which are the unique solutions to the optimal Bellman equations

$$Q_\star(s,a) = r(s,a) + \gamma \sum_{t \in \mathcal{S}} M_{s,t}^{(a)} \max_{a' \in \mathcal{A}} Q_\star(t, a'), \qquad V_\star(s) = \max_{a \in \mathcal{A}} Q_\star(s,a), \tag{3.2}$$

and the optimal policy satisfies $\pi_\star(s) \in \arg\max_{a \in \mathcal{A}} Q_\star(s,a)$ for each $s \in \mathcal{S}$. We obtain plug-in estimators $\hat{V}_\pi$ and $\hat{Q}_\pi$ for OPE by replacing $\mathbf{M}$ with $\hat{\mathbf{M}}$ in (3.1):

$$\hat{V}_\pi = (I - \gamma \Pi \hat{\mathbf{M}})^{-1} g, \qquad \hat{Q}_\pi = (I - \gamma \hat{\mathbf{M}} \Pi)^{-1} r. \tag{3.3}$$

For OPR, we obtain plug-in estimators $\hat{Q}_\star$, $\hat{V}_\star$, and $\hat{\pi}_\star$ by solving (3.2) with $M_{s,t}^{(a)}$ replaced by $\hat{M}_{s,t}^{(a)}$.

We use the bootstrap replicates $\hat{\mathbf{M}}^{*(1)}, \ldots, \hat{\mathbf{M}}^{*(B)}$ to obtain bootstrap replicates for OPE and OPR targets as follows. In the j-th replicate, we compute bootstrap plug-in value and Q-function estimators $\hat{V}_\pi^{*(j)}$ and $\hat{Q}_\pi^{*(j)}$ for OPE by replacing $\mathbf{M}$ with $\hat{\mathbf{M}}^{*(j)}$ in (3.1):

$$\hat{V}_\pi^{*(j)} = (I - \gamma \Pi \hat{\mathbf{M}}^{*(j)})^{-1} g, \qquad \hat{Q}_\pi^{*(j)} = (I - \gamma \hat{\mathbf{M}}^{*(j)} \Pi)^{-1} r. \tag{3.4}$$

For OPR, we obtain bootstrap plug-in estimators $\hat{Q}_\star^{*(j)}$, $\hat{V}_\star^{*(j)}$, and $\hat{\pi}_\star^{*(j)}$ by solving (3.2) with $M_{s,t}^{(a)}$ replaced by $\hat{M}_{s,t}^{*(a,j)}$. We repeat this process independently for $j = 1, \ldots, B$ to obtain B bootstrap replicates $\big\{\hat{V}_\pi^{*(j)}, \hat{Q}_\pi^{*(j)}, \hat{V}_\star^{*(j)}, \hat{Q}_\star^{*(j)}\big\}_{j=1}^B$.

We construct CIs for any coordinate of V_π from the B bootstrap replicates as follows. Recall that we denote the plug-in estimator by $\hat{V}_\pi$, and the bootstrap replicates by $\hat{V}_\pi^{*(1)}, \ldots, \hat{V}_\pi^{*(B)}$. We further denote the empirical α-quantile of $\{\hat{V}_\pi^{*(j)}(s)\}_{j=1}^B$ for some $s \in \mathcal{S}$ by q_α^*. The $100(1-\alpha)\%$ bootstrap percentile CI for $V_\pi(s)$ is $\big[q_{\alpha/2}^*,\ q_{1-\alpha/2}^*\big]$. The bootstrap pivot CI is $\big[2\hat{V}_\pi(s) - q_{1-\alpha/2}^*,\ 2\hat{V}_\pi(s) - q_{\alpha/2}^*\big]$. The CI constructions for Q_π, $V_\star$ and $Q_\star$ are analogous. Constructing CIs for OPE and OPR targets requires solving the Bellman equation for an additional $O(B)$ times, negligible compared to the $O(Bn)$ bootstrap simulation cost.

We now state the bootstrap distributional consistency results for the plug-in value and Q-function estimators for OPE and OPR, both in the single-chain and episodic settings. We define the asymptotic covariance matrices

$$\begin{aligned}
\Sigma_V^\pi &:= \gamma^2 \big[(I - \gamma \Pi \mathbf{M})^{-1} \Pi \otimes V_\pi^\top\big] \bar{\Lambda} \big[\Pi^\top (I - \gamma \mathbf{M}^\top \Pi^\top)^{-1} \otimes V_\pi\big], \\
\Sigma_Q^\pi &:= \gamma^2 \big[(I - \gamma \mathbf{M} \Pi)^{-1} \otimes Q_\pi^\top \Pi^\top\big] \bar{\Lambda} \big[(I - \gamma \Pi^\top \mathbf{M}^\top)^{-1} \otimes \Pi Q_\pi\big].
\end{aligned}$$

We further define $\Pi_\star$ as the policy matrix associated with $\pi_\star$, and π' as any policy with policy matrix Π'. The following proposition then establishes the bootstrap distributional consistency results to the plug-in value and Q-function estimators for OPE and OPR, both in the single-chain and episodic settings.

Proposition 2 (Bootstrap Validity for OPE and OPR)**.** *Under the conditions of Theorem 1 (single-chain) or Proposition 1 (episodic), for any fixed target policy* π*,*

$$\sqrt{n}\big(\hat{V}^*_\pi - \hat{V}_\pi\big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma^\pi_V), \qquad \sqrt{n}\big(\hat{Q}^*_\pi - \hat{Q}_\pi\big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma^\pi_Q),$$

conditionally on $\mathcal{D}_n$*, in probability, as* $n \to \infty$*. Matrices* Σ^π_V *and* Σ^π_Q *are the asymptotic covariances of* $\sqrt{n}(\hat{V}_\pi - V_\pi)$ *and* $\sqrt{n}(\hat{Q}_\pi - Q_\pi)$ *given in Su et al. [2026, Theorem 2], and* $\Sigma^{\pi_\star}_V$ *is the asymptotic covariance of* $\sqrt{n}(\hat{V}_\star - V_\star)$ *given in Su et al. [2026, Theorem 3].*

In addition, if the following non-degeneracy condition

$$\min_{s\in\mathcal{S}} \min_{a\neq\pi_\star(s)} \big\{Q_\star(s, \pi_\star(s)) - Q_\star(s, a)\big\} > 0$$

holds, then we have

$$\sqrt{n}\big(\hat{V}^*_\star - \hat{V}_\star\big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma^{\pi_\star}_V), \qquad \sqrt{n}\big(\hat{Q}^*_\star - \hat{Q}_\star\big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma^{\pi_\star}_Q),$$

conditionally on $\mathcal{D}_n$*, in probability, as* $n \to \infty$*.*

This proposition guarantees that percentile and pivot bootstrap CIs are asymptotically valid for V_π, Q_π, $V_\star$ and $Q_\star$. The proof of this proposition can be found in Appendix D.2.

4 Numerical Experiments

Following existing benchmarks, which broadly focus on the stationary setting, we replicate the small state-space RiverSwim experiment of Zhu et al. [2024] to evaluate finite-sample CI coverage for the proposed model-based bootstrap. The RiverSwim MDP has $S = 6$ states and $A = 2$ actions, with rewards $r(1, 0) = 1$, $r(6, 1) = 10$, and $r(s, a) = 0$ otherwise, and discount factor $\gamma = 0.95$. Figure 1 in Appendix E illustrates the transition and reward structure of the RiverSwim setup. The experiment is deliberately challenging for inference because the behavior policy visits upstream states (near state 6) rarely, while the optimal policy is nearly degenerate at the left-most state 1.

We use the stationary behavior policy $\pi_b(1 \mid s) = 0.8$ for all states, $B = 1{,}000$ bootstrap replicates, and $N_{\mathrm{reps}} = 1{,}000$ Monte Carlo replications. For OPE, we evaluate three fixed target policies including uniform ($\pi(1 \mid s) = 0.5$), mostly-right ($\pi(1 \mid s) = 0.8$), and mostly-left ($\pi(1 \mid s) = 0.2$) policies for horizon length $T = 50$. For OPR, we evaluate the optimal value and action-value targets across horizons $T \in \{10, 50, 100\}$. We do not evaluate all target policies with different episode lengths in the OPE study to keep the numerical results interpretable and tractable. Table 1 in Appendix E summarizes the settings of all experiments. We compare percentile and pivot CIs from the proposed model-based bootstrap with the episodic bootstrap [Hao et al., 2021] and plug-in CLT CIs [Zhu et al., 2024, Su et al., 2026].

Across the OPE and OPR tasks, the model-based bootstrap, especially the percentile CI, substantially improves coverage relative to the episodic bootstrap and plug-in CLT baselines. For OPE at $n \geq 500$, and for OPR with $T \in \{50, 100\}$ at $n \geq 500$, the model-based bootstrap percentile CI attains near-nominal coverage at nominal 50% (about 0.48–0.54), 90% (about 0.87–0.94), and 95% (about 0.92–0.97). The CI coverage for model-based bootstrap is weakest for rarely visited state–action pairs (near state 6) and for the short-horizon OPR setting $T = 10$, as expected from the strained asymptotic and non-degeneracy conditions. Detailed coverage tables and figures are reported in Appendix E.

5 Conclusion

We propose and analyze a model-based bootstrap for finite CMCs with possibly nonstationary or history-dependent control policies, motivated by the offline RL setting. We establish the bootstrap distributional consistency of the transition estimators in both the single long chain and episodic RL settings. We lift the bootstrap distributional consistency to OPE and OPR targets, exploiting the delta method by verifying Hadamard differentiability of the Bellman operators.

While our theoretical and numerical results are promising, we acknowledge several important considerations on the limitations and future directions of our work. First, we assume that the reward function is known in downstream RL tasks to avoid confusion in bootstrap targets. When the reward function is unknown and estimated from data, our bootstrap validity for OPE and OPR continues to hold if the reward estimate converges in probability to the mean reward, although finite-sample performance may be affected by the additional estimation error, which warrants further study. Second, we limit our current analysis to a finite state–action space. Extending our result to function approximation or continuous state-action spaces is an important future direction. Third, we only consider the asymptotic regime of fixed T and $K \to \infty$ for the episodic setting. We do not consider the asymptotic regime where K is fixed and greater than 1 and $T \to \infty$, and it remains an open question whether the bootstrap is consistent in this regime. Lastly, further experiments with non-stationary behavior policies, which the theory accommodates, remain to be conducted.

References

A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

K. Athreya and C.-D. Fuh. Bootstrapping markov chains: countable case. *Journal of Statistical Planning and Inference*, 33(3):311–331, 1992.

H. Babamoradi, F. van den Berg, and Å. Rinnan. Bootstrap based confidence limits in principal component analysis—a case study. *Chemometrics and Intelligent Laboratory Systems*, 120:97–105, 2013.

I. Banerjee. *Probably Approximately Correct Bounds for Estimating Markov Transition Kernels*. PhD thesis, Purdue University, 2023.

I. Banerjee and S. Chakrabarty. Small resamples, sharp guarantees: Convergence rates for resampled studentized quantile estimators. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

I. Banerjee, H. Honnappa, and V. Rao. Off-line estimation of controlled markov chains: Minimaxity and sample complexity. *Operations Research*, 2025.

P. Bertail and S. Clémençon. Regenerative block bootstrap for markov chains. *Bernoulli*, 12(4): 689–712, 2006.

D. P. Bertsekas. *Dynamic programming and optimal control, 3rd edition, volume II*. Belmont, MA: Athena Scientific, 2011.

P. Billingsley. Statistical methods in Markov chains. *The Annals of Mathematical Statistics*, pages 12–40, 1961.

V. S. Borkar. *Topics in controlled Markov chains*. Taylor & Francis, 1991.

H. Cramér and H. Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 1(4):290–294, 1936.

R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, pages 1–26, 1979.

M. K. S. Faradonbeh, A. Tewari, and G. Michailidis. On applications of bootstrap in continuous space reinforcement learning. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 1977–1984. IEEE, 2019.

S. Gonçalves, U. Hounyo, A. J. Patton, and K. Sheppard. Bootstrapping two-stage quasi-maximum likelihood estimators of time series models. *Journal of Business & Economic Statistics*, 41(3): 683–694, 2023.

P. Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, London, United Kingdom, 1980.

P. Hall and A. Padmanabhan. On the bootstrap and the trimmed mean. *Journal of Multivariate Analysis*, 41(1):132–153, 1992.

J. Han and Q. Liu. Bootstrap model aggregation for distributed statistical learning. *Advances in Neural Information Processing Systems*, 29, 2016.

J. Hanna, P. Stone, and S. Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

B. Hao, X. Ji, Y. Duan, H. Lu, C. Szepesvari, and M. Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR, 2021.

J. L. Horowitz. Bootstrap methods for markov processes. *Econometrica*, 71(4):1049–1082, 2003.

J. L. Horowitz. Bootstrap methods in econometrics. *Annual Review of Economics*, 11(1):193–224, 2019.

A. K. Jain and J. Moreau. Bootstrap technique in cluster analysis. *Pattern Recognition*, 20(5): 547–568, 1987.

Y. Jedra, J. Lee, A. Proutiere, and S.-Y. Yun. Nearly optimal latent state decoding in block mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 2805–2904. PMLR, 2023.

I. Kostrikov and O. Nachum. Statistical bootstrapping for uncertainty estimation in off-policy evaluation. *arXiv preprint arXiv:2007.13609*, 2020.

V. Krishnamurthy. *Partially observed Markov decision processes*. Cambridge university press, 2016.

R. Kulperger and B. P. Rao. Bootstrapping a finite state markov chain. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 178–191, 1989.

C. D. Meyer, Jr. The condition of a finite markov chain and perturbation bounds for the limiting probabilities. *SIAM Journal on Algebraic Discrete Methods*, 1(3):273–283, 1980.

J. Modayil and B. Kuipers. Towards bootstrap learning for object discovery. In *AAAI-2004 workshop on anchoring symbols to sensor data*, 2004.

P. Nakkiran, B. Neyshabur, and H. Sedghi. The deep bootstrap framework: Good online learners are good offline generalizers. *arXiv preprint arXiv:2010.08127*, 2020.

A. Rahoma, S. Imtiaz, and S. Ahmed. Sparse principal component analysis using bootstrap method. *Chemical Engineering Science*, 246:116890, 2021.

P. Ramprasad, Y. Li, Z. Yang, Z. Wang, W. W. Sun, and G. Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 118 (544):2901–2914, 2023.

E. Ruiz and L. Pascual. Bootstrapping financial time series. *Journal of Economic Surveys*, 16(3): 271–300, 2002.

A. L. Strehl and M. L. Littman. An empirical evaluation of interval estimation for markov decision processes. In *16th IEEE International Conference on Tools with Artificial Intelligence*, pages 128–135. IEEE, 2004.

Z. Su, I. Banerjee, and D. Klabjan. Central limit theorems for transition probabilities of controlled markov chains. *arXiv preprint arXiv:2508.01517*, 2026.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Y. Zhu, J. Dong, and H. Lam. Uncertainty quantification and exploration for reinforcement learning. *Operations Research*, 72(4):1689–1709, 2024.

A Sufficient Conditions for Assumption 2

Assumption 2 requires a uniform bound on the cumulative weak $\bar{\eta}$-mixing coefficients of the joint state–action process $\{(X_i, A_i)\}_{i\geq 0}$, which can be difficult to verify directly. Following Su et al. [2026], Banerjee et al. [2025], we give two separate conditions—one for the action process and one for the state process—whose conjunction implies Assumption 2.

Mixing of actions. For integers $p > i + j > i \geq 0$, we define the γ-mixing coefficient

$$\gamma_{p,j,i} := \sup_{\substack{s_p,\, h_{i+j}^{p-1},\, h_0^i \\ \mathbb{P}\left(X_p = s_p,\, H_{i+j}^{p-1} = h_{i+j}^{p-1},\, H_0^i = h_0^i\right) > 0}} \Big\| \mathbb{P}\left(A_p = \cdot \,\Big|\, X_p = s_p,\, H_{i+j}^{p-1} = h_{i+j}^{p-1},\, H_0^i = h_0^i\right) - \mathbb{P}\left(A_p = \cdot \,\Big|\, X_p = s_p,\, H_{i+j}^{p-1} = h_{i+j}^{p-1}\right) \Big\|_{\mathrm{TV}},$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance (TV). Thus $\gamma_{p,j,i}$ measures the worst-case sensitivity of the action distribution at time p to the distant past H_0^i, given the more recent history H_{i+j}^{p-1}. It thereby quantifies how much the dependence on old history of the behavior policy can alter future action selection. We now impose a summability condition on these coefficients.

Assumption 4 (Mixing of Actions)**.** *There exists a constant $C \geq 0$ such that*

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

When the sequence of actions a_p is Markovian, i.e. it depends only on the current state X_p, then $\gamma_{p,j,i} = 0$ for all p, j, i; thus, Assumption 4 holds with $C = 0$. This covers any stationary or non-stationary Markov behavior policy.

Mixing of states. For all integers $j \geq i \geq 0$, define the θ-mixing coefficient

$$\bar{\theta}_{i,j} := \sup_{\substack{(s_1,a_1),(s_2,a_2)\in\mathcal{S}\times\mathcal{A},\ (s_1,a_1)\neq(s_2,a_2) \\ \mathbb{P}(X_i=s_1,A_i=a_1)>0,\ \mathbb{P}(X_i=s_2,A_i=a_2)>0}} \Big\| \mathbb{P}\left(X_j = \cdot \mid X_i = s_1, A_i = a_1\right) - \mathbb{P}\left(X_j = \cdot \mid X_i = s_2, A_i = a_2\right) \Big\|_{\mathrm{TV}}.$$

Thus $\bar{\theta}_{i,j}$ measures the rate at which the state process forgets its initial state–action pair, extending Dobrushin's coefficients [Dobrushin, 1956] to the controlled setting. We now impose a summability condition on these coefficients.

Assumption 5 (Mixing of States)**.** *There exists a constant $C_\theta \geq 0$ such that*

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

Neither assumption implies the other. A CMC with deterministic actions satisfies Assumption 4 with $C = 0$ regardless of state mixing. Conversely, a CMC in which states are drawn i.i.d. satisfies Assumption 5 with $C_\theta = 0$ regardless of the behavior policy. The two conditions are therefore independent.

