SUPPLEMENTARY MATERIAL

Online Bootstrap Inference For Policy Evaluation in Reinforcement Learning

In this online supplementary material, we provide detailed proofs for the lemmas and main theorems in Sections 4.1 and 4.2, as well as additional experiments.

S1 Proofs for Section 4.1

The following lemma is a restatement of Theorem 2 from Chong et al. (1999). In the following, we say that a sequence ϵ_t is *small* with respect to another sequence α_t if there exist sequences $\{e_t\}$ and $\{r_t\}$ such that $\epsilon_t = e_t + r_t$ for all t, $r_t \to 0$, and $\sum_{k=1}^{t} \alpha_t \|e_t\|_2$ converges. Also, we say that a scalar sequence $\{a_t\}$ has *bounded variation* if $\sum_{t=1}^{\infty} |a_{t+1} - a_t| < \infty$. We refer to the cited article for further details on these conditions.

Lemma S1.1. *Consider the linear stochastic approximation update*

$$\theta_{t+1} = \theta_t + \alpha_{t+1}(\tilde{A}(X_{t+1})\theta_t - \tilde{b}(X_{t+1})).$$

Assume the following conditions hold:

(B1) The step size sequence $\{\alpha_t\}$ *satisfies* $\alpha_t > 0$, $\alpha_t \to 0$, *and* $\sum_{t=1}^{\infty} \alpha_t = \infty$.

(B2) $\bar{A}$ *is a bounded Hurwitz matrix.*

(B3) $\{\tilde{A}(X_t) - \bar{A}\}$ *is small with respect to* α_t.

(B4) Let $\{\rho_t\}$ *be a positive real sequence converging monotonically to 0, such that*

(i) $\{\rho_t^{-1}(\tilde{b}(X_t) - \bar{b})\}$ *is small with respect to* α_t.

(ii) $(\rho_t - \rho_{t+1})/(\alpha_t \rho_t) \to c < \infty$,

(iii) The sequences $\{\rho_{t+1}/\rho_t\}$ *and* $\{\rho_t/\rho_{t+1}\}$ *have bounded variation.*

Then $\theta_t - \theta_* = o(\rho_t)$.

Proof of Proposition 4.1

We verify that the conditions of Lemma S1.1 hold under our assumptions (A1), (A2), (A3).

Firstly, it is easy to see that (B1) holds under (A3), i.e., with a step size $\alpha_t = \alpha_0 t^{-\eta}$, $\eta \in (1/2, 1)$. Similarly, (B2) follows directly from assumption (A2).

For functions f defined over the state space $\mathcal{X}$, we define the t-step transition operator $\mathcal{P}^t f(x) = \int_{y \in \mathcal{X}} f(y) \mathcal{P}^t(x, dy)$, where $\mathcal{P}^t(x, y)$ denotes the t-step transition probability from state x to y. When $t = 1$, we write $\mathcal{P}^1 f(x) = \mathcal{P} f(x)$.

Next, define $\hat{A} : \mathcal{X} \to \mathbb{R}^{d \times d}$ and $\hat{b} : \mathcal{X} \to \mathbb{R}^d$ to be the solutions to the Poisson equations

$$\tilde{A}(x) - \bar{A}(x) = \hat{A}(x) - \mathcal{P}\hat{A}(x),$$
$$\tilde{b}(x) - \bar{b}(x) = \hat{b}(x) - \mathcal{P}\hat{b}(x),$$

for $x \in \mathcal{X}$. The existence of $\hat{A}$ and $\hat{b}$ is guaranteed under (A1). Furthermore, under (A2), there exist constants $\hat{A}_{\max}, \hat{b}_{\max > 0}$ such that $\|\hat{A}\|_F \leq \hat{A}_{\max}$ and $\|\hat{b}\|_2 \leq \hat{b}_{\max}$ (Delyon, 2000).

Then we can write $\tilde{A}(X_t) - \bar{A} = e_t + r_t$, where $e_t = \hat{A}(X_t) - \mathcal{P}\hat{A}(X_{t+1})$, and $r_t = \mathcal{P}\hat{A}(X_{t+1}) - \mathcal{P}\hat{A}(X_t)$. To verify condition (B3), it suffices to show that $\sum_{t=1}^{\infty} \alpha_t \|e_t\| < \infty$, and $\|r_t\| \to 0$, as $t \to \infty$ (Chong et al., 1999).

Let $\mathcal{F}_t = \sigma(\{X_t\})$ denote the natural filtration with respect to the Markov chain X_t. Then $\mathbb{E}[e_t | \mathcal{F}_t] = 0$, and e_t is a martingale difference sequence with respect to $\mathcal{F}_t$. Furthermore, e_t is a.s. uniformly bounded, since $\|e_t\|_F \leq 2\hat{A}_{\max}$, by construction. So $\sum_{t=1}^{\infty} \alpha_t^2 \mathbb{E}[\|e_t\|^2 | \mathcal{F}_t] < \infty$. Then, by Theorem 29 of Delyon (2000), $\sum_{t=1}^{\infty} \alpha_t \|e_t\|$ converges.

Also, since $\mathcal{P}(X_t, \cdot) \to \mu$ as $t \to \infty$, it follows that $\|r_t\| \to 0$, as $t \to \infty$. So $\{\tilde{A}(X_t) - \bar{A}\}$ is small with respect to α_t, and (B3) holds.