Finally, if Assumptions 4 and 5 hold, Banerjee et al. [2025, Lemma 3] implies that $\|\Delta_n\| \leq C + C_\theta + 1$ for all n. In particular, Assumption 2 holds with $C_\Delta = C + C_\theta + 1$.

B Auxiliary Lemmas for Theorem 1

B.1 Consistency of the Empirical Behavior Policy

Before we introduce the consistency of the empirical behavior policy, we introduce some requisite notation. We define a stationary behavior policy $\bar{\pi}_b(a \mid s) := p_s^{(a)}/p_s$, where $p_s := \sum_{a'\in\mathcal{A}} p_s^{(a')}$ is the sum of the ergodic occupation measure by actions. We now state the consistency of the empirical behavior policy.

Lemma 3 (Consistency of $\hat{\pi}_b$)**.** *Under Assumptions 1–3,*

$$\hat{\pi}_b(a \mid s) \xrightarrow{\text{a.s.}} \bar{\pi}_b(a \mid s) \qquad \textit{for every } (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Proof. As Assumptions 1–3 hold, by Su et al. [2026, Lemma 2], $N_s^{(a)}/n \xrightarrow{\text{a.s.}} p_s^{(a)}$ for every (s,a). Summing over $a \in \mathcal{A}$,

$$\frac{N_s}{n} = \sum_a \frac{N_s^{(a)}}{n} \xrightarrow{\text{a.s.}} p_s.$$

Assumption 3 gives $p_s^{(a)} \geq p_0 > 0$; thus, $p_s \geq p_0 > 0$. The continuous mapping theorem then yields

$$\hat{\pi}_b(a \mid s) = \frac{N_s^{(a)}/n}{N_s/n} \xrightarrow{\text{a.s.}} \frac{p_s^{(a)}}{p_s} = \bar{\pi}_b(a \mid s). \qquad \square$$

B.2 Properties of the Reference State–Action Chain

The analysis rests on a reference state–action Markov chain that couples the true transition kernel M with the stationary behavior policy $\bar{\pi}_b$. We begin by formally introducing the transition kernel of the reference state-action chain.

Definition 4 (Reference State–Action Chain)**.** *The reference state–action chain on $\mathcal{S} \times \mathcal{A}$ has one-step transition kernel*

$$K\big((s,a),(t,b)\big) := M_{s,t}^{(a)}\,\bar{\pi}_b(b \mid t), \qquad (s,a),(t,b) \in \mathcal{S} \times \mathcal{A}.$$

We record the key structural properties of K that are used throughout the proofs.

Lemma 4 (Properties of the Reference State–Action Chain)**.** *Under Assumptions 1–3 and the self-loop condition $M_{s_0,s_0}^{(a_0)} > 0$, we have*

(i) $p = (p_s^{(a)})_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ *is the unique stationary distribution of K,*

(ii) K *is irreducible and aperiodic on $\mathcal{S} \times \mathcal{A}$.*

Proof. (i) Unique stationarity. Since p is the ergodic occupation measure under (M, π_b), we have

$$p_s = \sum_{(t,l)} p_t^{(l)}\, M_{t,s}^{(l)}.$$

Therefore

$$\sum_{(t,l)\in\mathcal{S}\times\mathcal{A}} p_t^{(l)}\, K\big((t,l),(s,a)\big) = \Big(\sum_{(t,l)} p_t^{(l)} M_{t,s}^{(l)}\Big)\bar{\pi}_b(a \mid s) = p_s \cdot \frac{p_s^{(a)}}{p_s} = p_s^{(a)},$$

hence p satisfies $pK = p$. An irreducible Markov chain on a finite state space has a unique stationary distribution; uniqueness then follows from part (ii).

(ii) Irreducibility and aperiodicity. By Assumption 1, in the original CMC with control policy π_b, the hitting and return times to every state–action pair are finite a.s.; hence every state–action pair is hit and returned to infinitely often a.s. Therefore, for any two state–action pairs $(s,a),(t,b) \in \mathcal{S} \times \mathcal{A}$, there exists a finite path from (s,a) to (t,b) with positive probability under the original CMC.

By Assumption 3, $p_s^{(a)} \geq p_0 > 0$ for all (s,a); hence

$$\bar{\pi}_b(a \mid s) = \frac{p_s^{(a)}}{p_s} > 0$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Therefore, every positive-probability path segment under the original CMC has positive probability under the reference state–action kernel K, because each state transition along the path has positive transition probability and $\bar{\pi}_b$ assigns positive probability to every action at every state. Thus K is irreducible.

For aperiodicity, Assumption 3 gives

$$p_s = \sum_{a'} p_s^{(a')} \geq p_0 > 0 \qquad \text{and} \qquad \bar{\pi}_b(a \mid s) = \frac{p_s^{(a)}}{p_s} > 0 \qquad \text{for all } (s,a).$$

Then by the self-loop condition, we have

$$K\big((s_0,a_0),(s_0,a_0)\big) = M_{s_0,s_0}^{(a_0)}\, \bar{\pi}_b(a_0 \mid s_0) > 0.$$

A positive self-loop in a finite irreducible chain forces the period of the entire chain to be one; thus, K is aperiodic. □

B.3 Eventual Mixing Bound for the Bootstrap CMC

The bootstrap CMC $\{(X_i^*, A_i^*)\}$ replaces the true kernel M and the stationary policy $\bar{\pi}_b$ with their empirical counterparts $\hat{M}$ and $\hat{\pi}_b$, respectively. Let H_p^{*j} denote the history of the bootstrap CMC from time p to j, and $h_p^{*j} := \{X_p^* = s_p, A_p^* = a_p, \ldots, X_j^* = s_j, A_j^* = a_j\}$. We define the weak mixing coefficients for the bootstrap CMC $\{(X_i^*, A_i^*)\}$ as follows.

Definition 5 (Weak Mixing Coefficients for the Bootstrap CMC)**.** *For $i < j \leq n$, $\mathcal{T} \subseteq (\mathcal{S}\times\mathcal{A})^{n-j+1}$, $s_1, s_2 \in \mathcal{S}$, and $a_1, a_2 \in \mathcal{A}$, let $\eta_{i,j}^*(\mathcal{T}, s_1, s_2, a_1, a_2, h_0^{*i-1}, n) := \big|\mathbb{P}\big((X_j^*, A_j^*, \ldots, X_n^*, A_n^*) \in \mathcal{T} \mid X_i^* = s_1, A_i^* = a_1, H_0^{*i-1} = h_0^{*i-1}\big) - \mathbb{P}\big((X_j^*, A_j^*, \ldots, X_n^*, A_n^*) \in \mathcal{T} \mid X_i^* = s_2, A_i^* = a_2, H_0^{*i-1} = h_0^{*i-1}\big)\big|$. The weak $\bar{\eta}$-mixing coefficient for the bootstrap CMC is*

$$\bar{\eta}_{i,j}^*(n) := \sup_{\substack{\mathcal{T}, s_1, s_2, a_1, a_2, h_0^{*i-1}, \\ \mathbb{P}\left(X_i^*=s_1, A_i^*=a_1, H_0^{*i-1}=h_0^{*i-1}\right)>0, \\ \mathbb{P}\left(X_i^*=s_2, A_i^*=a_2, H_0^{*i-1}=h_0^{*i-1}\right)>0}} \eta_{i,j}^*(\mathcal{T}, s_1, s_2, a_1, a_2, h_0^{*i-1}, n).$$

Then, the bootstrap CMC counterpart of $\|\Delta_n\|$ in Assumption 2 is $\|\Delta_n^*\| := \max_{1\leq i\leq n}(1 + \bar{\eta}_{i,i+1}^* + \cdots + \bar{\eta}_{i,n}^*)$.

We formally introduce the transition kernel of the Markov chain of bootstrap state–action pairs that is used throughout the proofs.

Definition 6 (Bootstrap State–Action Chain)**.** *The* bootstrap state–action chain *on $\mathcal{S}\times\mathcal{A}$ has one-step transition kernel*

$$K_n^*\big((s,a),(t,b)\big) := \hat{M}_{s,t}^{(a)}\, \hat{\pi}_b(b \mid t), \qquad (s,a),(t,b) \in \mathcal{S}\times\mathcal{A}.$$

The following lemma shows that the bootstrap CMC $\{(X_i^*, A_i^*)\}$ inherits the mixing properties of the CMC $\{(X_i, A_i)\}$ in Assumption 2 for all sufficiently large n, almost surely.

Lemma 5 (Eventual Mixing Bound for Bootstrap Chain)**.** *There almost surely exist a finite (random) N_0 and a deterministic constant $C_\Delta^* < \infty$ such that*

$$\|\Delta_n^*\| \leq C_\Delta^* \qquad \textit{for all } n \geq N_0.$$

Proof. We proceed in five steps. Let N_0 be the (a.s. finite) random index beyond which all the entrywise convergences invoked in Steps 1–4 have taken effect simultaneously; its finiteness follows from the fact that each individual a.s. convergence has a finite threshold, and a finite maximum of a.s.-finite random variables is a.s. finite. All statements below hold for all $n \geq N_0$ a.s.

Step 1 (Entrywise convergence). By Lemma 3, $\hat{\pi}_b \to \bar{\pi}_b$ a.s. By Su et al. [2026, Lemma 2], $\hat{M}_{s,t}^{(a)} \to M_{s,t}^{(a)}$ a.s. for every (s,a,t). Hence $K_n^* \to K$ entrywise a.s.

Step 2 (Eventual irreducibility). By Assumption 3 and Su et al. [2026, Lemma 2], $N_s^{(a)}/n \to p_s^{(a)} \geq p_0 > 0$ a.s. Thus, $N_s^{(a)} > 0$ for all (s,a) and all $n \geq N_0$. When this holds, $\hat{\pi}_b(a \mid s) > 0$ and $\hat{M}_{s,t}^{(a)} > 0$ for all (s,a,t) such that $M_{s,t}^{(a)} > 0$.

Since K is irreducible by Lemma 4 (ii), for any $z, w \in \mathcal{S} \times \mathcal{A}$, there exists a finite sequence $z = z_0, z_1, \ldots, z_\ell = w$ such that $K(z_{i-1}, z_i) > 0$ for each $i = 1, \ldots, \ell$. Each such step satisfies $K(z_{i-1}, z_i) = M_{s,t}^{(a)} \bar{\pi}_b(l \mid t)$ for the corresponding (s, a, t, l), which requires $M_{s,t}^{(a)} > 0$. Since $\hat{M}_{s,t}^{(a)} > 0$ for all $n \geq N_0$ whenever $M_{s,t}^{(a)} > 0$, we have $K_n^*(z_{i-1}, z_i) > 0$ for every step. Thus, the same path has strictly positive probability under K_n^*. Hence K_n^* is irreducible for all $n \geq N_0$.

Step 3 (Eventual aperiodicity). Since $N_{s_0,s_0}^{(a_0)}/n \to p_{s_0}^{(a_0)} M_{s_0,s_0}^{(a_0)} > 0$ a.s., we have $N_{s_0,s_0}^{(a_0)} > 0$ for all $n \geq N_0$; thus, $\hat{M}_{s_0,s_0}^{(a_0)} > 0$, which gives

$$K_n^*\big((s_0, a_0), (s_0, a_0)\big) = \hat{M}_{s_0,s_0}^{(a_0)} \, \hat{\pi}_b(a_0 \mid s_0) > 0.$$

A positive self-loop in the irreducible K_n^* forces aperiodicity for all $n \geq N_0$.

Step 4 (Uniform Doeblin minorization and geometric ergodicity). Since K is finite, irreducible, and aperiodic by Lemma 4 (ii), it is primitive. Thus, there exists $r \geq 1$ such that $K^r(z, w) > 0$ for all $z, w \in \mathcal{S} \times \mathcal{A}$, where

$$K^r(z, w) \ := \sum_{z_1, \ldots, z_{r-1} \in \mathcal{S} \times \mathcal{A}} K(z, z_1) \, K(z_1, z_2) \cdots K(z_{r-1}, w)$$

is the r-step transition kernel ($K^1 := K$). We define

$$\eta := \min_{z, w \in \mathcal{S} \times \mathcal{A}} K^r(z, w) > 0.$$

Since entrywise convergence is preserved under matrix multiplication, $K_n^{*r} \to K^r$ entrywise a.s.; thus, for all $n \geq N_0$,

$$K_n^{*r}(z, w) \ \geq \ \frac{\eta}{2} \qquad \forall \, z, w \in \mathcal{S} \times \mathcal{A}.$$

Let $\nu(z) := 1/(SA)$ for all $z \in \mathcal{S} \times \mathcal{A}$ be the uniform distribution on $\mathcal{S} \times \mathcal{A}$, and set $\delta := SA \, \eta/2 > 0$. Then for all $n \geq N_0$ a.s., $K_n^{*r}(z, \cdot) \geq \delta \, \nu(\cdot)$ for all z. This is a uniform Doeblin minorization with deterministic r, δ, ν depending only on K.

We now derive the geometric ergodicity of K_n^* for all $n \geq N_0$ a.s. from the uniform Doeblin minorization. The first step is to derive a TV contraction for K_n^{*r} with contraction factor $1 - \delta$: for any two probability measures μ_1, μ_2 on $\mathcal{S} \times \mathcal{A}$,

$$\|\mu_1 K_n^{*r} - \mu_2 K_n^{*r}\|_{\mathrm{TV}} \ \leq \ (1 - \delta) \|\mu_1 - \mu_2\|_{\mathrm{TV}}.$$

We now derive the TV contraction explicitly. We write each row of K_n^{*r} as

$$K_n^{*r}(z, \cdot) = \delta \, \nu(\cdot) + (1 - \delta) \, R_z(\cdot), \qquad R_z(\cdot) := \frac{K_n^{*r}(z, \cdot) - \delta \, \nu(\cdot)}{1 - \delta},$$

where R_z is a probability measure. For any probability measure μ on $\mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}
(\mu K_n^{*r})(w) &= \sum_z \mu(z) \, K_n^{*r}(z, w) = \sum_z \mu(z) \big[\delta \, \nu(w) + (1 - \delta) \, R_z(w)\big] \\
&= \delta \, \nu(w) + (1 - \delta) \sum_z \mu(z) \, R_z(w),
\end{aligned}$$

where the last step uses $\sum_z \mu(z) = 1$. Then, for any two probability measures μ_1, μ_2 on $\mathcal{S} \times \mathcal{A}$, we subtract $\mu_1 K_n^{*r}$ from $\mu_2 K_n^{*r}$ to obtain

$$\mu_1 K_n^{*r} - \mu_2 K_n^{*r} = (1 - \delta) \left(\sum_z \mu_1(z) \, R_z - \sum_z \mu_2(z) \, R_z \right).$$

Taking the TV norm and applying the triangle inequality, we get

$$\begin{aligned}
\|\mu_1 K_n^{*r} - \mu_2 K_n^{*r}\|_{\mathrm{TV}} &= (1 - \delta) \Big\| \sum_z (\mu_1(z) - \mu_2(z)) \, R_z \Big\|_{\mathrm{TV}} \\
&= \frac{1 - \delta}{2} \sum_w \Big| \sum_z (\mu_1(z) - \mu_2(z)) \, R_z(w) \Big| \\
&\leq \frac{1 - \delta}{2} \sum_z |\mu_1(z) - \mu_2(z)| \sum_w R_z(w) \\
&= (1 - \delta) \, \|\mu_1 - \mu_2\|_{\mathrm{TV}},
\end{aligned}$$

where the final equality uses $\sum_w R_z(w) = 1$ since each R_z is a probability measure.

Let $\hat{p}^* := (\hat{p}_s^{*(a)})_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ denote the stationary distribution of K_n^*. For all $n \geq N_0$, since K_n^* is irreducible and aperiodic a.s., $\hat{p}^*$ is the unique stationary distribution of K_n^*, and $\hat{p}^* K_n^* = \hat{p}^*$.

We use this stationarity to bound $\|K_n^{*mr}(z,\cdot) - \hat{p}^*(\cdot)\|_{\mathrm{TV}}$ by induction on m. When $m = 0$, the bound holds trivially since $\|\delta_z - \hat{p}^*\|_{\mathrm{TV}} \leq 1 = (1-\delta)^0$. Suppose the bound holds for some $m \geq 0$, i.e. $\|\delta_z K_n^{*mr} - \hat{p}^*\|_{\mathrm{TV}} \leq (1-\delta)^m$. Applying the TV contraction with $\mu_1 = \delta_z K_n^{*mr}$ and $\mu_2 = \hat{p}^*$, and rewriting $K_n^{*(m+1)r} = K_n^{*mr} K_n^{*r}$, we have

$$\begin{aligned}\left\|\delta_z K_n^{*(m+1)r} - \hat{p}^*\right\|_{\mathrm{TV}} &= \left\|(\delta_z K_n^{*mr}) K_n^{*r} - \hat{p}^* K_n^{*r}\right\|_{\mathrm{TV}} \\ &\leq\ (1-\delta)\left\|\delta_z K_n^{*mr} - \hat{p}^*\right\|_{\mathrm{TV}} \\ &\leq\ (1-\delta)^{m+1},\end{aligned}$$

where the second inequality applies the induction hypothesis. Hence for all $m \geq 0$,

$$\sup_z \left\|K_n^{*mr}(z,\cdot) - \hat{p}^*(\cdot)\right\|_{\mathrm{TV}} \ \leq\ (1-\delta)^m.$$

For any $t \geq 0$, we can write $t = mr + s$ with $m \geq 0$ and $0 \leq s < r$. Applying the TV contraction to the pair $(\delta_z K_n^{*s},\, \hat{p}^*)$ over m further blocks of r steps, and using $\hat{p}^* K_n^{*mr} = \hat{p}^*$, we have

$$\begin{aligned}\left\|K_n^{*t}(z,\cdot) - \hat{p}^*(\cdot)\right\|_{\mathrm{TV}} &= \left\|\delta_z K_n^{*s} K_n^{*mr} - \hat{p}^* K_n^{*mr}\right\|_{\mathrm{TV}} \\ &\leq\ (1-\delta)^m \left\|\delta_z K_n^{*s} - \hat{p}^*\right\|_{\mathrm{TV}} \\ &\leq\ (1-\delta)^m.\end{aligned}$$

We define $\rho_* := (1-\delta)^{1/r} \in (0,1)$ and $C_* := (1-\delta)^{-1} < \infty$. They are deterministic constants depending only on K, since δ depends only on K. As $(1-\delta)^m = (1-\delta)^{(t-s)/r} \leq (1-\delta)^{-1}(1-\delta)^{t/r} = C_* \rho_*^t$, we conclude

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left\|K_n^{*t}\big((s,a),\cdot\big) - \hat{p}^*(\cdot)\right\|_{\mathrm{TV}} \ \leq\ C_* \rho_*^t, \qquad t \geq 0.$$

Step 5 (Applying Banerjee et al. [2025, Lemma 3]). Since $\hat{\pi}_b$ is a stationary Markov policy, Assumption 4 holds for the bootstrap CMC with $C = 0$. It remains to verify Assumption 5 for the bootstrap CMC.

We define the $\bar{\theta}$-mixing coefficient of the bootstrap CMC as

$$\begin{aligned}\bar{\theta}_{i,j}^* := \sup_{\substack{(s_1,a_1),(s_2,a_2)\in\mathcal{S}\times\mathcal{A},\ (s_1,a_1)\neq(s_2,a_2) \\ \mathbb{P}(X_i^*=s_1,A_i^*=a_1)>0,\ \mathbb{P}(X_i^*=s_2,A_i^*=a_2)>0}} &\Big\|\mathbb{P}\left(X_j^* = \cdot \mid X_i^* = s_1, A_i^* = a_1\right) \\ &- \mathbb{P}\left(X_j^* = \cdot \mid X_i^* = s_2, A_i^* = a_2\right)\Big\|_{\mathrm{TV}}.\end{aligned}$$

We connect the bound in Step 4 to the $\bar{\theta}$-mixing coefficient as follows.

For any two distributions μ, ν on $\mathcal{S}\times\mathcal{A}$ with $\mathcal{S}$-marginals μ_X, ν_X, marginalization cannot increase total variation and thus for every $B \subseteq \mathcal{S}$,

$$|\mu_X(B) - \nu_X(B)| = |\mu(B\times\mathcal{A}) - \nu(B\times\mathcal{A})| \leq \|\mu - \nu\|_{\mathrm{TV}}.$$

Hence $\|\mu_X - \nu_X\|_{\mathrm{TV}} \leq \|\mu - \nu\|_{\mathrm{TV}}$. Applying this to $\mu = K_n^{*(j-i)}((s_k,a_k),\cdot)$ and $\nu = \hat{p}^*$, whose $\mathcal{S}$-marginal is denoted by $\hat{\mu}_X^*$, gives

$$\left\|\mathbb{P}(X_j^* = \cdot \mid X_i^* = s_k, A_i^* = a_k) - \hat{\mu}_X^*(\cdot)\right\|_{\mathrm{TV}} \leq \left\|K_n^{*(j-i)}\big((s_k,a_k),\cdot\big) - \hat{p}^*(\cdot)\right\|_{\mathrm{TV}} \leq C_* \rho_*^{j-i}.$$

By the triangle inequality, we have

$$\bar{\theta}_{i,j}^* \leq 2C_* \rho_*^{j-i}, \qquad j > i \geq 1.$$

Therefore

$$\sup_{i\geq 1} \sum_{j>i} \bar{\theta}_{i,j}^* \ \leq\ 2\sum_{j=1}^{\infty} C_* \rho_*^j \ =\ \frac{2C_*\rho_*}{1-\rho_*} < \infty,$$

and Assumption 5 holds for the bootstrap CMC. Let $C_\Delta^* := \frac{2C_*\rho_*}{1-\rho_*} + 1$. Banerjee et al. [2025, Lemma 3] then implies that Assumption 2 holds for the bootstrap CMC, and $\|\Delta_n^*\| \leq 0 + \frac{2C_*\rho_*}{1-\rho_*} + 1 = C_\Delta^*$ for all $n \geq N_0$ a.s. This completes the proof. □

B.4 Continuity of the Stationary Distribution Map

Lemma 6 (Continuity of the Stationary Distribution Map)**.** *On the almost sure event of Lemma 5,*

$$\hat{p}_s^{*(a)} \xrightarrow{\text{a.s.}} p_s^{(a)} \qquad \textit{for every } (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Proof. By Lemma 4, K is irreducible and aperiodic on the finite state space $\mathcal{S} \times \mathcal{A}$. Since every irreducible finite-state chain is positive recurrent, K is ergodic, and its stationary distribution p is unique.

By Lemma 5, K_n^* is irreducible and aperiodic for all $n \geq N_0$ a.s. Therefore, K_n^* is eventually ergodic, with unique stationary distribution $\hat{p}^*$ for each $n \geq N_0$. Then by Meyer [1980, Theorem 4.1], the limiting probability vector of an ergodic chain is a continuous function of the transition matrix. Since $K_n^* \to K$ entrywise a.s. by Lemma 5 Step 1, we conclude $\hat{p}_s^{*(a)} \xrightarrow{\text{a.s.}} p_s^{(a)}$ for every $(s, a) \in \mathcal{S} \times \mathcal{A}$. □

B.5 Bootstrap Visitation Count Law of Large Numbers

Lemma 7 (Bootstrap Visitation Count LLN)**.** *For each $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

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

Proof. We begin by defining $\mu_n^* := \mathbb{E}[N_s^{*(a)} \mid \mathcal{D}_n]$ and writing

$$\frac{N_s^{*(a)}}{n} = \frac{N_s^{*(a)}}{\mu_n^*} \cdot \frac{\mu_n^*}{n} = \text{Term 1} \cdot \text{Term 2}.$$

We handle Term 2 first, then Term 1.

Term 2 (Expectation convergence). Let ν_n^* denote the initial distribution of (X_0^*, A_0^*) under K_n^*. Then

$$\mu_n^* = \sum_{i=0}^{n-1} (\nu_n^* K_n^{*i})(s, a),$$

hence $\mu_n^* - n\hat{p}_s^{*(a)} = \sum_{i=0}^{n-1} [(\nu_n^* K_n^{*i})(s, a) - \hat{p}_s^{*(a)}]$. By Lemma 5 (Step 4), $\sup_z \|K_n^{*i}(z, \cdot) - \hat{p}^*\|_{\text{TV}} \leq C_* \rho_*^i$ for all $n \geq N_0$ a.s. Since ν_n^* is a probability measure, we have

$$\|\nu_n^* K_n^{*i} - \hat{p}^*\|_{\text{TV}} \;\leq\; \sum_z \nu_n^*(z) \left\|K_n^{*i}(z, \cdot) - \hat{p}^*\right\|_{\text{TV}} \;\leq\; C_* \rho_*^i.$$

Therefore, for all $n \geq N_0$ a.s., we obtain

$$\left|\mu_n^* - n\hat{p}_s^{*(a)}\right| \;\leq\; \sum_{i=0}^{\infty} C_* \rho_*^i = \frac{C_*}{1 - \rho_*} < \infty.$$

Hence $\mu_n^*/n = \hat{p}_s^{*(a)} + O(1/n)$. Since $\hat{p}_s^{*(a)} \xrightarrow{\text{a.s.}} p_s^{(a)}$ by Lemma 6, we finally derive

$$\frac{\mu_n^*}{n} \xrightarrow{\text{a.s.}} p_s^{(a)}.$$

Term 1 (Concentration). On the probability-one event where $N_0 < \infty$ (Lemma 5), we choose $N_1 \geq N_0$ such that $\mu_n^* \geq (p_s^{(a)}/2)\, n$ and $\|\Delta_n^*\| \leq C_\Delta^*$ for all $n \geq N_1$ (which holds a.s. by Term 2 and Lemma 5). Applying Banerjee et al. [2025, Lemma 6] to the bootstrap chain with kernel K_n^*, we get that for every $\varepsilon > 0$ and $n \geq N_1$,

$$\mathbb{P}\left(\left|\frac{N_s^{*(a)}}{\mu_n^*} - 1\right| > \varepsilon\right) \;\leq\; 2\exp\left(-\frac{{\mu_n^*}^2 \varepsilon^2}{2n\|\Delta_n^*\|^2}\right) \;\leq\; 2\exp\left(-\frac{(p_s^{(a)})^2\, n\, \varepsilon^2}{8{C_\Delta^*}^2}\right).$$

The right-hand side is summable in n, and thus by the Borel–Cantelli lemma, we have $N_s^{*(a)}/\mu_n^* \xrightarrow{\text{a.s.}} 1$.

Since both Term 1 and Term 2 converge a.s., their product converges a.s. to $p_s^{(a)}$. This completes the proof. □

C Proof of Theorem 1

Proof. We work on the joint probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying both the observed data $\mathcal{D}_n$ and the bootstrap randomness; all convergence statements are under $\mathbb{P}$. We define the joint filtration $\mathcal{F}_{n,i} := \sigma(\mathcal{D}_n, X_0^*, A_0^*, \ldots, X_i^*, A_i^*)$.

We proceed in four steps. First, we decompose $\sqrt{n}(\hat{M}_{s,t}^{*(a)} - \hat{M}_{s,t}^{(a)})$ into a leading term and a remainder. Second, we show the remainder is negligible. Third, we show the leading term converges in distribution to $\mathcal{N}(0, \bar{\Lambda}_{sat,sat})$ conditionally on $\mathcal{D}_n$. This is the univariate, one-dimensional convergence of $\sqrt{n}(\hat{M}_{s,t}^{*(a)} - \hat{M}_{s,t}^{(a)})$. Finally, we use the Cramér–Wold theorem to extend the univariate convergence to the multivariate convergence of $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}})$ to $\mathcal{N}(0, \bar{\Lambda})$ conditionally on $\mathcal{D}_n$.

Step 1: Decomposition. Let $(s, a, t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ be fixed. We define

$$\xi_i^* := \mathbb{1}\{X_i^* = s, A_i^* = a\}\big(\mathbb{1}\{X_{i+1}^* = t\} - \hat{M}_{s,t}^{(a)}\big).$$

Since $\hat{M}_{s,t}^{*(a)} = N_{s,t}^{*(a)}/N_s^{*(a)}$, we have

$$\sqrt{n}\big(\hat{M}_{s,t}^{*(a)} - \hat{M}_{s,t}^{(a)}\big) = \sqrt{n}\left(\frac{N_{s,t}^{*(a)}}{N_s^{*(a)}} - \hat{M}_{s,t}^{(a)}\right) = \frac{\sqrt{n}}{N_s^{*(a)}}\big(N_{s,t}^{*(a)} - \hat{M}_{s,t}^{(a)} N_s^{*(a)}\big) = \frac{\sqrt{n}}{N_s^{*(a)}} \sum_{i=0}^{n-1} \xi_i^*.$$

Since $\sqrt{n}/N_s^{*(a)} = 1/(\sqrt{n}\, p_s^{(a)}) + (1/\sqrt{n})(n/N_s^{*(a)} - 1/p_s^{(a)})$, we write

$$\sqrt{n}\big(\hat{M}_{s,t}^{*(a)} - \hat{M}_{s,t}^{(a)}\big) = S_n^{*,sat} + R_n^{*,sat},$$

where

$$S_n^{*,sat} := \frac{1}{\sqrt{n}\, p_s^{(a)}} \sum_{i=0}^{n-1} \xi_i^*, \qquad R_n^{*,sat} := \Big(\frac{n}{N_s^{*(a)}} - \frac{1}{p_s^{(a)}}\Big)\frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} \xi_i^*.$$

Step 2: Remainder is negligible. We define

$$Z_n^* := \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} \xi_i^*,$$

such that $R_n^{*,sat} = \big(n/N_s^{*(a)} - 1/p_s^{(a)}\big) \cdot Z_n^*$. Our goal is to show $R_n^{*,sat} = o_{\mathbb{P}}(1)$, by first showing $Z_n^* = O_{\mathbb{P}}(1)$ and $n/N_s^{*(a)} - 1/p_s^{(a)} \xrightarrow{\text{a.s.}} 0$.

We start by bounding the variance of Z_n^*. On the event $\{X_i^* = s, A_i^* = a\}$, the bootstrap chain reaches next state t with probability $\hat{M}_{s,t}^{(a)}$, hence

$$\mathbb{E}[\xi_i^* \mid \mathcal{F}_{n,i}] = \mathbb{1}\{X_i^* = s, A_i^* = a\}\big(\hat{M}_{s,t}^{(a)} - \hat{M}_{s,t}^{(a)}\big) = 0.$$

Therefore (ξ_i^*) is a martingale difference sequence under $(\mathcal{F}_{n,i})$ and $\mathbb{P}$. For $i < j$, the tower property gives

$$\mathbb{E}[\xi_i^* \xi_j^*] = \mathbb{E}\big[\xi_i^*\, \mathbb{E}[\xi_j^* \mid \mathcal{F}_{n,j}]\big] = 0,$$

thus cross-terms in $\mathrm{Var}[Z_n^*]$ vanish. The conditional second moment of ξ_i^* is

$$\mathbb{E}\big[(\xi_i^*)^2 \mid \mathcal{F}_{n,i}\big] = \mathbb{1}\{X_i^* = s, A_i^* = a\}\, \hat{M}_{s,t}^{(a)}\big(1 - \hat{M}_{s,t}^{(a)}\big) \le \frac{1}{4},$$

since $\mathbb{1}\{X_{i+1}^* = t\}$ given $\mathcal{F}_{n,i}$ on $\{X_i^* = s, A_i^* = a\}$ is Bernoulli$(\hat{M}_{s,t}^{(a)})$. Summing over i, we get

$$\begin{aligned}
\mathrm{Var}[Z_n^*] &= \frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\big[(\xi_i^*)^2\big] \\
&= \frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\mathbb{E}\big[(\xi_i^*)^2 \mid \mathcal{F}_{n,i}\big]\right] \\
&\leq \ \frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{4} \\
&= \frac{1}{4}.
\end{aligned}$$

For any $M > 0$, by the Chebyshev's inequality, we have

$$\mathbb{P}\big(|Z_n^*| > M\big) \leq \frac{\mathrm{Var}[Z_n^*]}{M^2} \leq \frac{1}{4M^2},$$

hence $Z_n^* = O_{\mathbb{P}}(1)$. Since $N_s^{*(a)}/n \xrightarrow{\text{a.s.}} p_s^{(a)} > 0$, the continuous mapping theorem gives $n/N_s^{*(a)} - 1/p_s^{(a)} \xrightarrow{\text{a.s.}} 0$.

To show $R_n^{*,sat} = o_{\mathbb{P}}(1)$, we consider any $\varepsilon, \delta > 0$ and choose M large enough that $1/(4M^2) < \delta/2$. We define the event $\mathcal{E}_n := \{|n/N_s^{*(a)} - 1/p_s^{(a)}| > \varepsilon/M\} \cup \{|Z_n^*| > M\}$. On the event $\mathcal{E}_n^c$, both $|n/N_s^{*(a)} - 1/p_s^{(a)}| \leq \varepsilon/M$ and $|Z_n^*| \leq M$ hold. Thus, we have $|R_n^{*,sat}| = |n/N_s^{*(a)} - 1/p_s^{(a)}| \cdot |Z_n^*| \leq (\varepsilon/M) \cdot M = \varepsilon$. Therefore,

$$\begin{aligned}
\mathbb{P}\big(|R_n^{*,sat}| > \varepsilon\big) &= \mathbb{P}\big(|R_n^{*,sat}| > \varepsilon,\, \mathcal{E}_n^c\big) + \mathbb{P}\big(|R_n^{*,sat}| > \varepsilon,\, \mathcal{E}_n\big) \\
&= 0 + \mathbb{P}\big(|R_n^{*,sat}| > \varepsilon,\, \mathcal{E}_n\big) \\
&\leq \ \mathbb{P}\big(|Z_n^*| > M\big) \\
&\leq \ \frac{1}{4M^2} \\
&< \ \frac{\delta}{2}.
\end{aligned}$$

Since δ is arbitrary, we have $R_n^{*,sat} = o_{\mathbb{P}}(1)$. It remains to show $S_n^{*,sat} \xrightarrow{\text{d}} \mathcal{N}(0, \bar{\Lambda}_{sat,sat})$ under $\mathbb{P}$.

Step 3: One Dimensional Martingale CLT. We define

$$d_{n,i+1}^{*,sat} := \frac{\xi_i^*}{\sqrt{n}\, p_s^{(a)}},$$

so that $S_n^{*,sat} = \sum_{i=0}^{n-1} d_{n,i+1}^{*,sat}$. Since $\mathbb{E}[\xi_i^* \mid \mathcal{F}_{n,i}] = 0$, we have $\mathbb{E}[d_{n,i+1}^{*,sat} \mid \mathcal{F}_{n,i}] = 0$, and $(d_{n,i+1}^{*,sat})$ is a martingale difference array under $\mathbb{P}$ with respect to the joint filtration $(\mathcal{F}_{n,i})$.

To prove the CLT for $S_n^{*,sat}$, we verify the Lindeberg condition. Since $|\xi_i^*| \leq 1$, we have $|d_{n,i+1}^{*,sat}| \leq 1/(\sqrt{n}\, p_s^{(a)})$. For any $\varepsilon > 0$, we set $n_0 := \lceil (\varepsilon\, p_s^{(a)})^{-2} \rceil$, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. For all $n \geq n_0$, we have $|d_{n,i+1}^{*,sat}| \leq \varepsilon$ and thus $\{|d_{n,i+1}^{*,sat}| > \varepsilon\} = \emptyset$ and

$$\sum_{i=0}^{n-1}\mathbb{E}\Big[(d_{n,i+1}^{*,sat})^2\, \mathbb{1}\big\{|d_{n,i+1}^{*,sat}| > \varepsilon\big\} \,\Big|\, \mathcal{F}_{n,i}\Big] = 0.$$

Hence, the Lindeberg condition holds for all $n \geq n_0$.