Next, set $\rho_t = t^{-\gamma}$, with $\gamma \in (0, \eta - 1/2)$. Conditions (B4)(ii) and (B4)(iii) hold under this definition, with $c = 0$ in (B4)(ii) (Chong et al., 1999). It remains to verify (B4)(i).

Define $\tilde{b}(X_t) - \bar{b} = e_t + r_t$, where $e_t = \hat{b}(X_t) - \mathcal{P}\hat{b}(X_{t+1})$, and $r_t = \mathcal{P}\hat{b}(X_{t+1}) - \mathcal{P}\hat{b}(X_t)$. It suffices to show that $\sum_{t=1}^{\infty} \alpha_t \rho_t^{-1} \|e_t\| < \infty$, and that $\rho_t^{-1} \|r_t\| \to 0$, as $t \to \infty$.

By the same argument used above for $\{\tilde{A} - \bar{A}\}$, e_t is an a.s. uniformly bounded martingale

difference sequence, with $\|e_t\|_F \leq 2\hat{b}_{\max}$ for all t. Then $\sum_{t=1}^{\infty} \alpha_t^2 \rho_t^{-2} \mathbb{E}[\|e_t\|_2 | \mathcal{F}_t] < \infty$, since $\eta - \gamma > 1/2$. So, by Theorem 29 of Delyon (2000), $\sum_{t=1}^{\infty} \alpha_t \rho_t \|e_t\|$ converges.

Next, we have

$$\begin{aligned}
\|\mathcal{P}\hat{b}(X_{t+1}) - \mathcal{P}\hat{b}(X_t)\| &= \left\| \int_{y \in \mathcal{X}} \hat{b}(y)(\mathcal{P}(X_{t+1}, dy) - \mathcal{P}(X_t, dy)) \right\| \\
&\leq \int_{y \in \mathcal{X}} \|\hat{b}(y)\| \|\mathcal{P}(X_{t+1}, dy) - \mathcal{P}(X_t, dy)\| \\
&\leq \hat{b}_{\max} \int_{y \in \mathcal{X}} \|\mathcal{P}(X_{t+1}, dy) - \mathcal{P}(X_t, dy)\|.
\end{aligned}$$

Consider the integrand in the above expression. For any bounded initial distribution ν_0, i.e., with $\sup_{x \in \mathcal{X}} \|\nu_0(x)\| \leq \nu_{\max}$, for some constant $\nu_{\max} < \infty$, we have

$$\begin{aligned}
\|\mathcal{P}(X_{t+1}, dy) - \mathcal{P}(X_t, dy)\| &\leq \sup_{x_1, x_2 \in \mathcal{X}} \|\nu_0 \mathcal{P}^{t+1}(x_1, dy) - \nu_0 \mathcal{P}^t(x_2, dy)\| \\
&\leq \nu_{\max} \sup_{x_1, x_2 \in \mathcal{X}} \|\mathcal{P}^{t+1}(x_1, dy) - \mathcal{P}^t(x_2, dy)\| \\
&= \nu_{\max} \sup_{x_1, x_2 \in \mathcal{X}} \|(\mathcal{P}^{t+1}(x_1, dy) - \pi(dy)) - (\mathcal{P}^t(x_2, dy) - \pi(dy))\| \\
&\leq \nu_{\max} \sup_{x_1, x_2 \in \mathcal{X}} \left(\|\mathcal{P}^{t+1}(x_1, dy) - \pi(dy)\| + \|\mathcal{P}^t(x_2, dy) - \pi(dy)\| \right) \\
&\leq \nu_{\max} \left(M\kappa^{t+1} + M\kappa^t \right) \\
&\leq 2\nu_{\max} M \kappa^t,
\end{aligned}$$

where the penultimate inequality holds by (4.1). It follows that $\|\mathcal{P}\hat{b}(X_{t+1}) - \mathcal{P}\hat{b}(X_t)\| \leq 2\hat{b}_{\max}\nu_{\max} M\kappa^t$, and so $\rho_t^{-1}\|r_t\| \leq 2\hat{b}_{\max}\nu_{\max} M t^{\gamma} \kappa^t \to 0$, as $t \to \infty$. □

Proof of Proposition 4.2

First, we list the conditions required for our central limit theorem, Proposition 4.2, to hold. The assumptions listed below are from Liang (2010), who proved a central limit theorem for the varying truncation stochastic approximation MCMC algorithm. This is a general form of algorithm (2.1), and is designed to solve the equation

$$h(\theta) = \int_{\mathcal{X}} H(\theta, x) f_\theta(x) dx = 0,$$

where $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ is a parameter vector and $f_\theta(x), x \in \mathcal{X} \subset \mathbb{R}^{d_x}$ is a density function depending on θ. The function $h(\theta)$ is called the mean field function, and $H(\theta, x)$ is a noisy observation of $h(\theta)$.

The stochastic approximation algorithm is designed to iteratively estimate θ from a sequence of noisy observations that depend on the current estimate of θ (hence forming a controlled Markov chain). The main update step for this algorithm is given by

$$\begin{aligned} \theta_{t+1} &= \theta_t + \alpha_{t+1} H(\theta_t, X_{t+1}) \\ &= \theta_t + \alpha_{t+1} h(\theta_t) + \alpha_{t+1} \epsilon_{t+1}, \end{aligned} \tag{S1.1}$$

where $h(\theta) = \int H(\theta, x) f_\theta(x) dx$, f_θ being the invariant distribution of the controlled Markov transition kernel $\mathcal{P}_\theta$, and $\epsilon_{t+1} = H(\theta_t, X_{t+1}) - h(\theta_t)$ is the residual noise term.