We now derive the asymptotic variance of $S_n^{*,sat}$ by analyzing the predictable quadratic variation $\sum_{i=0}^{n-1}\mathbb{E}[(d_{n,i+1}^{*,sat})^2 \mid \mathcal{F}_{n,i}]$. Using

$$\mathbb{E}[(\xi_i^*)^2 \mid \mathcal{F}_{n,i}] = \mathbb{1}\{X_i^* = s, A_i^* = a\}\hat{M}_{s,t}^{(a)}(1 - \hat{M}_{s,t}^{(a)}),$$

we have

$$\mathbb{E}\big[(d_{n,i+1}^{*,sat})^2 \mid \mathcal{F}_{n,i}\big] = \frac{\mathbb{1}\{X_i^* = s, A_i^* = a\}}{n\,(p_s^{(a)})^2}\,\hat{M}_{s,t}^{(a)}\big(1-\hat{M}_{s,t}^{(a)}\big).$$

Summing over $i = 0,\dots,n-1$, we get

$$\sum_{i=0}^{n-1}\mathbb{E}\big[(d_{n,i+1}^{*,sat})^2 \mid \mathcal{F}_{n,i}\big] = \frac{N_s^{*(a)}}{n\,(p_s^{(a)})^2}\,\hat{M}_{s,t}^{(a)}\big(1-\hat{M}_{s,t}^{(a)}\big). \tag{C.1}$$

By Lemma 7, $N_s^{*(a)}/n \xrightarrow{\text{a.s.}} p_s^{(a)}$, and by Su et al. [2026, Lemma 2], $\hat{M}_{s,t}^{(a)} \xrightarrow{\text{a.s.}} M_{s,t}^{(a)}$. Therefore, the right-hand side of (C.1) converges a.s. to

$$\frac{M_{s,t}^{(a)}\big(1-M_{s,t}^{(a)}\big)}{p_s^{(a)}} = \bar{\Lambda}_{sat,sat}.$$

By the martingale CLT [Hall and Heyde, 1980, Corollary 3.1], applied conditionally on $\mathcal{D}_n$ for $\mathbb{P}$-a.e. $\mathcal{D}_n$, we obtain that, conditional on $\mathcal{D}_n$, $S_n^{*,sat} \xrightarrow{\text{d}} \mathcal{N}(0,\bar{\Lambda}_{sat,sat})$ for $\mathbb{P}$-almost every $(\mathcal{D}_n)$. Combined with $R_n^{*,sat} = o_{\mathbb{P}}(1)$, we conclude that, conditional on $\mathcal{D}_n$, $\sqrt{n}(\hat{M}_{s,t}^{*(a)} - \hat{M}_{s,t}^{(a)}) \xrightarrow{\text{d}} \mathcal{N}(0,\bar{\Lambda}_{sat,sat})$ for $\mathbb{P}$-almost every $(\mathcal{D}_n)$.

Step 4: Multivariate CLT via Cramér–Wold. Step 3 establishes the marginal CLT for each scalar coordinate $\sqrt{n}(\hat{M}_{s,t}^{*(a)} - \hat{M}_{s,t}^{(a)})$. We now extend this to the joint distribution over all coordinates simultaneously.

Let $S_n^* := (S_n^{*,sat})_{(s,a,t)}$ denote the stacked vector, such that $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}}) = S_n^* + o_{\mathbb{P}}(1)$, using $R_n^{*,sat} = o_{\mathbb{P}}(1)$ from Step 2. By the Cramér–Wold theorem, a random vector converges in distribution to a multivariate normal if and only if every fixed linear combination of its coordinates does. It therefore suffices to show that for every fixed $u \in \mathbb{R}^{S^2A}$,

$$u^\top S_n^* \xrightarrow{\text{d}} \mathcal{N}(0, u^\top \bar{\Lambda}\, u), \quad \text{conditional on } \mathcal{D}_n, \text{ for } \mathbb{P}\text{-a.e. } (\mathcal{D}_n).$$

We fix such u and let $p_0 := \min_{s,a} p_s^{(a)} > 0$. We define the projected increments

$$\tilde{d}_{n,i+1}^* := \sum_{s,a,t} u_{sat}\, d_{n,i+1}^{*,sat}$$

and the projected partial sum $\tilde{S}_n^* := \sum_{i=0}^{n-1}\tilde{d}_{n,i+1}^*$, so that $u^\top S_n^* = \tilde{S}_n^*$. Since each $d_{n,i+1}^{*,sat}$ has zero $\mathbb{P}$-conditional mean given $\mathcal{F}_{n,i}$, the same holds for $\tilde{d}_{n,i+1}^*$. Thus, $(\tilde{d}_{n,i+1}^*)$ is a martingale difference array under $\mathbb{P}$ with respect to $(\mathcal{F}_{n,i})$. We now verify the Lindeberg condition. Since $|d_{n,i+1}^{*,sat}| \le 1/(\sqrt{n}\,p_s^{(a)}) \le 1/(\sqrt{n}\,p_0)$, the triangle inequality gives $|\tilde{d}_{n,i+1}^*| \le \|u\|_1/(\sqrt{n}\,p_0)$. For any $\varepsilon > 0$, we set $n_0 := \lceil \|u\|_1^2/(\varepsilon p_0)^2\rceil$. For all $n \ge n_0$, we have $|\tilde{d}_{n,i+1}^*| \le \varepsilon$, and thus

$$\sum_{i=0}^{n-1}\mathbb{E}\Big[(\tilde{d}_{n,i+1}^*)^2\,\mathbb{1}\big\{|\tilde{d}_{n,i+1}^*| > \varepsilon\big\}\;\Big|\;\mathcal{F}_{n,i}\Big] = 0.$$

Hence, the Lindeberg condition holds for all $n \ge n_0$.

We next compute the predictable quadratic variation. We expand $(\tilde{d}_{n,i+1}^*)^2 = \big(\sum_{s,a,t} u_{sat}\, d_{n,i+1}^{*,sat}\big)^2$ to get

$$\mathbb{E}\big[(\tilde{d}_{n,i+1}^*)^2 \mid \mathcal{F}_{n,i}\big] = \sum_{s,a,t}\sum_{s',a',t'} u_{sat}\,u_{s'a't'}\,\mathbb{E}\big[d_{n,i+1}^{*,sat}\,d_{n,i+1}^{*,s'a't'} \mid \mathcal{F}_{n,i}\big].$$

We observe that when $(s,a) \neq (s',a')$, we have

$$\mathbb{1}\{X_i^* = s, A_i^* = a\}\cdot\mathbb{1}\{X_i^* = s', A_i^* = a'\} = 0,$$

since the chain occupies a single state–action pair at each step. Hence $d_{n,i+1}^{*,sat}\cdot d_{n,i+1}^{*,s'a't'} = 0$ and those cross-row terms vanish. For the diagonal terms $(s,a) = (s',a')$, we condition on $\mathcal{F}_{n,i}$ and use that

$(\mathbb{1}\{X^*_{i+1} = t\})_{t\in\mathcal{S}}$ given $(X^*_i = s, A^*_i = a)$ follows a multinomial distribution with 1 trial and event probabilities $(\hat{M}^{(a)}_{s,1}, \ldots, \hat{M}^{(a)}_{s,S})$. The second moment formula of the multinomial distribution gives

$$\begin{aligned}
&\mathbb{E}\big[\xi^{*,sat}_i \xi^{*,sat'}_i \mid \mathcal{F}_{n,i}\big] \\
&= \mathbb{1}\{X^*_i = s, A^*_i = a\} \times \mathbb{E}\big[(\mathbb{1}\{X^*_{i+1} = t\} - \hat{M}^{(a)}_{s,t})(\mathbb{1}\{X^*_{i+1} = t'\} - \hat{M}^{(a)}_{s,t'}) \\
&\qquad \mid X^*_i = s, A^*_i = a, \mathcal{D}_n\big] \\
&= \mathbb{1}\{X^*_i = s, A^*_i = a\}\, \hat{M}^{(a)}_{s,t}\big(\mathbb{1}\{t = t'\} - \hat{M}^{(a)}_{s,t'}\big).
\end{aligned}$$

Therefore

$$\begin{aligned}
\sum_{i=0}^{n-1} \mathbb{E}\big[(\tilde{d}^*_{n,i+1})^2 \mid \mathcal{F}_{n,i}\big] = \sum_{s,a} \frac{N^{*(a)}_s}{n\,(p^{(a)}_s)^2} \sum_{t,t'} u_{sat}\, u_{sat'} \\
\times \hat{M}^{(a)}_{s,t}\big(\mathbb{1}\{t = t'\} - \hat{M}^{(a)}_{s,t'}\big).
\end{aligned}$$

Since $N^{*(a)}_s/n \xrightarrow{\text{a.s.}} p^{(a)}_s$ and $\hat{M}^{(a)}_{s,t} \to M^{(a)}_{s,t}$ a.s., the quadratic variation converges a.s. to

$$\sum_{s,a} \frac{1}{p^{(a)}_s} \sum_{t,t'} u_{sat}\, u_{sat'}\big(M^{(a)}_{s,t}\mathbb{1}\{t = t'\} - M^{(a)}_{s,t} M^{(a)}_{s,t'}\big) = u^\top \bar{\Lambda}\, u.$$

Hence, conditional on $\mathcal{D}_n$, $\tilde{S}^*_n \xrightarrow{\text{d}} \mathcal{N}(0, u^\top \bar{\Lambda} u)$ for $\mathbb{P}$-almost every $(\mathcal{D}_n)$. Since u is arbitrary, the Cramér–Wold theorem gives that, for $\mathbb{P}$-almost every sequence of datasets $(\mathcal{D}_n)_{n\geq 1}$, conditional on $\mathcal{D}_n$, yields

$$\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}}) \ \xrightarrow{\text{d}}\ \mathcal{N}(0, \bar{\Lambda}).$$

This completes the proof. □

D Proofs of Propositions

D.1 Proof of Proposition 1

Proof. We use the notation introduced in Section 3.2.

We recall that the concatenated single-chain dataset is $\tilde{\mathcal{D}} := \{(\widetilde{X}_i, \widetilde{A}_i, \widetilde{X}_{i+1})\}_{i=0}^{n'-1}$ with $n' = K(T+1)-1$. The concatenated dataset is defined by setting $\widetilde{X}_{(k-1)(T+1)+j} := X^{(k)}_j$ and $\widetilde{A}_{(k-1)(T+1)+j} := A^{(k)}_j$ for $j = 0, \ldots, T-1$, and inserting at each episode boundary $k = 1, \ldots, K-1$ a pseudo-transition $\widetilde{A}_{k(T+1)-1} := a_\dagger$ mapping $X^{(k)}_T$ to $X^{(k+1)}_0$. The extended action space is $\widetilde{\mathcal{A}} = \mathcal{A} \cup \{a_\dagger\}$, and the resulting process $\{(\widetilde{X}_i, \widetilde{A}_i)\}$ is a CMC on $\mathcal{S} \times \widetilde{\mathcal{A}}$ with transition kernel $\widetilde{M}$, where $\widetilde{M}^{(a)}_{s,t} = M^{(a)}_{s,t}$ for $a \in \mathcal{A}$ and $\widetilde{M}^{(a_\dagger)}_{s,t} > 0$ for all $s, t \in \mathcal{S}$.

By hypotheses, $\{(\widetilde{X}_i, \widetilde{A}_i)\}$ satisfies Assumptions 1–3 on $\mathcal{S} \times \widetilde{\mathcal{A}}$, and there exists $(s_0, a_0) \in \mathcal{S} \times \mathcal{A}$ with $M^{(a_0)}_{s_0,s_0} > 0$. Applying Theorem 1 to $\tilde{\mathcal{D}}$ gives that, for $\mathbb{P}$-almost every $(\mathcal{D}_n)_{n\geq 1}$, conditional on $\mathcal{D}_n$, we have

$$\sqrt{n'}\,\mathrm{vec}(\hat{\widetilde{\mathbf{M}}}^* - \hat{\widetilde{\mathbf{M}}}) \ \xrightarrow{\text{d}}\ \mathcal{N}(0, \bar{\tilde{\Lambda}}), \quad \text{as } K \to \infty,$$

where $\bar{\tilde{\Lambda}}$ is the asymptotic covariance matrix. Among the $T+1$ steps per episode, exactly T involve actions in $\mathcal{A}$. Thus the ergodic occupation measure of the concatenated chain satisfies

$$\tilde{p}^{(a)}_s = \tfrac{T}{T+1}\, p^{(a)}_s \text{ for } a \in \mathcal{A}, \text{ giving } \bar{\tilde{\Lambda}}_{sat,sat} = \tfrac{T+1}{T}\, \bar{\Lambda}_{sat,sat} \text{ for all } (s,a,t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}.$$

Since $a_\dagger$-transitions are excluded from $\hat{\mathbf{M}}$ (the empirical estimator uses only $a \in \mathcal{A}$), we take the marginals of the Gaussian limit to get

$$\sqrt{n'}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}}) \xrightarrow{\text{d}} \mathcal{N}(0, \bar{\tilde{\Lambda}}|_{\mathcal{S}\times\mathcal{A}\times\mathcal{S}}),$$

where $\bar{\tilde{\Lambda}}|_{\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$ denotes the restriction of $\bar{\tilde{\Lambda}}$ to the entries corresponding to $(s, a, t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Among the $T+1$ steps per episode, exactly T involve actions in $\mathcal{A}$. Thus the ergodic occupation measure of the concatenated chain satisfies $\tilde{p}_s^{(a)} = \frac{T}{T+1}\, p_s^{(a)}$ for $a \in \mathcal{A}$, giving $\bar{\tilde{\Lambda}}_{sat,sat} = \frac{T+1}{T}\, \bar{\Lambda}_{sat,sat}$. Since $n' = K(T+1) - 1$ and $n = KT$, we have

$$\frac{n}{n'} = \frac{KT}{K(T+1)-1} = \frac{T}{T+1-1/K} \to \frac{T}{T+1}, \quad \text{as } K \to \infty.$$

Therefore, re-normalizing $\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}})$ by $\sqrt{n}$ yields

$$\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}}) = \sqrt{\frac{n}{n'}}\,\sqrt{n'}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}}) \ \xrightarrow{\mathrm{d}}\ \mathcal{N}\Big(0, \tfrac{T}{T+1}\,\bar{\tilde{\Lambda}}|_{\mathcal{S}\times\mathcal{A}\times\mathcal{S}}\Big) = \mathcal{N}(0, \bar{\Lambda}). \qquad \square$$

D.2 Proof of Proposition 2

The proof applies the bootstrap delta method to Theorem 1 (single-chain) or Proposition 1 (episodic). Before we provide the formal proof, we recall Hadamard differentiability [Van der Vaart, 2000, Chapter 20.2].

A map $\phi : \mathbb{D}_\phi \to \mathbb{E}$, defined on a subset $\mathbb{D}_\phi$ of a normed space $\mathbb{D}$ and taking values in another normed space $\mathbb{E}$, is Hadamard differentiable at $\theta \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}$ if there exists a continuous linear map $\phi'_\theta : \mathbb{D}_0 \to \mathbb{E}$ such that

$$\left\|\frac{\phi(\theta + th_n) - \phi(\theta)}{t} - \phi'_\theta(h)\right\|_{\mathbb{E}} \to 0 \ \text{ as } t \to 0, \text{ for every } h_n \to h \in \mathbb{D}_0.$$

The bootstrap delta method [Van der Vaart, 2000, Theorem 23.9] then states that if $\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \xrightarrow{\mathrm{d}} L$ conditionally on $\mathcal{D}_n$ in probability, and ϕ is Hadamard differentiable at θ tangentially to a subspace $\mathbb{D}_0$, then $\sqrt{n}(\phi(\hat{\theta}^*) - \phi(\hat{\theta})) \xrightarrow{\mathrm{d}} \phi'_\theta(L)$ conditionally on $\mathcal{D}_n$ in probability.

The proof of Proposition 2 is as follows.

Proof. **OPE targets V_π and Q_π.** We first derive the bootstrap consistency results for the OPE targets V_π. The bootstrap consistency for Q_π is entirely analogous.

We consider any fixed stationary target policy π with block-diagonal policy matrix Π. The value function is given by the Bellman equation

$$V_\pi(\mathbf{M}) = (I - \gamma\Pi\mathbf{M})^{-1} g.$$

We show that the map $V_\pi : \mathbb{R}^{S^2 A} \to \mathbb{R}^S$ is Hadamard differentiable at $\mathbf{M}$. Let H be any perturbation matrix and $t > 0$ sufficiently small. Using the matrix property $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, we have

$$\begin{aligned} V_\pi(\mathbf{M} + tH) - V_\pi(\mathbf{M}) &= \big[(I - \gamma\Pi(\mathbf{M} + tH))^{-1} - (I - \gamma\Pi\mathbf{M})^{-1}\big] g \\ &= (I - \gamma\Pi(\mathbf{M} + tH))^{-1}(\gamma\Pi t H)(I - \gamma\Pi\mathbf{M})^{-1} g \\ &= t\gamma(I - \gamma\Pi(\mathbf{M} + tH))^{-1}\Pi H\, V_\pi(\mathbf{M}). \end{aligned}$$

Dividing by t and taking $t \to 0$, we get

$$\frac{V_\pi(\mathbf{M} + tH) - V_\pi(\mathbf{M})}{t} \ \to\ \gamma(I - \gamma\Pi\mathbf{M})^{-1}\Pi H\, V_\pi(\mathbf{M}).$$

We define the linear map $V'_{\mathbf{M}} : \mathbb{R}^{S^2 A} \to \mathbb{R}^S$ by $V'_{\mathbf{M}}(H) := \gamma(I - \gamma\Pi\mathbf{M})^{-1}\Pi H\, V_\pi(\mathbf{M})$. The derivative $V'_{\mathbf{M}}$ is linear and continuous in H, hence for any $H_n \to H$, we have

$$\left\|\frac{V_\pi(\mathbf{M} + tH_n) - V_\pi(\mathbf{M})}{t} - V'_{\mathbf{M}}(H)\right\|_{\mathbb{E}} \to 0.$$

Therefore, $\mathbf{M} \mapsto V_\pi(\mathbf{M})$ is Hadamard differentiable at $\mathbf{M}$. Since Theorem 1 and Proposition 1 establish that $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \bar{\Lambda})$ conditionally on $\mathcal{D}_n$—the same Gaussian law as

$\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}} - \mathbf{M})$ in Su et al. [2026, Corollary 1]—the bootstrap delta method with derivative $V'_{\mathbf{M}}$ gives, conditionally on $\mathcal{D}_n$ in probability,

$$\sqrt{n}(\hat{V}^*_\pi - \hat{V}_\pi) \ \xrightarrow{\mathrm{d}}\ \mathcal{N}(0, \Sigma^\pi_V),$$

where Σ^π_V is the asymptotic covariance of $\sqrt{n}(\hat{V}_\pi - V_\pi)$ in Su et al. [2026, Theorem 2]. The statement for Q_π is analogous with $V'_{\mathbf{M}}$ replaced by the analogous $Q'_{\mathbf{M}}$ and Σ^π_V replaced by Σ^π_Q.