In order to ensure the convergence of the iterates in (S1.1), Liang (2010) imposes a varying truncation scheme, whereby the iterates θ_t are constrained within an increasing sequence of compact sets $\{\mathcal{K}_s\}_{s \geq 0}$. Under this scheme, Andrieu et al. (2005) showed that there exists a time step $t_{\sigma_s} < \infty$ such that $\theta_t \in \mathcal{K}_{\sigma_s}$ for all $t \geq t_{\sigma_s}$, and there are no further truncations beyond time step t_{σ_s}. The central limit theorem applies to the averaged iterate $\bar{\theta}_t := \frac{1}{t - t_{\sigma_s}} \sum_{i=t_{\sigma_s}+1}^{t} \theta_i$.

The following conditions are assumed by Liang (2010):

(C1) Θ is an open set, the function $h : \Theta \to \mathbb{R}^d$ is continuous, and there exists a continuously differential function $v : \Theta \to [0, \infty)$ such that

(i) There exists $M_0 > 0$ such that

$$\mathcal{L} = \{\theta \in \Theta, \langle \nabla v(\theta), h(\theta) \rangle = 0\} \subset \{\theta \in \Theta, v(\theta) < M_0\}.$$

(ii) There exists $M_1 \in (M_0, \infty)$ such that $\mathcal{V}_{M_1}$ is a compact set, where $\mathcal{V}_M = \{\theta \in \Theta, v(\theta) \leq M\}$.

(iii) For any $\theta \in \Theta \backslash \mathcal{L}$, $\langle \nabla v(\theta), h(\theta) \rangle < 0$.

(iv) The closure of $v(\mathcal{L})$ has an empty interior.

(C2) The mean field $h(\theta)$ is measurable and locally bounded. There exists a Hurwitz matrix F, $\gamma > 0, \rho \in (0, 1]$, and a constant c such that, for any $\theta_* \in \mathcal{L}$,

$$\|h(\theta) - F(\theta - \theta_*)\| \leq c\|\theta - \theta^*\|^{1+\rho} \quad \forall \theta \in \{\theta : \|\theta - \theta_*\| \leq \gamma\},$$

where $\mathcal{L}$ is defined in (B1)(i).

(C3) For any $\theta \in \Theta$, the transition kernel $\mathcal{P}_\theta$ is irreducible and aperiodic. In addition, there exists a function $V : \mathcal{X} \to [1, \infty)$, and a constant $\alpha \geq 2$ such that for any compact set $\mathcal{K} \subset \Theta$:

(i) There exists a set $\mathbf{C} \subset \mathcal{X}$, and integer l, constants $0 < \lambda < 1, b, \zeta, \delta > 0$ and a probability measure ν such that

$$\begin{aligned}
&\sup_{\theta \in \mathcal{K}} \mathcal{P}_\theta^l V^\alpha(x) \leq \lambda V^\alpha(x) + bI(x \in \mathbf{C}) \quad \forall x \in \mathcal{X}, \\
&\sup_{\theta \in \mathcal{K}} \mathcal{P}_\theta V^\alpha(x) \leq \zeta V^\alpha(x) \quad \forall x \in \mathcal{X}, \\
&\inf_{\theta \in \mathcal{K}} \mathcal{P}_\theta^l(x, A) \geq \delta \nu(A) \quad \forall x \in \mathbf{C}, \forall A \in \mathcal{B}_\mathcal{X}.
\end{aligned}$$

(ii) There exists a constant $c > 0$ such that, for all $x \in \mathcal{X}$,

$$\begin{aligned}
&\sup_{\theta \in \mathcal{K}} \|H(\theta, x)\|_V \leq c, \\
&\sup_{\theta, \theta' \in \mathcal{K}} \|H(\theta, x) - H(\theta', x)\|_V \leq c\|\theta - \theta'\|.
\end{aligned}$$

(iii) There exists a constant $c > 0$ such that, for all $\theta, \theta' \in \mathcal{K}$,

$$\begin{aligned}
&\|\mathcal{P}_\theta g - \mathcal{P}_{\theta'}\|_V \leq c\|g\|_V\|\theta - \theta'\| \quad \forall g \in \mathcal{L}_V, \\
&\|\mathcal{P}_\theta g - \mathcal{P}_{\theta'} g\|_{V^\alpha} \leq c\|g\|_{V^\alpha}\|\theta - \theta'\| \quad \forall g \in \mathcal{L}_{V^\alpha}.
\end{aligned}$$

(C4) The step sizes $\{\alpha_t\}$ are non-increasing, positive sequences that satisfy the conditions

$$\sum_{t=1}^{\infty} \alpha_t = \infty, \quad \lim_{t \to \infty}(t\alpha_t) = \infty, \quad \frac{\alpha_{t+1} - \alpha_t}{\alpha_t} = o(\alpha_{t+1}), \quad \sum_{t=1}^{\infty} \frac{\alpha_t^{(1+\tau)/2}}{\sqrt{t}} < \infty,$$

for some $\tau \in (0, 1]$ and a constant $\alpha \geq 2$ defined in (B3).

We refer to Liang (2010) for further details on these conditions.

We now verify that (C1)-(C4) hold under assumptions (A1)-(A3). We also show that, under our assumptions, the iterates of the update (2.1) are constrained within a compact set $\mathcal{K} \subset \Theta$, thereby avoiding the need for the varying truncation scheme. Then the result directly follows. Using the notation of (S1.1), in our case, we have $H(\theta, x) = \tilde{A}(x)\theta - \tilde{b}$, with mean field $h(\theta) = \bar{A}\theta - \bar{b}$. By (A2)(iii), we have $\mathbb{E}_{X\sim\mu}[H(\theta, X)] = h(\theta)$, for all $\theta \in \Theta$.

(C1) assumes the existence of a global Lyapunov function v. We may choose $v(\theta) = \theta^\intercal \bar{b} - \frac{1}{2}\theta^\intercal \bar{A}\theta$. Then v is a global Lyupanov function for the mean field h (Andrieu et al., 2005; Liang, 2010). $\mathcal{L}$ denotes the set of all valid solutions θ^* for the equation $h(\theta) = 0$. In our case, since $\bar{A}$ is Hurwitz by (A2)(ii), there exists a unique solution θ^* for the linear system $\bar{A}\theta = \bar{b}$, and $\mathcal{L}$ is a singleton set.

For (C2), the measurability and local boundedness of h follows directly from linearity. For the latter part, we may choose $F = A$. Then for any $\theta \in \Theta$, we have $\|h(\theta) - F(\theta - \theta^*)\| \equiv 0$, so (C2) holds.

For (C3), in our case the function $H(\theta, x) = \tilde{A}(x)\theta - \tilde{b}(x)$ is bounded by (A2)(ii), so we can choose the drift function $V \equiv 1$. Then the first two conditions of (C3)(i) hold trivially.

The third condition in (C3)(i) is a standard assumption in the Markov Chain Monte Carlo (MCMC) literature, and is referred to as the minorization condition. By Theorem 5.2.2 of Meyn and Tweedie (2009), for φ-irreducible Markov chains, "small sets" for which the minorization condition holds exist. By (A1), the Markov chain is irreducible, and so, by definition, is φ-irreducible for some irreducibility measure φ. Hence the condition holds in our case.

(C3)(ii) follows directly from (A2)(ii). (C3)(iii) does not apply in our case as we are dealing with a homogeneous Markov chain that does not depend on θ_k (not have a controlled Markov chain). The conditions of (C4) hold trivially under (A3).

Finally, by Proposition 4.1, we can choose a large enough constant $R_\theta > 0$ such that $\|\theta_t - \theta_*\|_2 \leq R_\theta$ for all $t \geq 0$. Then $\theta_t \in \mathcal{K}$ for all $t \geq 0$, for some compact set $\mathcal{K} \subset \Theta$. $\square$

S2 Proofs for Section 4.2

Proof of Proposition 4.3

To verify the conditions for Lemma S1.1, it suffices to check that $\{W_t\tilde{A}(X_t) - \bar{A}\}$ and $\{\rho_t^{-1}(W_t\tilde{b}(X_t) - \bar{b})\}$ are small with respect to the step sizes $\{\alpha_t\}$. By (A2)(ii) and the boundedness of W_t, we have $\|W_t\tilde{A}(X_t)\|_F \le W_{\max}A_{\max} < \infty$, and $\|W_t\tilde{b}(X_t)\| \le W_{\max}b_{\max} < \infty$, for all $t \ge 1$. By independence of W_t, we have $\mathbb{E}_\mu[W_t\tilde{A}(X_t) - \bar{A}] = 0$ and $\mathbb{E}_\mu[W_t\tilde{b}(X_t) - \bar{b}] = 0$. The rest of the argument is identical to the proof of Proposition 4.1. □

Lemma S2.1. *Assume (A1)-(A3) hold. Then*

$$\sqrt{t}(\bar{\hat{\theta}}_t - \theta_*) = -\bar{A}^{-1}\frac{1}{\sqrt{t}}\sum_{i=1}^{t}\hat{\epsilon}_{i+1} + o_p(1).$$

Proof. An argument similar to above may be used to verify that conditions (C1)-(C4) also hold for the perturbed SA update (3.1) under assumptions (A1)-(A3). Then the result follows as an intermediate step in the proof of Theorem 2.2 by Liang (2010). □

The following lemma is adapted from Lemma 5 of Xu et al. (2020).

Lemma S2.2. *Assume (A1)-(A3) hold. Then, for any $i > j$, we have*

$$\left\|\mathbb{E}\left[\tilde{A}(X_i)|\mathcal{F}_j\right] - \bar{A}\right\|_F \le A_{\max}M\kappa^{i-j},$$

where M and κ refer to the constants from (4.1).

Proof. By (4.1), for any $i > j$, the following holds:

$$\left\|\mathcal{P}^{i-j}(\cdot|\mathcal{F}_j) - \mu\right\| \le M\kappa^{i-j}, \tag{S2.1}$$

Then we have

$$\begin{aligned}
\left\|\mathbb{E}\left[\tilde{A}(X_i)|\mathcal{F}_j\right] - \bar{A}\right\|_F &= \left\|\int_{x\in\mathcal{X}}\tilde{A}(x)\mathcal{P}^{i-j}(dx|\mathcal{F}_j) - \int_{x\in\mathcal{X}}\tilde{A}(x)\mu(dx)\right\|_F \\
&\le \int_{x\in\mathcal{X}}\left\|\tilde{A}(x)\mathcal{P}^{i-j}(dx|\mathcal{F}_j) - \tilde{A}(x)\mu(dx)\right\|_F \\
&\le \int_{x\in\mathcal{X}}\left\|\tilde{A}(x)\right\|_F\left\|\mathcal{P}^{i-j}(dx|\mathcal{F}_j) - \mu(dx)\right\| \\
&\le A_{\max}M\kappa^{i-j}.
\end{aligned}$$