OPR targets $V_\star$ and $Q_\star$. We next derive the bootstrap consistency result for the OPR target $V_\star$. The bootstrap consistency for $Q_\star$ is entirely analogous.

The optimal-value functional $\mathbf{M} \mapsto V_\star(\mathbf{M})$ is not directly Hadamard differentiable, because the policy-selection map $\mathbf{M} \mapsto \pi_\star(\mathbf{M})$ involves an argmax. We resolve this via the non-degeneracy condition

$$\min_{s\in\mathcal{S}} \min_{a\neq\pi_\star(s)} \big\{Q_\star(s, \pi_\star(s)) - Q_\star(s,a)\big\} > 0.$$

We claim that this condition implies that $\pi_\star$ is locally constant in a neighborhood of $\mathbf{M}$. We denote the optimal Q-function under $\mathbf{M}'$ by $Q_\star(\mathbf{M}')$.

By the non-degeneracy condition, we define

$$\Delta := \min_{s\in\mathcal{S}} \min_{a\neq\pi_\star(s)} \{Q_\star(\mathbf{M}; s, \pi_\star(\mathbf{M}; s)) - Q_\star(\mathbf{M}; s, a)\} > 0.$$

Since the finite discounted optimal Bellman equation has a unique fixed point and the Bellman optimality operator is continuous in $\mathbf{M}'$, the map $\mathbf{M}' \mapsto Q_\star(\mathbf{M}')$ is continuous. Hence there exists a neighborhood $\mathcal{U}$ of $\mathbf{M}$ such that, for every $\mathbf{M}' \in \mathcal{U}$, we have

$$\max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} |Q_\star(\mathbf{M}'; s, a) - Q_\star(\mathbf{M}; s, a)| < \Delta/4.$$

Therefore, for every $s \in \mathcal{S}$ and every $a \neq \pi_\star(\mathbf{M}; s)$, we have

$$\begin{aligned}
Q_\star(\mathbf{M}'; s, \pi_\star(\mathbf{M}; s)) - Q_\star(\mathbf{M}'; s, a) &= Q_\star(\mathbf{M}; s, \pi_\star(\mathbf{M}; s)) - Q_\star(\mathbf{M}; s, a) \\
&\quad + [Q_\star(\mathbf{M}'; s, \pi_\star(\mathbf{M}; s)) - Q_\star(\mathbf{M}; s, \pi_\star(\mathbf{M}; s))] \\
&\quad - [Q_\star(\mathbf{M}'; s, a) - Q_\star(\mathbf{M}; s, a)] \\
&\geq Q_\star(\mathbf{M}; s, \pi_\star(\mathbf{M}; s)) - Q_\star(\mathbf{M}; s, a) \\
&\quad - 2\max_{s,a} |Q_\star(\mathbf{M}'; s, a) - Q_\star(\mathbf{M}; s, a)| \\
&> \Delta - \Delta/2 = \Delta/2 > 0.
\end{aligned}$$

Thus $\pi_\star(\mathbf{M}) = \pi_\star(\mathbf{M}')$ for all $\mathbf{M}' \in \mathcal{U}$.

By Su et al. [2026, Lemma 1], we have $\hat{\mathbf{M}} \xrightarrow{\text{a.s.}} \mathbf{M}$. Then $\hat{\mathbf{M}} \in \mathcal{U}$ and $\hat{\Pi}_\star = \Pi_\star$ for large n a.s. On this probability one event, the OPR functional locally coincides with the fixed-policy map $\mathbf{M}' \mapsto V_\star(\mathbf{M}')$, which is Hadamard differentiable by the aforementioned OPE argument. We can therefore apply the bootstrap delta method with the same derivative $V'_{\mathbf{M}}$ as in the OPE case.

Since Theorem 1 and Proposition 1 establish that $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{M}}^* - \hat{\mathbf{M}}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \bar{\Lambda})$ conditionally on $\mathcal{D}_n$—the same Gaussian law as Su et al. [2026, Corollary 1]—applying the bootstrap delta method yields, conditional on $\mathcal{D}_n$ in probability, we obtain

$$\sqrt{n}\big(\hat{V}^*_\star - \hat{V}_\star\big) \ \xrightarrow{\mathrm{d}}\ \mathcal{N}(0, \Sigma^{\pi_\star}_V), \quad \text{as } n \to \infty,$$

where $\Sigma^{\pi_\star}_V$ is the asymptotic covariance of $\sqrt{n}(\hat{V}_\star - V_\star)$ in Su et al. [2026, Theorem 3]. The statement for $Q_\star$ is analogous with $V'_{\mathbf{M}}$ replaced by the analogous $Q'_{\mathbf{M}}$ and $\Sigma^{\pi_\star}_V$ replaced by $\Sigma^{\pi_\star}_Q$. $\square$

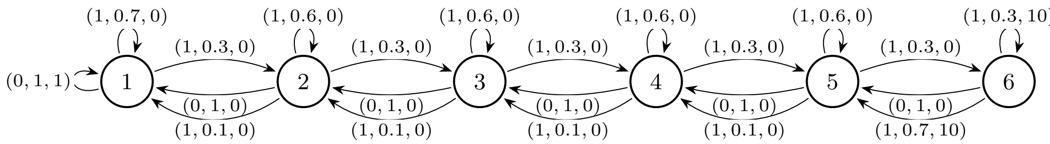

Figure 1: The RiverSwim MDP with $S = 6$ states. Arc labels (a, p, r): action (0 = swim left, 1 = swim right), transition probability p, immediate reward r.

E Details of Numerical Experiments

We evaluate finite-sample CI coverage of the proposed bootstrap on the RiverSwim problem [Strehl and Littman, 2004], replicating the experimental setup of the small state-space regime of Zhu et al. [2024]. Our RiverSwim state–action space has $S = 6$ states $\{1, \ldots, 6\}$ and $A = 2$ actions (0: swim left, 1: swim right). We set the rewards as $r(1, 0) = 1$, $r(6, 1) = 10$, $r(s, a) = 0$ otherwise, and discount factor $\gamma = 0.95$. Figure 1 illustrates the transition and reward structure of our RiverSwim setup. RiverSwim is designed to be difficult for exploration: the transition structure induces highly imbalanced visitation frequencies across state–action pairs, with upstream states (near state 6) visited far less frequently than downstream states (near state 1). Moreover, at its left-most state 1, the optimal policy is nearly degenerate, with $Q_\star(1, 0) \approx Q_\star(1, 1)$, thus the optimal action is only weakly identifiable. This places the problem near the boundary of optimal policy degeneracy and the transition probabilities for rarely visited state–action pairs are weakly estimated, thus the conditions for asymptotic normality in Proposition 2 are most strained.

In all experiments, we employ a stationary behavior policy $\pi_b(1 \mid s) = 0.8$ for all s. We use $B = 1{,}000$ bootstrap replicates and $N_{\mathrm{reps}} = 1{,}000$ Monte Carlo replications, and compare the model-based bootstrap (percentile and pivot CIs) against the episodic bootstrap [Hao et al., 2021] and plug-in CLT CIs [Zhu et al., 2024, Su et al., 2026]. We evaluate coverage at the nominal 95% level ($\alpha = 0.05$) for the entries $Q(1, 0)$, $Q(3, 1)$, $Q(6, 0)$, and $V(1)$–$V(6)$. Full results appear in Tables 2–7. Figures 2 and 3 display the coverages of the $n \in \{500, 1{,}000\}$ subset for the selected entries $Q(1, 0)$, $Q(3, 1)$, $Q(6, 0)$, $V(2)$, $V(4)$, and $V(5)$ in OPE and OPR, respectively. Appendices E.3 and E.4 report additional experiments at nominal 50% and 90%, respectively. Table 1 summarizes the settings of all experiments and the corresponding coverage tables. All coverage tables use a green cell-shading scheme described below.

Cell shading. In all coverage tables, we shade the cells green in proportion to empirical coverage relative to the nominal level. White indicates coverage at or below a floor value; full green indicates coverage at or above nominal; intermediate shades indicate coverages between the floor and nominal. The floor values are 0.75 (nominal 95%), 0.70 (nominal 90%), and 0.30 (nominal 50%).

For OPE, we conclude that the model-based bootstrap substantially outperforms the episodic bootstrap and plug-in CLT across all three target policies, with the percentile CI delivering the strongest overall performance and near-nominal coverage at $n = 500$–$1{,}000$. The performance of the model-based bootstrap is weakest for entries associated with state 6, where visitation is sparse and the corresponding transition probabilities are weakly identified. Undercoverage at these entries persists at $n = 100$ and is most severe for the mostly-left policy. The episodic bootstrap fails to produce valid intervals at $n = 50$, where a single episode eliminates all bootstrap variation, and suffers undercoverage at $n \leq 500$. The plug-in CI is moderately undercovered at $n \leq 100$ for all policies. Coverage improves monotonically with n for all methods. The same conclusions hold at nominal 50% and 90% (Appendix E.3).

For OPR, we conclude that the model-based bootstrap with percentile CI performs markedly better than the baselines and is the strongest method in the OPR study, achieving near-nominal coverage at $n = 500$–$1{,}000$ for $T = 50$ and $T = 100$. The performance of the model-based bootstrap is weakest at $T = 10$, where coverage at $n = 1{,}000$ remains below nominal, and coverage improves monotonically with T. The pivot CI undercovers at all (n, T) combinations, most severely at $Q_\star(1, 0)$, where $Q_\star(1, 0) \approx Q_\star(1, 1)$ and the non-degeneracy condition in Proposition 2 is weakest. The episodic bootstrap collapses to zero coverage at $n = 50$ when $T \geq 50$, where at most one complete episode is available. At $T = 10$ the collapse is avoided but coverage remains well below nominal. The plug-in CI severely undercovers at $n \leq 100$ and approaches near-nominal performance only at $n = 1{,}000, T = 100$ for the rightmost states and state-action pairs. Coverage improves

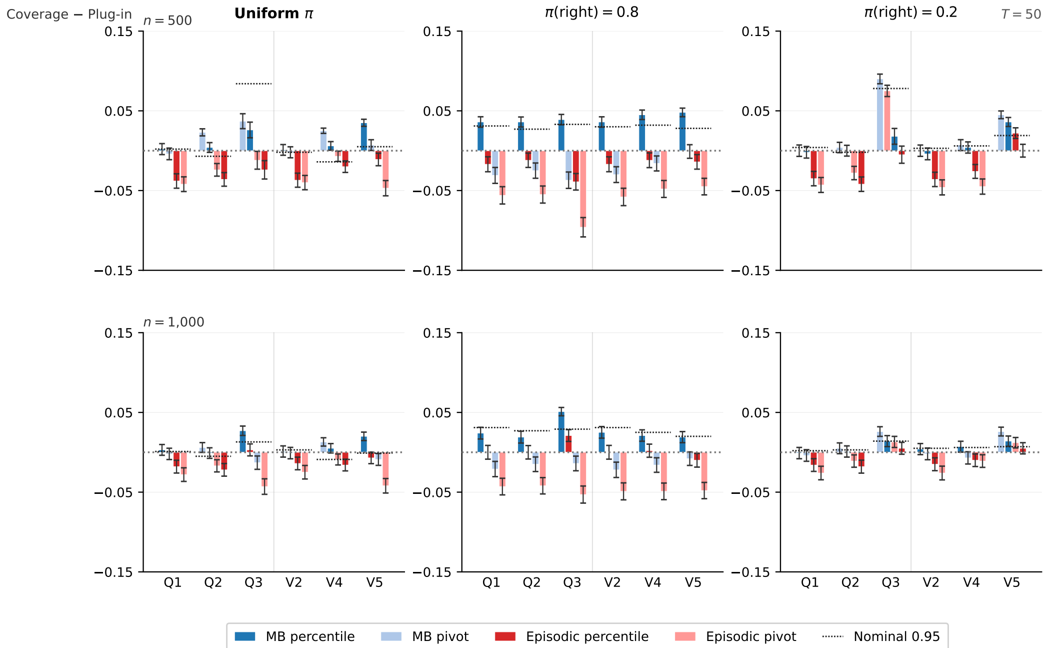

Figure 2: Empirical coverage relative to the Plug-in (CLT) baseline for $Q^\pi(1,0)$, $Q^\pi(3,1)$, $Q^\pi(6,0)$, $V^\pi(2)$, $V^\pi(4)$, $V^\pi(5)$ in RiverSwim. Horizontal lines mark the nominal 0.95 level for each entry. Columns correspond to the uniform, mostly-right, and mostly-left target policies, rows correspond to $n \in \{500, 1{,}000\}$, and all panels use $T = 50$. This figure visualizes the $n \in \{500, 1{,}000\}$ subset of Tables 2–4. Labels Q1–Q3 and V2, V4, V5 correspond to these entries respectively. Full results appear in Tables 2–4.

monotonically with n for all methods, and with T for the model-based bootstrap and plug-in CI. The same conclusions hold at nominal 50% and 90% (Appendix E.4).

Handling zero counts. We handle state-action pairs where $N_s^{(a)} = 0$ in the bootstrap simulation as follows. In bootstrap simulation, we use the repaired kernel

$$\hat{M}_{s,\cdot}^{\mathrm{sim},(a)} = \begin{cases} \hat{M}_{s,\cdot}^{(a)}, & N_s^{(a)} > 0, \\ \delta_s, & N_s^{(a)} = 0, \end{cases}$$

where δ_s is the point mass at s. We thus treat an unobserved state-action pair as a self-loop in bootstrap simulation rather than assigning fictitious transition probabilities. Similarly, if $N_s = 0$, we set $\hat{\pi}_b(\cdot \mid s)$ to the uniform distribution. Under Assumption 3, these rules are used with probability tending to zero and hence have no asymptotic effect.

We organize the remainder of this appendix as follows. Appendix E.1 reports OPE (V^π, Q^π) results at nominal 95%; Appendix E.2 reports OPR ($V_\star$, $Q_\star$) results at nominal 95%; Appendices E.3 and E.4 report OPE and OPR results at nominal 50% and 90%, respectively. Figures 4 and 5 display OPE coverages for the $n \in \{500, 1{,}000\}$ subset at nominal 50% and 90%; Figures 6 and 7 display the corresponding OPR results.

E.1 OPE (V^π, Q^π)

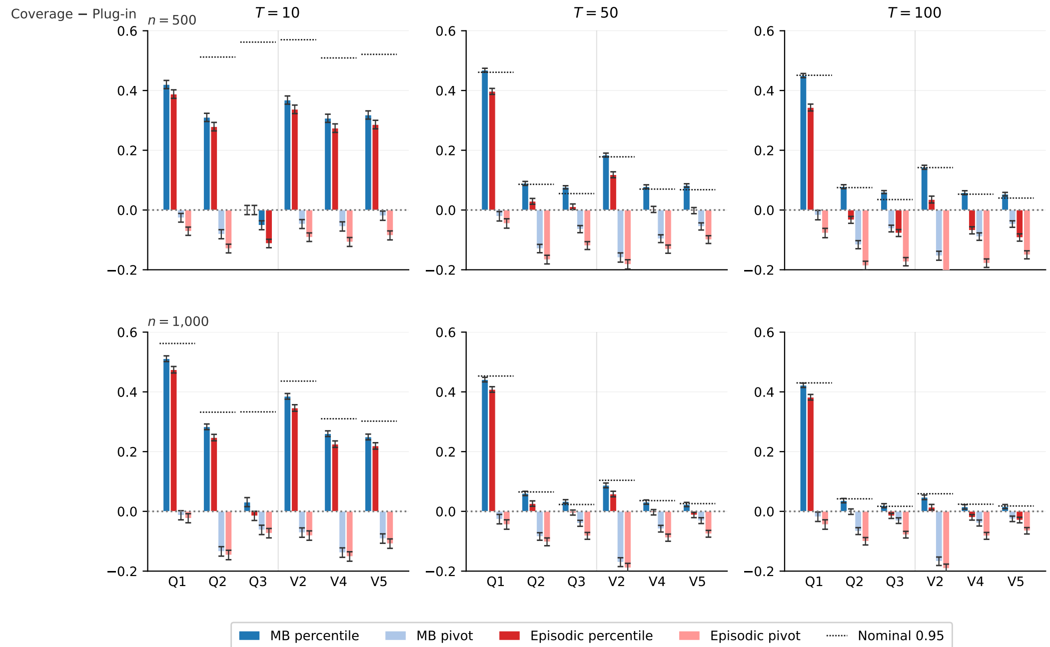

Figure 3: Empirical coverage relative to the Plug-in (CLT) baseline for $Q_\star(1,0)$, $Q_\star(3,1)$, $Q_\star(6,0)$, $V_\star(2)$, $V_\star(4)$, $V_\star(5)$ in RiverSwim. Horizontal lines mark the nominal 0.95 level for each entry. Columns correspond to $T \in \{10, 50, 100\}$, rows correspond to $n \in \{500, 1{,}000\}$, and the target policy is optimal. This figure visualizes the $n \in \{500, 1{,}000\}$ subset of Tables 5–7. Labels Q1–Q3 and V2, V4, V5 correspond to these entries respectively. Full results for all n appear in Tables 5–7.