The first equality follows from the definition of $\bar{A}$ in (A2)(i), the second step holds by Jensen's inequality, and the final step follows from (A2)(iii) and (S2.1). □

Proof of Lemma 4.1

Starting with (4.4), we have

$$\begin{aligned}\hat{\epsilon}_{t+1} &= (W_{t+1}\tilde{A}(X_{t+1}) - \bar{A})\hat{\theta}_t - (W_{t+1}\tilde{b}(X_{t+1}) - \bar{b}) \\ &= W_{t+1}(\tilde{A}(X_{t+1})\theta_* - \tilde{b}(X_{t+1})) + (W_{t+1}\tilde{A}(X_{t+1}) - \bar{A})(\hat{\theta}_t - \theta_*). \qquad \text{(S2.2)}\end{aligned}$$

using the fact that $\bar{A}\theta_* = \bar{b}$.

By Lemma S2.1 and (S2.2), we have

$$\begin{aligned}&\sqrt{t}(\bar{\hat{\theta}}_t - \theta_*) = -\bar{A}^{-1}\frac{1}{\sqrt{t}}\sum_{i=1}^{t}\hat{\epsilon}_{i+1} + o_p(1) \\ &= -\bar{A}^{-1}\frac{1}{\sqrt{t}}\sum_{i=1}^{t}W_{i+1}(\tilde{A}(X_{i+1})\theta_* - \tilde{b}(X_{i+1})) - \bar{A}^{-1}\frac{1}{\sqrt{t}}\sum_{i=1}^{t}(W_{i+1}\tilde{A}(X_{i+1}) - \bar{A})(\hat{\theta}_i - \theta_*) + o_p(1).\end{aligned}$$

Consider the second term in the above expression. We want to show that this term is $o_p(1)$. It suffices to show that its second moment vanishes as $t \to \infty$. First we expand the second moment and split it into square and cross terms. We have

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}\sum_{i=1}^{t}\left(W_{i+1}\tilde{A}(X_{i+1}) - \bar{A}\right)\left(\hat{\theta}_i - \theta_*\right)\right\|_2^2\right] \\ &= \frac{1}{t}\sum_{i=1}^{t}\sum_{j=1}^{t}\mathbb{E}\left[\left\langle\left(W_{i+1}\tilde{A}(X_{i+1}) - \bar{A}\right)\left(\hat{\theta}_i - \theta_*\right), \left(W_{j+1}\tilde{A}(X_{j+1}) - \bar{A}\right)\left(\hat{\theta}_j - \theta_*\right)\right\rangle\right] \\ &= \frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[\left\|\left(W_{i+1}\tilde{A}(X_{i+1}) - \bar{A}\right)\left(\hat{\theta}_i - \theta_*\right)\right\|_2^2\right] \\ &+ \frac{1}{t}\sum_{i\neq j}\mathbb{E}\left[\left\langle\left(W_{i+1}\tilde{A}(X_{i+1}) - \bar{A}\right)\left(\hat{\theta}_i - \theta_*\right), \left(W_{j+1}\tilde{A}(X_{j+1}) - \bar{A}\right)\left(\hat{\theta}_j - \theta_*\right)\right\rangle\right] \\ &= I_1 + I_2.\end{aligned}$$

We deal with each term separately. First, we have

$$I_1 = \frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[(\hat{\theta}_i - \theta_*)^\intercal (W_{i+1}\tilde{A}(X_{i+1}) - \bar{A})^\intercal (W_{i+1}\tilde{A}(X_{i+1}) - \bar{A})(\hat{\theta}_i - \theta_*)\right]$$
$$\leq \frac{\lambda_A}{t}\sum_{i=1}^{t}\mathbb{E}\left[\left\|\hat{\theta}_i - \theta_*\right\|_2^2\right] \to 0,$$

since $\hat{\theta}_i \to \theta_*$ a.s.-$\mathbb{P}_{\mathcal{W}|\mathcal{D}}$, by Proposition 4.3. Here, $\lambda_A = \sup_{x\in\mathcal{X}}\left\|W_1\tilde{A}(x) - \bar{A}\right\|_2^2 < \infty$, by Assumption (A2)(ii) and the boundedness of W.

Now consider the term within the sum in I_2. Without loss of generality, assume $i > j$. Let $\mathcal{F}_j$ denote the natural filtration with respect to the Markov chain $\{X_k\}$, upto index j. Then, we have

$$\mathbb{E}\left[\left\langle\left(W_{i+1}\tilde{A}(X_{i+1}) - \bar{A}\right)\left(\hat{\theta}_i - \theta_*\right), \left(W_{j+1}\tilde{A}(X_{j+1}) - \bar{A}\right)\left(\hat{\theta}_j - \theta_*\right)\right\rangle\right]$$
$$\leq \frac{R_\theta^2}{i^\gamma j^\gamma}\mathbb{E}\left[\left\langle W_{i+1}\tilde{A}(X_{i+1}) - \bar{A}, W_{j+1}\tilde{A}(X_{j+1}) - \bar{A}\right\rangle\right]$$
$$= \frac{R_\theta^2}{i^\gamma j^\gamma}\mathbb{E}\left[\mathbb{E}\left[\langle W_{i+1}\tilde{A}(X_{i+1}) - \bar{A}, W_{j+1}\tilde{A}(X_{j+1}) - \bar{A}\rangle \,|\mathcal{F}_{j+1}\right]\right]$$
$$= \frac{R_\theta^2}{i^\gamma j^\gamma}\mathbb{E}\left[\left\langle\mathbb{E}\left[W_{i+1}\tilde{A}(X_{i+1}\,|\mathcal{F}_{j+1}\right] - \bar{A}, W_{j+1}\tilde{A}(X_{j+1}) - \bar{A}\right\rangle\right],$$

where the first step uses $\left\|\hat{\theta}_i - \theta_*\right\| \leq R_\theta i^{-\gamma}$, for some $\gamma \in (0, \eta - 1/2)$ and $R_\theta < \infty$, by Proposition 4.3, while the second step follows from the tower property, conditioning on the filtration $\mathcal{F}_{j+1}$.

Proceeding from here, we have

$$\frac{R_\theta^2}{i^\gamma j^\gamma}\mathbb{E}\left[\left\langle\mathbb{E}\left[W_{i+1}\tilde{A}(X_{i+1})\,|\mathcal{F}_{j+1}\right] - \bar{A}, W_{j+1}\tilde{A}(X_{j+1}) - \bar{A}\right\rangle\right]$$
$$\leq \frac{R_\theta^2}{i^\gamma j^\gamma}\mathbb{E}\left[\left\|\mathbb{E}\left[W_{i+1}\tilde{A}(X_{i+1})\,|\mathcal{F}_{j+1}\right] - \bar{A}\right\|_F\left\|W_{j+1}\tilde{A}(X_{j+1}) - \bar{A}\right\|_F\right]$$
$$\leq \frac{R_\theta^2}{i^\gamma j^\gamma}\mathbb{E}\left[\left\|\mathbb{E}\left[W_{i+1}A(X_{i+1})\,|\mathcal{F}_{j+1}\right] - \bar{A}\right\|_F\left(\|W_{j+1}A(X_{j+1})\|_F + \left\|\bar{A}\right\|_F\right)\right]$$
$$\leq \frac{(1+W_{\max})A_{\max}R_\theta^2}{i^\gamma j^\gamma}\mathbb{E}\left[\left\|\mathbb{E}\left[W_{i+1}A(X_{i+1})\,|\mathcal{F}_{j+1}\right] - \bar{A}\right\|_F\right]$$
$$\leq (1+W_{\max})A_{\max}^2R_\theta^2M\frac{\kappa^{i-j}}{i^\gamma j^\gamma}.$$

We first bound the inner product using Frobenius norms. In the third step, we bound the second term within the expectation using Assumption (A2)(ii) and the boundedness of W_j. The final step follows from Lemma S2.2.

So far, we have shown that

$$I_2 \leq (1 + W_{\max}) A_{\max}^2 R_\theta^2 M \cdot \frac{1}{t} \sum_{i \neq j}^{t} \frac{\kappa^{i-j}}{i^\gamma j^\gamma}.$$

Consider the double sum above. Grouping terms by $l = |i - j|$, we have

$$\sum_{i \neq j}^{t} \frac{\kappa^{i-j}}{i^\gamma j^\gamma} = 2 \sum_{l=1}^{t-1} S_{t,l} \kappa^l, \quad \text{where} \quad S_{t,l} = \sum_{j=1}^{t-l} \frac{1}{j^\gamma (j+l)^\gamma}.$$

Then $S_{t,l} \leq \sum_{j=1}^{t-l} \frac{1}{j^{2\gamma}}$. For any fixed l, $\lim_{t\to\infty} \sum_{j=1}^{t-l} \frac{1}{j^{1+2\gamma}} < \infty$, and so $\lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{t-l} \frac{1}{j^{2\gamma}} = 0$, by Kronecker's lemma. Hence, $S_{t,l}/t \to 0$, as $t \to \infty$. Then, by the Dominated Convergence Theorem, we have $\lim_{t\to\infty} \frac{1}{t} \sum_{l=1}^{t-1} S_{t,l} \kappa^l = 0$. It follows that $I_2 \to 0$ as $t \to \infty$, and so $\frac{1}{\sqrt{t}} \sum_{i=1}^{t} (W_{i+1} \tilde{A}(X_{i+1}) - \bar{A})(\hat{\theta}_i - \theta_*) = o_p(1)$. This concludes the proof. □

The following is a restatement of Lemma 2.11 from van der Vaart (1998).

Lemma S2.3. *Suppose that $X_n \implies X$ for a random vector X with a continuous distribution function. Then* $\sup_x |P(X_n \leq x) - P(X \leq x)| \to 0$.

Proof of Theorem 4.2

Let $f(x) = \tilde{A}(x)\theta_* - \tilde{b}(x)$. Then, by Assumption (A2)(ii), f is bounded, and

$$\lim_{t\to\infty} \mathbb{E}[f(X_t)] = \bar{A}\theta_* - \bar{b} = 0$$

under the stationary distribution μ.