Table 1: Settings for all RiverSwim coverage tables in this appendix. $K = n/T$ is the number of episodes. Policy abbreviations: uniform policy $\pi(a \mid s) = 0.5$; mostly-right policy $\pi(1 \mid s) = 0.8$; mostly-left policy $\pi(0 \mid s) = 0.8$; optimal policy $\pi^\star$. All experiments use $B = 1{,}000$ bootstrap replicates and $N_{\text{reps}} = 1{,}000$ Monte Carlo replications. Tables 2–4 report OPE coverages and Tables 5–7 report OPR coverages at nominal 95%. The 50% and 90% rows refer to additional-nominal-level tables in Appendices E.3 and E.4.

Table	n	T	K	Nominal	Policy
2	$\{50, 100, 500, 1{,}000\}$	50	$\{1, 2, 10, 20\}$	95%	uniform
3	$\{50, 100, 500, 1{,}000\}$	50	$\{1, 2, 10, 20\}$	95%	mostly-right
4	$\{50, 100, 500, 1{,}000\}$	50	$\{1, 2, 10, 20\}$	95%	mostly-left
5	$\{50, 100, 500, 1{,}000\}$	50	$\{1, 2, 10, 20\}$	95%	optimal
6	$\{50, 100, 500, 1{,}000\}$	10	$\{5, 10, 50, 100\}$	95%	optimal
7	$\{100, 500, 1{,}000\}$	100	$\{1, 5, 10\}$	95%	optimal
8 / 11	$\{50, 100, 500, 1{,}000\}$	50	$\{1, 2, 10, 20\}$	50%, 90%	uniform
9 / 12	$\{50, 100, 500, 1{,}000\}$	50	$\{1, 2, 10, 20\}$	50%, 90%	mostly-right
10 / 13	$\{50, 100, 500, 1{,}000\}$	50	$\{1, 2, 10, 20\}$	50%, 90%	mostly-left
14 / 15	$\{50, 100, 500, 1{,}000\}$	10	$\{5, 10, 50, 100\}$	50%, 90%	optimal
16 / 17	$\{50, 100, 500, 1{,}000\}$	50	$\{1, 2, 10, 20\}$	50%, 90%	optimal
18 / 19	$\{100, 500, 1{,}000\}$	100	$\{1, 5, 10\}$	50%, 90%	optimal

Table 2: Empirical coverage (nominal 95%) for a fixed target policy π (uniform over actions) in RiverSwim. Comparison of model-based bootstrap (ours), episodic bootstrap, and plug-in CI, episode length $T = 50$.

	Q^π coverage			V^π coverage					
n	$Q^\pi(1,0)$	$Q^\pi(3,1)$	$Q^\pi(6,0)$	$V^\pi(1)$	$V^\pi(2)$	$V^\pi(3)$	$V^\pi(4)$	$V^\pi(5)$	$V^\pi(6)$
Model-based bootstrap with pivot CI									
50	0.873	0.806	0.373	0.873	0.856	0.794	0.664	0.494	0.374
100	0.948	0.965	0.635	0.948	0.957	0.959	0.946	0.804	0.629
500	0.944	0.980	0.903	0.944	0.950	0.959	0.989	0.952	0.874
1,000	0.947	0.961	0.924	0.947	0.946	0.952	0.972	0.942	0.927
Model-based bootstrap with percentile CI									
50	0.896	0.886	0.182	0.896	0.882	0.870	0.629	0.435	0.229
100	0.937	0.958	0.333	0.937	0.938	0.946	0.905	0.669	0.362
500	0.950	0.961	0.892	0.950	0.953	0.956	0.970	0.980	0.896
1,000	0.952	0.954	0.964	0.952	0.948	0.952	0.964	0.971	0.966
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.546	0.575	0.535	0.546	0.551	0.566	0.575	0.431	0.535
500	0.906	0.933	0.854	0.906	0.912	0.926	0.957	0.898	0.818
1,000	0.921	0.938	0.894	0.921	0.922	0.935	0.950	0.909	0.891
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.515	0.539	0.227	0.515	0.519	0.530	0.512	0.476	0.228
500	0.910	0.921	0.842	0.910	0.915	0.915	0.944	0.934	0.829
1,000	0.931	0.933	0.940	0.931	0.933	0.936	0.943	0.944	0.936
Plug-in (CLT) CI									
50	0.849	0.790	0.206	0.849	0.810	0.777	0.619	0.358	0.315
100	0.929	0.926	0.371	0.929	0.929	0.922	0.899	0.603	0.405
500	0.948	0.957	0.866	0.948	0.952	0.952	0.964	0.945	0.822
1,000	0.949	0.955	0.937	0.949	0.947	0.951	0.959	0.951	0.922

Table 3: Empirical coverage (nominal 95%) for a fixed target policy π with $\pi(\text{right} \mid s) = 0.8$ for all s in RiverSwim. Comparison of model-based bootstrap (ours), episodic bootstrap, and plug-in CI, episode length $T = 50$.

	Q^π coverage			V^π coverage					
n	$Q^\pi(1,0)$	$Q^\pi(3,1)$	$Q^\pi(6,0)$	$V^\pi(1)$	$V^\pi(2)$	$V^\pi(3)$	$V^\pi(4)$	$V^\pi(5)$	$V^\pi(6)$
Model-based bootstrap with pivot CI									
50	0.430	0.407	0.403	0.430	0.430	0.408	0.395	0.371	0.358
100	0.646	0.630	0.608	0.646	0.642	0.632	0.623	0.591	0.558
500	0.888	0.898	0.880	0.888	0.890	0.895	0.902	0.921	0.879
1,000	0.898	0.908	0.907	0.898	0.897	0.906	0.909	0.922	0.925
Model-based bootstrap with percentile CI									
50	0.430	0.431	0.095	0.430	0.430	0.431	0.432	0.433	0.263
100	0.670	0.671	0.299	0.670	0.670	0.671	0.672	0.673	0.500
500	0.955	0.959	0.956	0.955	0.956	0.958	0.963	0.970	0.980
1,000	0.943	0.942	0.972	0.943	0.944	0.945	0.946	0.949	0.965
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.436	0.398	0.522	0.436	0.441	0.410	0.362	0.291	0.431
500	0.863	0.868	0.821	0.863	0.862	0.865	0.870	0.877	0.818
1,000	0.876	0.881	0.868	0.876	0.870	0.877	0.876	0.882	0.886
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.492	0.481	0.098	0.492	0.491	0.485	0.480	0.469	0.204
500	0.902	0.911	0.878	0.902	0.903	0.910	0.906	0.908	0.907
1,000	0.919	0.923	0.942	0.919	0.919	0.925	0.927	0.920	0.929
Plug-in (CLT) CI									
50	0.410	0.397	0.343	0.410	0.405	0.396	0.384	0.367	0.349
100	0.654	0.647	0.565	0.654	0.653	0.645	0.631	0.594	0.566
500	0.919	0.923	0.917	0.919	0.920	0.925	0.918	0.922	0.911
1,000	0.919	0.923	0.921	0.919	0.919	0.922	0.925	0.930	0.931

Table 4: Empirical coverage (nominal 95%) for a fixed target policy π with $\pi(\text{right} \mid s) = 0.2$ for all s in RiverSwim. Comparison of model-based bootstrap (ours), episodic bootstrap, and plug-in CI, episode length $T = 50$.

	Q^π coverage			V^π coverage					
n	$Q^\pi(1,0)$	$Q^\pi(3,1)$	$Q^\pi(6,0)$	$V^\pi(1)$	$V^\pi(2)$	$V^\pi(3)$	$V^\pi(4)$	$V^\pi(5)$	$V^\pi(6)$
Model-based bootstrap with pivot CI									
50	0.845	0.721	0.281	0.845	0.825	0.705	0.501	0.331	0.276
100	0.921	0.933	0.573	0.921	0.919	0.908	0.830	0.718	0.571
500	0.946	0.956	0.962	0.946	0.947	0.945	0.951	0.976	0.959
1,000	0.944	0.952	0.962	0.944	0.943	0.944	0.937	0.969	0.963
Model-based bootstrap with percentile CI									
50	0.780	0.710	0.392	0.780	0.762	0.638	0.408	0.308	0.402
100	0.910	0.867	0.550	0.910	0.899	0.875	0.796	0.680	0.588
500	0.944	0.952	0.890	0.944	0.943	0.944	0.948	0.967	0.891
1,000	0.947	0.948	0.950	0.947	0.949	0.949	0.951	0.957	0.958
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.560	0.598	0.490	0.560	0.564	0.578	0.545	0.419	0.490
500	0.903	0.924	0.947	0.903	0.901	0.906	0.899	0.931	0.937
1,000	0.922	0.936	0.949	0.922	0.919	0.915	0.933	0.955	0.941
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.504	0.501	0.371	0.504	0.501	0.484	0.451	0.406	0.379
500	0.911	0.910	0.867	0.911	0.911	0.907	0.918	0.953	0.868
1,000	0.932	0.929	0.941	0.932	0.930	0.931	0.934	0.948	0.944
Plug-in (CLT) CI									
50	0.856	0.675	0.158	0.856	0.810	0.685	0.475	0.302	0.175
100	0.936	0.882	0.304	0.936	0.933	0.898	0.822	0.663	0.319
500	0.946	0.952	0.872	0.946	0.947	0.948	0.944	0.931	0.867
1,000	0.948	0.947	0.936	0.948	0.945	0.947	0.944	0.943	0.946

E.2 OPR ($V_\star$, $Q_\star$)

Table 5: Empirical coverage (nominal 95%) for selected entries of $Q_\star$ and $V_\star$ in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap and plug-in CI, episode length $T = 50$.

	$Q_\star$ coverage			$V_\star$ coverage					
n	$Q_\star(1,0)$	$Q_\star(3,1)$	$Q_\star(6,0)$	$V_\star(1)$	$V_\star(2)$	$V_\star(3)$	$V_\star(4)$	$V_\star(5)$	$V_\star(6)$
Model-based bootstrap with pivot CI									
50	0.246	0.720	0.306	0.246	0.609	0.742	0.316	0.309	0.286
100	0.361	0.555	0.503	0.361	0.498	0.554	0.503	0.509	0.466
500	0.468	0.735	0.831	0.468	0.613	0.707	0.784	0.827	0.834
1,000	0.471	0.801	0.887	0.471	0.676	0.781	0.856	0.893	0.901
Model-based bootstrap with percentile CI									
50	0.433	0.433	0.289	0.433	0.433	0.433	0.430	0.429	0.325
100	0.667	0.670	0.501	0.667	0.668	0.670	0.668	0.669	0.555
500	0.957	0.953	0.971	0.957	0.956	0.953	0.958	0.964	0.970
1,000	0.938	0.945	0.960	0.938	0.933	0.945	0.945	0.947	0.950
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.159	0.251	0.369	0.159	0.225	0.247	0.213	0.225	0.290
500	0.444	0.698	0.776	0.444	0.590	0.676	0.749	0.783	0.776
1,000	0.453	0.783	0.845	0.453	0.657	0.764	0.826	0.849	0.861
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.434	0.438	0.225	0.434	0.432	0.438	0.430	0.428	0.293
500	0.886	0.893	0.906	0.886	0.890	0.893	0.882	0.880	0.899
1,000	0.905	0.911	0.923	0.905	0.904	0.911	0.911	0.912	0.911
Plug-in (CLT) CI									
50	0.251	0.321	0.313	0.251	0.306	0.318	0.323	0.321	0.306
100	0.373	0.521	0.514	0.373	0.465	0.521	0.526	0.515	0.500
500	0.489	0.864	0.895	0.489	0.772	0.864	0.880	0.882	0.880
1,000	0.497	0.885	0.927	0.497	0.846	0.885	0.914	0.924	0.925

Table 6: Empirical coverage (nominal 95%) for selected entries of $Q_\star$ and $V_\star$ in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap and plug-in CI, episode length $T = 10$.

	$Q_\star$ coverage			$V_\star$ coverage					
n	$Q_\star(1,0)$	$Q_\star(3,1)$	$Q_\star(6,0)$	$V_\star(1)$	$V_\star(2)$	$V_\star(3)$	$V_\star(4)$	$V_\star(5)$	$V_\star(6)$
Model-based bootstrap with pivot CI									
50	0.047	0.517	0.049	0.047	0.546	0.667	0.052	0.052	0.046
100	0.102	0.468	0.110	0.102	0.221	0.589	0.120	0.118	0.104
500	0.302	0.357	0.388	0.302	0.333	0.350	0.386	0.410	0.347
1,000	0.375	0.484	0.555	0.375	0.443	0.476	0.502	0.557	0.500
Model-based bootstrap with percentile CI									
50	0.115	0.115	0.014	0.115	0.115	0.115	0.115	0.115	0.036
100	0.237	0.237	0.048	0.237	0.237	0.237	0.237	0.237	0.086
500	0.748	0.748	0.337	0.748	0.748	0.748	0.748	0.747	0.417
1,000	0.899	0.901	0.648	0.899	0.899	0.901	0.900	0.897	0.690
Episodic bootstrap with pivot CI									
50	0.014	0.447	0.048	0.014	0.454	0.586	0.010	0.012	0.042
100	0.048	0.419	0.111	0.048	0.165	0.535	0.059	0.049	0.086
500	0.257	0.309	0.388	0.257	0.289	0.303	0.334	0.344	0.337
1,000	0.365	0.472	0.544	0.365	0.433	0.464	0.489	0.540	0.487
Episodic bootstrap with percentile CI									
50	0.115	0.114	0.006	0.115	0.115	0.114	0.113	0.109	0.011
100	0.234	0.233	0.032	0.234	0.234	0.233	0.233	0.233	0.056
500	0.716	0.717	0.276	0.716	0.717	0.717	0.715	0.715	0.347
1,000	0.862	0.865	0.602	0.862	0.860	0.865	0.865	0.867	0.650
Plug-in (CLT) CI									
50	0.043	0.043	0.037	0.043	0.047	0.043	0.039	0.041	0.041
100	0.092	0.105	0.084	0.092	0.104	0.105	0.101	0.092	0.092
500	0.328	0.438	0.388	0.328	0.380	0.438	0.441	0.429	0.389
1,000	0.388	0.618	0.617	0.388	0.514	0.618	0.640	0.648	0.601

Table 7: Empirical coverage (nominal 95%) for selected entries of $Q_\star$ and $V_\star$ in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in CI, episode length $T = 100$.

	$Q_\star$ coverage			$V_\star$ coverage					
n	$Q_\star(1,0)$	$Q_\star(3,1)$	$Q_\star(6,0)$	$V_\star(1)$	$V_\star(2)$	$V_\star(3)$	$V_\star(4)$	$V_\star(5)$	$V_\star(6)$
Model-based bootstrap with pivot CI									
100	0.350	0.549	0.524	0.350	0.521	0.549	0.527	0.545	0.492
500	0.482	0.759	0.853	0.482	0.655	0.742	0.808	0.863	0.872
1,000	0.502	0.842	0.902	0.502	0.724	0.830	0.887	0.907	0.908
Model-based bootstrap with percentile CI									
100	0.689	0.693	0.566	0.689	0.691	0.693	0.692	0.695	0.611
500	0.949	0.953	0.975	0.949	0.951	0.953	0.955	0.963	0.967
1,000	0.942	0.944	0.952	0.942	0.938	0.944	0.941	0.948	0.953
Episodic bootstrap with pivot CI									
100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
500	0.422	0.689	0.742	0.422	0.589	0.673	0.719	0.760	0.764
1,000	0.476	0.808	0.855	0.476	0.700	0.797	0.844	0.867	0.868
Episodic bootstrap with percentile CI									
100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
500	0.842	0.842	0.838	0.842	0.843	0.842	0.829	0.818	0.823
1,000	0.902	0.907	0.918	0.902	0.905	0.907	0.906	0.903	0.896
Plug-in (CLT) CI									
100	0.366	0.560	0.571	0.366	0.482	0.560	0.575	0.561	0.544
500	0.499	0.875	0.915	0.499	0.808	0.875	0.897	0.910	0.904
1,000	0.520	0.908	0.933	0.520	0.891	0.908	0.926	0.932	0.933

E.3 OPE: Additional Nominal Levels (50% and 90%)

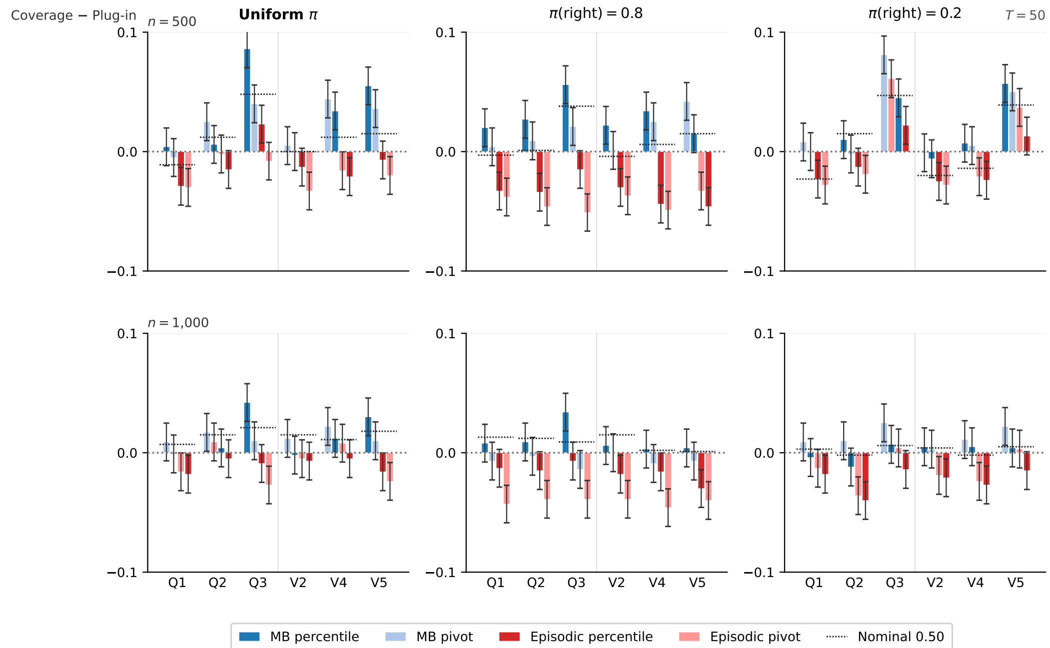

Figure 4: Empirical coverage relative to the Plug-in (CLT) baseline for $Q^\pi(1,0)$, $Q^\pi(3,1)$, $Q^\pi(6,0)$, $V^\pi(2)$, $V^\pi(4)$, $V^\pi(5)$ in RiverSwim at nominal 50%. Horizontal lines mark the nominal 0.50 level for each entry. Columns correspond to the uniform, mostly-right, and mostly-left target policies, rows correspond to $n \in \{500, 1{,}000\}$, and all panels use $T = 50$. Full results appear in Tables 8–10.