By the Poisson equation (see e.g., Douc et al. (2018)), there exists a bounded function u such that

$$u(x) - \mathcal{P}u(x) = f(x).$$

For $t \geq 0$, we define the following terms:

$$e_{t+1} = u(X_{t+1}) - \mathcal{P}u(X_t),$$
$$r_{t+1} = \mathcal{P}u(X_t) - \mathcal{P}u(X_{t+1}).$$

Let $\mathcal{F}_t = \sigma(\{X_i\}_{i=1}^t)$ denote the natural filtration induced by the Markov chain $\{X_t\}$. Then $f(X_t) = e_t + r_t$, where e_t is a martingale difference sequence, since

$$\mathbb{E}[e_{t+1}|\mathcal{F}_t] = \mathbb{E}[u(X_{t+1}) \mid \mathcal{F}_t] - \mathcal{P}u(X_t) = 0,$$

and

$$\frac{1}{\sqrt{t}}\sum_{i=1}^{t} r_i = \frac{1}{\sqrt{t}}(\mathcal{P}u(X_0) - \mathcal{P}u(X_t)) \to 0 \quad a.s., \tag{S2.3}$$

as $t \to \infty$, by a telescoping sum argument. Then from (4.6) we have

$$\begin{aligned}\sqrt{t}(\bar{\theta}_t - \theta_*) &= -\frac{1}{\sqrt{t}}\bar{A}^{-1}\sum_{i=1}^{t} f(X_{i+1}) + o_p(1) \\ &= -\frac{1}{\sqrt{t}}\bar{A}^{-1}\sum_{i=1}^{t} e_{i+1} + o_p(1),\end{aligned} \tag{S2.4}$$

by (S2.3). Combined with Proposition 4.2, this implies that

$$\frac{1}{\sqrt{t}}\bar{A}^{-1}\sum_{i=1}^{t} e_{i+1} \implies \mathcal{N}(0, \bar{A}^{-1}Q(\bar{A}^{-1})^{\mathsf{T}}). \tag{S2.5}$$

On the other hand, since e_{i+1} is uniformly bounded (as $f(x)$ is uniformly bounded for all $x \in \mathcal{X}$), the Lindenberg condition is satisfied, that is,

$$\sum_{i=1}^{t} \mathbb{E}\left[\frac{\|e_i\|_2^2}{t} I_{\{\|e_i\|_2/\sqrt{t} \geq \epsilon\}} \,|\mathcal{F}_{i-1}\right] \to 0,$$

in probability, as $t \to \infty$. So, by the martingale central limit theorem (e.g., Lemma A.3. of Liang (2010)), we have

$$\frac{1}{\sqrt{t}}\bar{A}^{-1}\sum_{i=1}^{t} e_{i+1} \implies \mathcal{N}(0, \Lambda), \tag{S2.6}$$

where Λ is a positive definite matrix with

$$\bar{A}^{-1}\sum_{i=1}^{t}\mathbb{E}[e_i e_i^{\mathsf{T}}/t|\mathcal{F}_{i-1}]\left(\bar{A}^{-1}\right)^{\mathsf{T}} \to \Lambda, \tag{S2.7}$$

in probability as $t \to \infty$. It follows from (S2.5) and (S2.6) that $\Lambda = \bar{A}^{-1}Q\left(\bar{A}^{-1}\right)^{\mathsf{T}}$, and so, by (S2.7), we have

$$\sum_{i=1}^{t}\mathbb{E}[e_i e_i^{\mathsf{T}}/t|\mathcal{F}_{i-1}] \to Q, \tag{S2.8}$$

in probability, as $t \to \infty$.

Next, from (4.7), we have

$$\begin{aligned}\sqrt{t}(\bar{\hat{\theta}}_t - \bar{\theta}_t) &= -\frac{1}{\sqrt{t}}\bar{A}^{-1}\sum_{i=1}^{t}(W_{i+1}-1)f(X_{i+1}) + o_p(1)\\ &= -\frac{1}{\sqrt{t}}\bar{A}^{-1}\sum_{i=1}^{t}(W_{i+1}-1)e_{i+1} + o_p(1),\end{aligned}$$

using (S2.3). Let $\xi_t = (W_t - 1)e_t$. Then ξ_t is a martingale difference sequence, since

$$\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = \mathbb{E}[W_{t+1}-1]\mathbb{E}[e_{t+1} \mid \mathcal{F}_t] = 0.$$

Since ξ_t is uniformly bounded, the Lindenberg condition holds. Then, by the martingale central limit theorem, conditional on the data $\mathcal{D}$, the term $\frac{1}{\sqrt{t}}\bar{A}^{-1}\sum_{i=1}^{t}\xi_{i+1}$ is asymptotically normal with mean 0 and variance

$$\begin{aligned}\underset{t\to\infty}{p\text{-lim}}\,\bar{A}^{-1}\sum_{i=1}^{t}\mathbb{E}[\xi_i\xi_i^{\mathsf{T}}/t|\mathcal{F}_{i-1}]\left(\bar{A}^{-1}\right)^{\mathsf{T}} &= \underset{t\to\infty}{p\text{-lim}}\,\bar{A}^{-1}\mathrm{Var}(W_1)\sum_{i=1}^{t}\mathbb{E}[e_i e_i^{\mathsf{T}}/t|\mathcal{F}_{i-1}]\left(\bar{A}^{-1}\right)^{\mathsf{T}}\\ &= \bar{A}^{-1}Q\left(\bar{A}^{-1}\right)^{-1},\end{aligned}$$

where the first equality follows by independence of W_i and e_i and the fact that the W_i's are i.i.d., while the second equality follows from (S2.8) and $\mathrm{Var}(W_1) = 1$. So, we have

$$\sqrt{t}(\bar{\hat{\theta}}_t - \bar{\theta}) = -\frac{1}{\sqrt{t}}\bar{A}^{-1}\sum_{i=1}^{t}\xi_{i+1} + o_p(1) \implies \mathcal{N}(0, \bar{A}^{-1}Q(\bar{A}^{-1})^{\mathsf{T}}), \tag{S2.9}$$

as $t \to \infty$, where the asymptotic normality holds conditional on data $\mathcal{D}$.

Let $X \sim \mathcal{N}(0, \bar{A}^{-1}Q(\bar{A}^{-1})^\intercal)$ denote the random variable with the limiting distribution of $\sqrt{t}(\bar{\theta}_t - \theta_*)$ as $t \to \infty$. Applying Lemma S2.3 to the result of Proposition 4.2 and equation (S2.9), respectively, we get

$$\sup_{v \in \mathbb{R}^d} \left| \mathbb{P}_{\mathcal{D}}(\sqrt{t}(\bar{\theta}_t - \theta_*) \leq v) - \mathbb{P}(X \leq v) \right| \to 0, \text{and}$$
$$\sup_{v \in \mathbb{R}^d} \left| \mathbb{P}_{\mathcal{W}|\mathcal{D}}(\sqrt{t}(\bar{\tilde{\theta}}_t - \theta_*) \leq v) - \mathbb{P}(X \leq v) \right| \to 0, \text{in probability},$$

as $t \to \infty$. Then

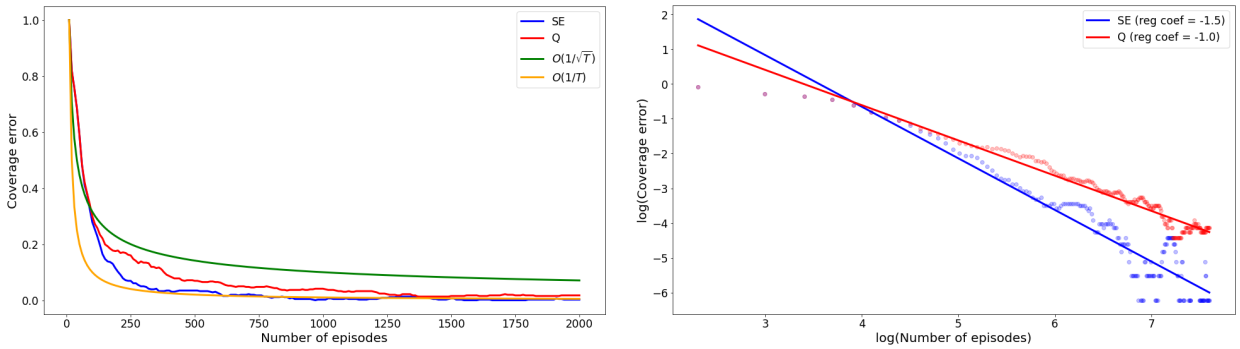
$$\begin{aligned}
&\sup_{v \in \mathbb{R}^d} \left| \mathbb{P}_{\mathcal{W}|\mathcal{D}}(\sqrt{t}(\bar{\tilde{\theta}}_t - \theta_*) \leq v) - \mathbb{P}_{\mathcal{D}}(\sqrt{t}(\bar{\theta}_t - \theta_*) \leq v) \right| \\
&\leq \sup_{v \in \mathbb{R}^d} \left| \mathbb{P}_{\mathcal{D}}(\sqrt{t}(\bar{\theta}_t - \theta_*) \leq v) - \mathbb{P}(X \leq v) \right| \\
&+ \sup_{v \in \mathbb{R}^d} \left| \mathbb{P}_{\mathcal{W}|\mathcal{D}}(\sqrt{t}(\bar{\tilde{\theta}}_t - \theta_*) \leq v) - \mathbb{P}(X \leq v) \right| \\
&\to 0,
\end{aligned}$$

in probability, as $t \to \infty$. □

S3 Additional Experiments

In this section, we provide the study of the second-order accuracy of our bootstrap method in the Frozenlake environment considered in Section 5.1.

To empirically evaluate the second-order accuracy, we measured the coverage error rates of the 95% confidence intervals for the value function of the initial state in the Frozenlake environment. We use TD learning to estimate the value function, and the quantile and standard error estimators, computed from the online bootstrap estimates, to generate the confidence intervals. Figure 8a shows the empirical coverage errors of the quantile and standard error estimators as a function of the number of episodes in the RL Frozenlake environment. The rates are re-scaled to start from 1 at the first time step. In both cases, we can see that the coverage error decreases at a rate faster than $O(1/\sqrt{t})$ initially, and eventually reaches a rate of $O(1/t)$ or better.



(a) Coverage error rates

(b) Least squares estimates

Figure 8: Figure 8a shows the coverage error rates for the quantile and SE confidence intervals, and Figure 8b shows the linear regression coefficients for the log coverage error rates against the log number of episodes.

We then computed estimates of the coverage error rate by regression the log of the coverage errors against the log of the number of episodes. We would expect a first-order accurate method to have a regression coefficient of $-1/2$ or lower (corresponding to a coverage error rate of $O(1/\sqrt{t})$), while a second-order accurate method would have a coefficient of -1 or lower. As shown in Figure 8b, both the quantile and standard error have regression coefficients of -1 or lower, which demonstrates that they both achieved second-order accuracy.