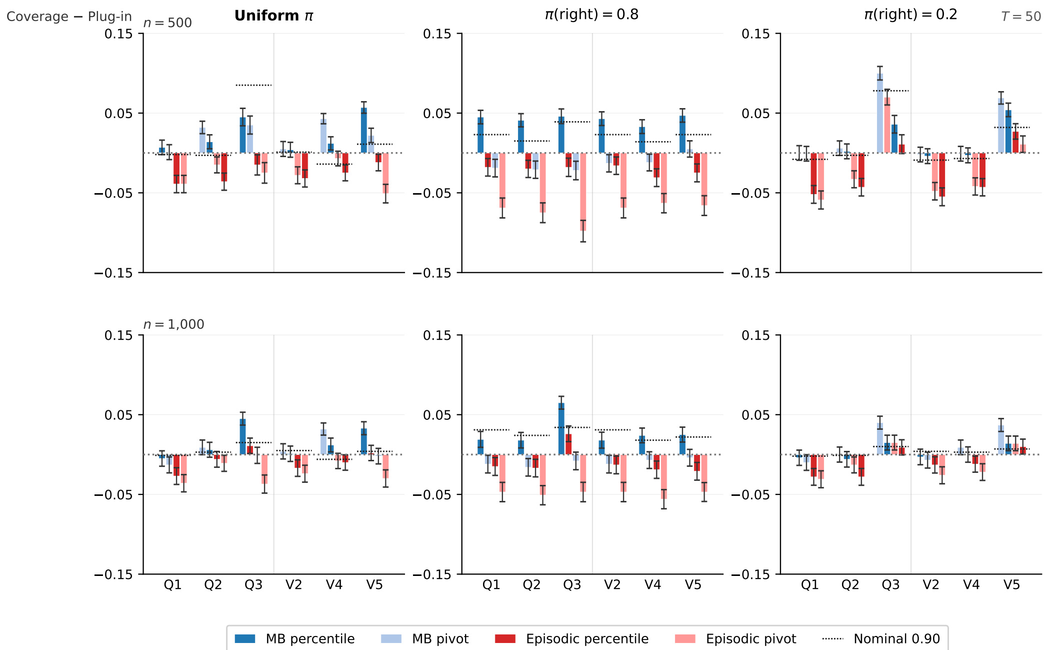

Figure 5: Empirical coverage relative to the Plug-in (CLT) baseline for $Q^{\pi}(1,0)$, $Q^{\pi}(3,1)$, $Q^{\pi}(6,0)$, $V^{\pi}(2)$, $V^{\pi}(4)$, $V^{\pi}(5)$ in RiverSwim at nominal 90%. Horizontal lines mark the nominal 0.90 level for each entry. Columns correspond to the uniform, mostly-right, and mostly-left target policies, rows correspond to $n \in \{500, 1{,}000\}$, and all panels use $T = 50$. Full results appear in Tables 11–13.

Table 8: Empirical coverage (nominal 50%) for a fixed target policy π (uniform over actions) in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 50$.

	Q^π coverage			V^π coverage					
n	$Q^\pi(1,0)$	$Q^\pi(3,1)$	$Q^\pi(6,0)$	$V^\pi(1)$	$V^\pi(2)$	$V^\pi(3)$	$V^\pi(4)$	$V^\pi(5)$	$V^\pi(6)$
Model-based bootstrap with pivot CI									
50	0.543	0.478	0.253	0.543	0.548	0.457	0.375	0.262	0.255
100	0.550	0.602	0.435	0.550	0.571	0.586	0.582	0.504	0.453
500	0.506	0.513	0.492	0.506	0.505	0.515	0.532	0.521	0.502
1,000	0.502	0.502	0.489	0.502	0.497	0.496	0.511	0.492	0.506
Model-based bootstrap with percentile CI									
50	0.432	0.471	0.104	0.432	0.434	0.444	0.239	0.301	0.084
100	0.499	0.570	0.184	0.499	0.507	0.536	0.541	0.503	0.155
500	0.515	0.494	0.538	0.515	0.500	0.494	0.522	0.540	0.519
1,000	0.492	0.489	0.521	0.492	0.483	0.476	0.501	0.512	0.551
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.302	0.361	0.286	0.302	0.306	0.338	0.386	0.273	0.292
500	0.481	0.486	0.444	0.481	0.467	0.479	0.472	0.465	0.422
1,000	0.477	0.494	0.452	0.477	0.480	0.488	0.497	0.458	0.446
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.300	0.319	0.130	0.300	0.312	0.323	0.309	0.284	0.127
500	0.482	0.473	0.475	0.482	0.487	0.477	0.467	0.478	0.445
1,000	0.475	0.480	0.470	0.475	0.478	0.464	0.484	0.466	0.490
Plug-in (CLT) CI									
50	0.405	0.392	0.083	0.405	0.416	0.392	0.284	0.174	0.077
100	0.485	0.489	0.185	0.485	0.479	0.484	0.456	0.305	0.140
500	0.511	0.488	0.452	0.511	0.500	0.486	0.488	0.485	0.440
1,000	0.493	0.485	0.479	0.493	0.485	0.484	0.489	0.482	0.497

Table 9: Empirical coverage (nominal 50%) for a fixed target policy π (mostly-right: $\pi(1 \mid s) = 0.8$) in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 50$.

	Q^π coverage			V^π coverage					
n	$Q^\pi(1,0)$	$Q^\pi(3,1)$	$Q^\pi(6,0)$	$V^\pi(1)$	$V^\pi(2)$	$V^\pi(3)$	$V^\pi(4)$	$V^\pi(5)$	$V^\pi(6)$
Model-based bootstrap with pivot CI									
50	0.271	0.224	0.232	0.271	0.261	0.229	0.198	0.180	0.219
100	0.427	0.406	0.422	0.427	0.424	0.415	0.389	0.352	0.390
500	0.507	0.508	0.483	0.507	0.505	0.509	0.519	0.527	0.533
1,000	0.480	0.485	0.477	0.480	0.485	0.486	0.489	0.492	0.487
Model-based bootstrap with percentile CI									
50	0.304	0.310	0.005	0.304	0.306	0.310	0.307	0.306	0.041
100	0.434	0.445	0.036	0.434	0.434	0.443	0.447	0.448	0.163
500	0.523	0.526	0.518	0.523	0.526	0.524	0.528	0.500	0.526
1,000	0.495	0.497	0.525	0.495	0.491	0.494	0.501	0.503	0.515
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.229	0.205	0.309	0.229	0.229	0.213	0.193	0.155	0.250
500	0.465	0.453	0.411	0.465	0.467	0.454	0.445	0.452	0.437
1,000	0.444	0.449	0.452	0.444	0.446	0.450	0.452	0.459	0.465
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.240	0.237	0.077	0.240	0.241	0.239	0.236	0.237	0.140
500	0.470	0.465	0.447	0.470	0.474	0.464	0.450	0.439	0.447
1,000	0.474	0.473	0.484	0.474	0.467	0.470	0.482	0.469	0.467
Plug-in (CLT) CI									
50	0.195	0.188	0.119	0.195	0.195	0.186	0.184	0.185	0.161
100	0.332	0.324	0.208	0.332	0.331	0.323	0.316	0.285	0.255
500	0.503	0.499	0.462	0.503	0.504	0.502	0.494	0.485	0.493
1,000	0.487	0.488	0.491	0.487	0.485	0.492	0.498	0.499	0.496

Table 10: Empirical coverage (nominal 50%) for a fixed target policy π (mostly-left: $\pi(0 \mid s) = 0.8$) in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 50$.

	Q^π coverage			V^π coverage					
n	$Q^\pi(1,0)$	$Q^\pi(3,1)$	$Q^\pi(6,0)$	$V^\pi(1)$	$V^\pi(2)$	$V^\pi(3)$	$V^\pi(4)$	$V^\pi(5)$	$V^\pi(6)$
Model-based bootstrap with pivot CI									
50	0.533	0.459	0.234	0.533	0.530	0.449	0.294	0.145	0.227
100	0.557	0.573	0.440	0.557	0.559	0.560	0.500	0.359	0.445
500	0.531	0.483	0.534	0.531	0.519	0.514	0.519	0.511	0.520
1,000	0.506	0.512	0.519	0.506	0.499	0.505	0.513	0.517	0.521
Model-based bootstrap with percentile CI									
50	0.389	0.296	0.229	0.389	0.364	0.297	0.087	0.036	0.238
100	0.511	0.444	0.312	0.511	0.506	0.475	0.404	0.325	0.324
500	0.523	0.495	0.498	0.523	0.514	0.511	0.521	0.518	0.501
1,000	0.493	0.490	0.501	0.493	0.501	0.494	0.507	0.499	0.512
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.318	0.375	0.249	0.318	0.317	0.351	0.354	0.313	0.255
500	0.495	0.466	0.514	0.495	0.492	0.487	0.493	0.498	0.473
1,000	0.484	0.466	0.498	0.484	0.477	0.477	0.478	0.498	0.497
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.293	0.303	0.206	0.293	0.310	0.296	0.322	0.308	0.203
500	0.500	0.472	0.475	0.500	0.495	0.482	0.490	0.474	0.462
1,000	0.479	0.462	0.480	0.479	0.475	0.469	0.475	0.480	0.480
Plug-in (CLT) CI									
50	0.407	0.284	0.073	0.407	0.391	0.331	0.230	0.143	0.077
100	0.505	0.466	0.154	0.505	0.510	0.480	0.426	0.323	0.142
500	0.523	0.485	0.453	0.523	0.520	0.504	0.514	0.461	0.432
1,000	0.497	0.502	0.494	0.497	0.496	0.503	0.502	0.495	0.488

Table 11: Empirical coverage (nominal 90%) for a fixed target policy π (uniform over actions) in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 50$.

	Q^π coverage			V^π coverage					
n	$Q^\pi(1,0)$	$Q^\pi(3,1)$	$Q^\pi(6,0)$	$V^\pi(1)$	$V^\pi(2)$	$V^\pi(3)$	$V^\pi(4)$	$V^\pi(5)$	$V^\pi(6)$
Model-based bootstrap with pivot CI									
50	0.842	0.784	0.365	0.842	0.834	0.772	0.649	0.479	0.362
100	0.912	0.952	0.621	0.912	0.925	0.936	0.934	0.785	0.614
500	0.903	0.935	0.850	0.903	0.904	0.919	0.957	0.911	0.829
1,000	0.888	0.906	0.884	0.888	0.899	0.905	0.938	0.894	0.879
Model-based bootstrap with percentile CI									
50	0.821	0.836	0.168	0.821	0.814	0.811	0.560	0.433	0.183
100	0.894	0.922	0.322	0.894	0.893	0.917	0.861	0.662	0.324
500	0.909	0.917	0.860	0.909	0.903	0.911	0.926	0.946	0.871
1,000	0.896	0.903	0.930	0.896	0.896	0.899	0.918	0.929	0.935
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.546	0.575	0.535	0.546	0.551	0.566	0.575	0.431	0.535
500	0.863	0.888	0.790	0.863	0.871	0.875	0.907	0.838	0.752
1,000	0.865	0.886	0.848	0.865	0.871	0.876	0.898	0.866	0.841
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.515	0.539	0.227	0.515	0.519	0.530	0.512	0.476	0.228
500	0.863	0.867	0.800	0.863	0.867	0.866	0.889	0.877	0.781
1,000	0.874	0.891	0.896	0.874	0.878	0.884	0.896	0.898	0.877
Plug-in (CLT) CI									
50	0.791	0.738	0.174	0.791	0.770	0.723	0.570	0.342	0.246
100	0.878	0.893	0.341	0.878	0.872	0.881	0.855	0.565	0.344
500	0.902	0.903	0.815	0.902	0.899	0.906	0.914	0.889	0.772
1,000	0.901	0.897	0.885	0.901	0.895	0.897	0.906	0.896	0.878

Table 12: Empirical coverage (nominal 90%) for a fixed target policy π (mostly-right: $\pi(1 \mid s) = 0.8$) in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 50$.

	Q^π coverage			V^π coverage					
n	$Q^\pi(1,0)$	$Q^\pi(3,1)$	$Q^\pi(6,0)$	$V^\pi(1)$	$V^\pi(2)$	$V^\pi(3)$	$V^\pi(4)$	$V^\pi(5)$	$V^\pi(6)$
Model-based bootstrap with pivot CI									
50	0.424	0.390	0.393	0.424	0.418	0.393	0.372	0.349	0.349
100	0.637	0.618	0.595	0.637	0.632	0.622	0.605	0.566	0.542
500	0.858	0.864	0.839	0.858	0.864	0.866	0.874	0.882	0.846
1,000	0.857	0.860	0.858	0.857	0.857	0.861	0.875	0.874	0.880
Model-based bootstrap with percentile CI									
50	0.421	0.425	0.057	0.421	0.421	0.423	0.426	0.426	0.208
100	0.654	0.654	0.218	0.654	0.655	0.655	0.661	0.664	0.440
500	0.922	0.926	0.907	0.922	0.920	0.926	0.919	0.924	0.942
1,000	0.888	0.894	0.931	0.888	0.887	0.896	0.906	0.903	0.923
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.436	0.398	0.522	0.436	0.441	0.410	0.362	0.291	0.431
500	0.808	0.810	0.763	0.808	0.808	0.813	0.823	0.811	0.766
1,000	0.822	0.825	0.819	0.822	0.822	0.827	0.826	0.831	0.839
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.492	0.481	0.098	0.492	0.491	0.485	0.480	0.469	0.204
500	0.859	0.865	0.843	0.859	0.861	0.866	0.855	0.852	0.862
1,000	0.854	0.859	0.892	0.854	0.856	0.857	0.863	0.857	0.874
Plug-in (CLT) CI									
50	0.393	0.372	0.308	0.393	0.387	0.372	0.359	0.347	0.318
100	0.632	0.612	0.515	0.632	0.628	0.611	0.594	0.560	0.515
500	0.877	0.885	0.861	0.877	0.877	0.884	0.886	0.877	0.862
1,000	0.869	0.876	0.866	0.869	0.869	0.873	0.882	0.878	0.886

Table 13: Empirical coverage (nominal 90%) for a fixed target policy π (mostly-left: $\pi(0 \mid s) = 0.8$) in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 50$.

	Q^π coverage			V^π coverage					
n	$Q^\pi(1,0)$	$Q^\pi(3,1)$	$Q^\pi(6,0)$	$V^\pi(1)$	$V^\pi(2)$	$V^\pi(3)$	$V^\pi(4)$	$V^\pi(5)$	$V^\pi(6)$
Model-based bootstrap with pivot CI									
50	0.821	0.694	0.280	0.821	0.802	0.684	0.468	0.292	0.274
100	0.878	0.911	0.567	0.878	0.884	0.873	0.794	0.668	0.565
500	0.907	0.909	0.922	0.907	0.907	0.911	0.906	0.937	0.910
1,000	0.892	0.901	0.930	0.892	0.889	0.897	0.906	0.930	0.922
Model-based bootstrap with percentile CI									
50	0.721	0.654	0.377	0.721	0.696	0.594	0.356	0.224	0.391
100	0.865	0.808	0.501	0.865	0.850	0.819	0.740	0.644	0.545
500	0.908	0.905	0.858	0.908	0.905	0.905	0.904	0.922	0.871
1,000	0.898	0.895	0.905	0.898	0.893	0.897	0.897	0.907	0.916
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.560	0.598	0.490	0.560	0.564	0.578	0.545	0.419	0.490
500	0.849	0.870	0.892	0.849	0.861	0.861	0.865	0.879	0.877
1,000	0.871	0.888	0.905	0.871	0.870	0.866	0.875	0.907	0.895
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.504	0.501	0.371	0.504	0.501	0.484	0.451	0.406	0.379
500	0.856	0.860	0.833	0.856	0.854	0.851	0.864	0.895	0.829
1,000	0.874	0.873	0.899	0.874	0.883	0.881	0.885	0.903	0.898
Plug-in (CLT) CI									
50	0.793	0.621	0.144	0.793	0.769	0.635	0.421	0.276	0.155
100	0.889	0.842	0.288	0.889	0.885	0.861	0.777	0.620	0.276
500	0.908	0.903	0.822	0.908	0.909	0.906	0.907	0.868	0.818
1,000	0.902	0.901	0.890	0.902	0.896	0.903	0.897	0.893	0.889

E.4 OPR: Additional Nominal Levels (50% and 90%)

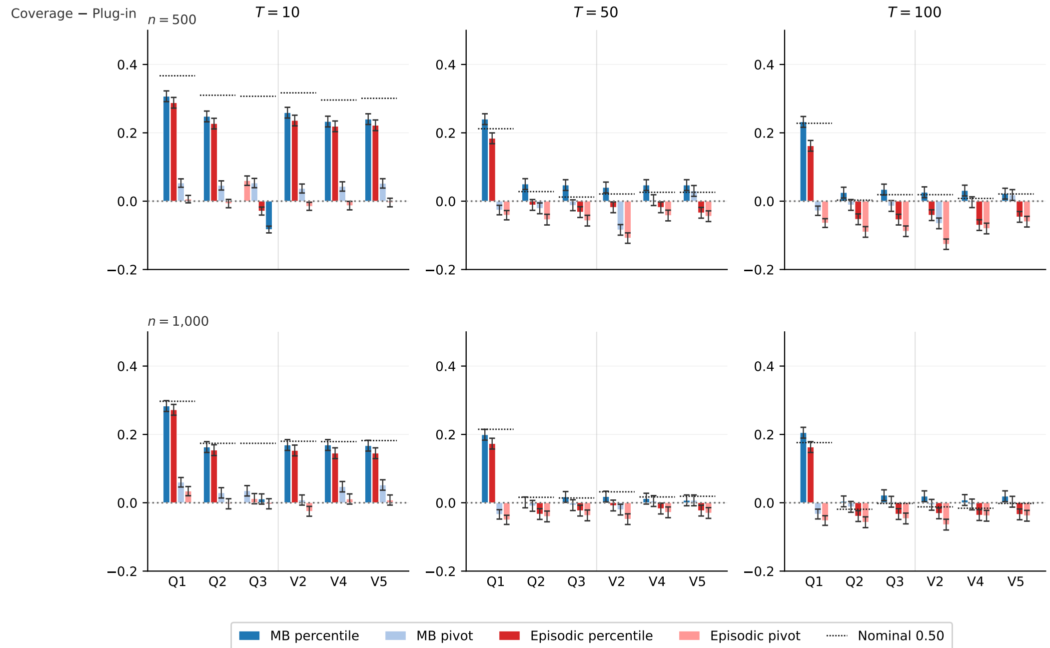

Figure 6: Empirical coverage relative to the Plug-in (CLT) baseline for $Q_\star(1,0)$, $Q_\star(3,1)$, $Q_\star(6,0)$, $V_\star(2)$, $V_\star(4)$, $V_\star(5)$ in RiverSwim at nominal 50%. Horizontal lines mark the nominal 0.50 level for each entry. Columns correspond to $T \in \{10, 50, 100\}$, rows correspond to $n \in \{500, 1{,}000\}$. Full results appear in Tables 14–18.

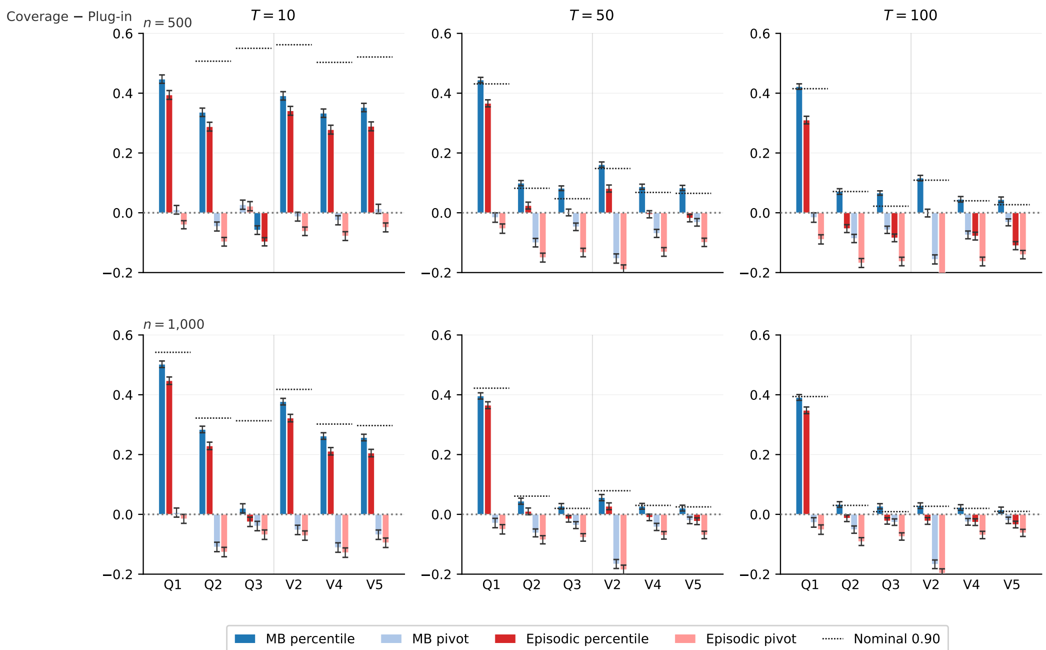

Figure 7: Empirical coverage relative to the Plug-in (CLT) baseline for $Q_\star(1,0)$, $Q_\star(3,1)$, $Q_\star(6,0)$, $V_\star(2)$, $V_\star(4)$, $V_\star(5)$ in RiverSwim at nominal 90%. Horizontal lines mark the nominal 0.90 level for each entry. Columns correspond to $T \in \{10, 50, 100\}$, rows correspond to $n \in \{500, 1{,}000\}$. Full results appear in Tables 15–19.

Table 14: Empirical coverage (nominal 50%) for $Q_\star$ and $V_\star$ in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 10$.

	$Q_\star$ coverage			$V_\star$ coverage					
n	$Q_\star(1,0)$	$Q_\star(3,1)$	$Q_\star(6,0)$	$V_\star(1)$	$V_\star(2)$	$V_\star(3)$	$V_\star(4)$	$V_\star(5)$	$V_\star(6)$
Model-based bootstrap with pivot CI									
50	0.013	0.177	0.028	0.013	0.102	0.245	0.017	0.016	0.027
100	0.037	0.087	0.066	0.037	0.053	0.108	0.053	0.055	0.057
500	0.186	0.236	0.246	0.186	0.220	0.234	0.247	0.251	0.223
1,000	0.263	0.355	0.361	0.263	0.328	0.351	0.368	0.370	0.326
Model-based bootstrap with percentile CI									
50	0.102	0.102	0.000	0.102	0.101	0.102	0.102	0.102	0.001
100	0.203	0.201	0.002	0.203	0.202	0.201	0.201	0.196	0.015
500	0.440	0.438	0.110	0.440	0.442	0.438	0.437	0.439	0.206
1,000	0.486	0.489	0.337	0.486	0.489	0.489	0.490	0.485	0.427
Episodic bootstrap with pivot CI									
50	0.002	0.222	0.043	0.002	0.106	0.293	0.008	0.010	0.040
100	0.019	0.098	0.095	0.019	0.036	0.125	0.030	0.034	0.074
500	0.139	0.183	0.253	0.139	0.168	0.180	0.191	0.195	0.210
1,000	0.237	0.323	0.338	0.237	0.295	0.320	0.332	0.326	0.298
Episodic bootstrap with percentile CI									
50	0.105	0.106	0.006	0.105	0.105	0.106	0.104	0.101	0.011
100	0.205	0.205	0.019	0.205	0.204	0.205	0.206	0.203	0.045
500	0.421	0.417	0.164	0.421	0.419	0.417	0.423	0.421	0.220
1,000	0.475	0.480	0.323	0.475	0.473	0.480	0.466	0.463	0.402
Plug-in (CLT) CI									
50	0.011	0.014	0.011	0.011	0.016	0.014	0.013	0.026	0.025
100	0.036	0.049	0.044	0.036	0.049	0.049	0.049	0.055	0.048
500	0.133	0.190	0.193	0.133	0.183	0.190	0.204	0.199	0.183
1,000	0.203	0.326	0.326	0.203	0.320	0.326	0.321	0.318	0.301

Table 15: Empirical coverage (nominal 90%) for $Q_\star$ and $V_\star$ in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 10$.

	$Q_\star$ coverage			$V_\star$ coverage					
n	$Q_\star(1,0)$	$Q_\star(3,1)$	$Q_\star(6,0)$	$V_\star(1)$	$V_\star(2)$	$V_\star(3)$	$V_\star(4)$	$V_\star(5)$	$V_\star(6)$
Model-based bootstrap with pivot CI									
50	0.047	0.479	0.047	0.047	0.407	0.620	0.052	0.052	0.045
100	0.098	0.364	0.106	0.098	0.155	0.464	0.118	0.118	0.098
500	0.293	0.347	0.377	0.293	0.325	0.341	0.372	0.392	0.337
1,000	0.364	0.469	0.548	0.364	0.430	0.461	0.487	0.535	0.492
Model-based bootstrap with percentile CI									
50	0.115	0.115	0.011	0.115	0.115	0.115	0.115	0.115	0.025
100	0.237	0.237	0.035	0.237	0.237	0.237	0.237	0.237	0.067
500	0.730	0.729	0.292	0.730	0.729	0.729	0.730	0.731	0.386
1,000	0.860	0.862	0.607	0.860	0.859	0.862	0.860	0.860	0.664
Episodic bootstrap with pivot CI									
50	0.011	0.441	0.047	0.011	0.413	0.574	0.010	0.012	0.042
100	0.041	0.288	0.108	0.041	0.106	0.395	0.056	0.049	0.083
500	0.243	0.296	0.372	0.243	0.276	0.291	0.319	0.330	0.324
1,000	0.343	0.452	0.519	0.343	0.411	0.444	0.470	0.508	0.470
Episodic bootstrap with percentile CI									
50	0.115	0.114	0.006	0.115	0.114	0.114	0.113	0.109	0.011
100	0.232	0.232	0.030	0.232	0.232	0.232	0.232	0.231	0.054
500	0.677	0.681	0.253	0.677	0.679	0.681	0.675	0.668	0.333
1,000	0.805	0.807	0.562	0.805	0.804	0.807	0.809	0.808	0.629
Plug-in (CLT) CI									
50	0.034	0.038	0.034	0.034	0.040	0.038	0.034	0.034	0.041
100	0.078	0.096	0.078	0.078	0.090	0.096	0.089	0.078	0.090
500	0.283	0.393	0.350	0.283	0.338	0.393	0.397	0.379	0.370
1,000	0.358	0.578	0.587	0.358	0.482	0.578	0.598	0.603	0.558

Table 16: Empirical coverage (nominal 50%) for $Q_\star$ and $V_\star$ in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 50$.

	$Q_\star$ coverage			$V_\star$ coverage					
n	$Q_\star(1,0)$	$Q_\star(3,1)$	$Q_\star(6,0)$	$V_\star(1)$	$V_\star(2)$	$V_\star(3)$	$V_\star(4)$	$V_\star(5)$	$V_\star(6)$
Model-based bootstrap with pivot CI									
50	0.100	0.234	0.191	0.100	0.181	0.251	0.154	0.161	0.176
100	0.164	0.237	0.335	0.164	0.212	0.232	0.256	0.264	0.309
500	0.262	0.451	0.476	0.262	0.395	0.451	0.477	0.504	0.490
1,000	0.251	0.475	0.479	0.251	0.448	0.475	0.488	0.488	0.482
Model-based bootstrap with percentile CI									
50	0.292	0.289	0.101	0.292	0.289	0.289	0.291	0.288	0.151
100	0.445	0.437	0.225	0.445	0.449	0.437	0.436	0.441	0.298
500	0.528	0.522	0.535	0.528	0.519	0.522	0.521	0.521	0.538
1,000	0.484	0.485	0.503	0.484	0.486	0.485	0.495	0.488	0.492
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.082	0.121	0.199	0.082	0.116	0.121	0.109	0.122	0.154
500	0.247	0.418	0.431	0.247	0.371	0.417	0.432	0.430	0.424
1,000	0.235	0.444	0.449	0.235	0.420	0.444	0.455	0.451	0.447
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.225	0.221	0.139	0.225	0.214	0.221	0.221	0.219	0.174
500	0.472	0.461	0.456	0.472	0.461	0.461	0.456	0.440	0.451
1,000	0.458	0.451	0.463	0.458	0.460	0.451	0.466	0.458	0.457
Plug-in (CLT) CI									
50	0.102	0.171	0.163	0.102	0.156	0.171	0.166	0.173	0.165
100	0.180	0.278	0.261	0.180	0.269	0.278	0.278	0.275	0.271
500	0.288	0.472	0.488	0.288	0.479	0.472	0.474	0.474	0.473
1,000	0.285	0.484	0.486	0.285	0.468	0.484	0.483	0.481	0.484

Table 17: Empirical coverage (nominal 90%) for $Q_\star$ and $V_\star$ in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 50$.

	$Q_\star$ coverage			$V_\star$ coverage					
n	$Q_\star(1,0)$	$Q_\star(3,1)$	$Q_\star(6,0)$	$V_\star(1)$	$V_\star(2)$	$V_\star(3)$	$V_\star(4)$	$V_\star(5)$	$V_\star(6)$
Model-based bootstrap with pivot CI									
50	0.230	0.633	0.298	0.230	0.531	0.650	0.300	0.295	0.278
100	0.333	0.476	0.487	0.333	0.422	0.474	0.453	0.468	0.461
500	0.453	0.718	0.806	0.453	0.599	0.694	0.763	0.803	0.809
1,000	0.449	0.777	0.844	0.449	0.655	0.762	0.828	0.855	0.861
Model-based bootstrap with percentile CI									
50	0.421	0.418	0.251	0.421	0.419	0.418	0.421	0.415	0.303
100	0.648	0.649	0.466	0.648	0.651	0.649	0.652	0.651	0.520
500	0.913	0.917	0.935	0.913	0.913	0.917	0.919	0.918	0.922
1,000	0.874	0.883	0.907	0.874	0.877	0.883	0.897	0.896	0.902
Episodic bootstrap with pivot CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.159	0.251	0.369	0.159	0.225	0.247	0.213	0.225	0.290
500	0.416	0.668	0.720	0.416	0.562	0.652	0.701	0.736	0.733
1,000	0.428	0.754	0.803	0.428	0.636	0.742	0.800	0.806	0.819
Episodic bootstrap with percentile CI									
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	0.434	0.438	0.225	0.434	0.432	0.438	0.430	0.428	0.293
500	0.835	0.842	0.854	0.835	0.833	0.842	0.827	0.817	0.847
1,000	0.843	0.849	0.865	0.843	0.848	0.849	0.860	0.852	0.858
Plug-in (CLT) CI									
50	0.231	0.300	0.296	0.231	0.288	0.299	0.302	0.295	0.294
100	0.349	0.490	0.484	0.349	0.440	0.490	0.495	0.477	0.468
500	0.469	0.818	0.853	0.469	0.752	0.818	0.832	0.835	0.823
1,000	0.478	0.839	0.880	0.478	0.821	0.839	0.870	0.875	0.880

Table 18: Empirical coverage (nominal 50%) for $Q_\star$ and $V_\star$ in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 100$.

	$Q_\star$ coverage			$V_\star$ coverage					
n	$Q_\star(1,0)$	$Q_\star(3,1)$	$Q_\star(6,0)$	$V_\star(1)$	$V_\star(2)$	$V_\star(3)$	$V_\star(4)$	$V_\star(5)$	$V_\star(6)$
Model-based bootstrap with pivot CI									
100	0.156	0.226	0.343	0.156	0.207	0.219	0.263	0.278	0.311
500	0.244	0.486	0.467	0.244	0.416	0.484	0.489	0.497	0.496
1,000	0.291	0.507	0.505	0.291	0.506	0.507	0.511	0.505	0.506
Model-based bootstrap with percentile CI									
100	0.447	0.444	0.276	0.447	0.451	0.444	0.452	0.453	0.339
500	0.504	0.522	0.515	0.504	0.507	0.522	0.523	0.501	0.511
1,000	0.529	0.523	0.524	0.529	0.531	0.523	0.524	0.521	0.516
Episodic bootstrap with pivot CI									
100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
500	0.208	0.407	0.393	0.208	0.355	0.404	0.412	0.419	0.410
1,000	0.272	0.462	0.456	0.272	0.448	0.462	0.478	0.464	0.461
Episodic bootstrap with percentile CI									
100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
500	0.434	0.444	0.427	0.434	0.440	0.444	0.422	0.433	0.421
1,000	0.487	0.480	0.469	0.487	0.481	0.480	0.480	0.468	0.469
Plug-in (CLT) CI									
100	0.176	0.284	0.301	0.176	0.277	0.284	0.294	0.303	0.276
500	0.272	0.497	0.481	0.272	0.481	0.497	0.492	0.479	0.478
1,000	0.324	0.519	0.502	0.324	0.512	0.519	0.516	0.502	0.506

Table 19: Empirical coverage (nominal 90%) for $Q_\star$ and $V_\star$ in RiverSwim. Comparison of model-based bootstrap, episodic bootstrap, and plug-in (CLT) CI, episode length $T = 100$.

	$Q_\star$ coverage			$V_\star$ coverage					
n	$Q_\star(1,0)$	$Q_\star(3,1)$	$Q_\star(6,0)$	$V_\star(1)$	$V_\star(2)$	$V_\star(3)$	$V_\star(4)$	$V_\star(5)$	$V_\star(6)$
Model-based bootstrap with pivot CI									
100	0.333	0.492	0.512	0.333	0.460	0.485	0.477	0.515	0.483
500	0.469	0.743	0.821	0.469	0.635	0.726	0.786	0.841	0.843
1,000	0.479	0.819	0.865	0.479	0.706	0.811	0.855	0.870	0.875
Model-based bootstrap with percentile CI									
100	0.662	0.662	0.529	0.662	0.661	0.662	0.660	0.667	0.580
500	0.907	0.900	0.944	0.907	0.907	0.900	0.905	0.917	0.937
1,000	0.897	0.903	0.918	0.897	0.902	0.903	0.903	0.905	0.899
Episodic bootstrap with pivot CI									
100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
500	0.396	0.661	0.715	0.396	0.566	0.648	0.697	0.733	0.733
1,000	0.455	0.779	0.817	0.455	0.676	0.774	0.811	0.828	0.813
Episodic bootstrap with percentile CI									
100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
500	0.795	0.776	0.794	0.795	0.790	0.776	0.782	0.763	0.786
1,000	0.854	0.857	0.869	0.854	0.851	0.857	0.854	0.856	0.857
Plug-in (CLT) CI									
100	0.355	0.539	0.537	0.355	0.468	0.539	0.540	0.524	0.516
500	0.485	0.829	0.878	0.485	0.791	0.829	0.860	0.873	0.861
1,000	0.506	0.870	0.891	0.506	0.873	0.870	0.880	0.890	0.885