Supplement to "Reinforcement Learning in Possibly Nonstationary Environments"

Mengbing Li*[1], Chengchun Shi*[2], Zhenke Wu[3], and Piotr Fryzlewicz[4]

[1,3]University of Michigan, Ann Arbor
[2,4]London School of Economics and Political Science

This supplement is organised as follows. We begin with a list of commonly used notations in the supplement. We next introduce some technical definitions in Appendix A. In Appendix B, we present the proofs of Lemma 1, Theorems 1, 2 and 3. Finally, in Appendix C, we detail the simulation setting and present some additional simulation results.

Notations

S_t	The state vector at time t
A_t	The action taken at time t
R_t	The immediate reward at time t
F_t	The state transition function at time t, i.e., $S_{t+1} = F_t(S_t, A_t, \varepsilon_t)$
V^π	The state value function under π
η^π	The value under π, aggregated over different initial states
Q^{opt}	The optimal Q-function
$\widehat{Q}_{[T_1,T_2]}$	The estimated optimal Q-function using data collected from the interval $[T_1, T_2]$
$\widehat{\beta}_{[T_1,T_2]}$	The estimated regression coefficients using data collected from the interval $[T_1, T_2]$

*The first two authors contributed equally to this paper

$\beta^*_{[T_1,T_2]}$	The population limit of $\widehat{\beta}_{[T_1,T_2]}$
$\phi_L(a,s)$	The set of sieve basis functions of length L
u	A candidate change point location
γ	The discounted factor, between 0 and 1
ϵ	The boundary removal parameter
TS	The test statistic
TS^b	The bootstrap test statistic

A Some Technical Definitions

We first introduce the class of p-smoothness functions. For a J-tuple $\alpha = (\alpha_1, \dots, \alpha_J)^\top$ of nonnegative integers and a given function h on $\mathcal{S}$, let D^α denote the differential operator:

$$D^\alpha h(s) = \frac{\partial^{\|\alpha\|_1} h(s)}{\partial s_1^{\alpha_1} \cdots \partial s_J^{\alpha_J}}.$$

Here, s_j denotes the jth element of s. For any $p > 0$, let $\lfloor p \rfloor$ denote the largest integer that is smaller than p. The class of p-smooth functions is defined as follows:

$$\Lambda(p,c) = \left\{ h : \sup_{\|\alpha\|_1 \le \lfloor p \rfloor} \sup_{s \in \mathcal{S}} |D^\alpha h(s)| \le c, \sup_{\|\alpha\|_1 = \lfloor p \rfloor} \sup_{\substack{s_1, s_2 \in \mathcal{S} \\ s_1 \ne s_2}} \frac{|D^\alpha h(s_1) - D^\alpha h(s_2)|}{\|s_1 - s_2\|_2^{p - \lfloor p \rfloor}} \le c \right\},$$

for some constant $c > 0$.

B Proofs

Throughout the proof, we use c, $\bar{c}$, C, $\bar{C}$ to denote some generic constants whose values are allowed to vary from place to place. For any two positive sequences $\{a_{N,T}\}_{N,T}$, $\{b_{N,T}\}_{N,T}$, the notation $a_{N,T} \preceq b_{N,T}$ means that there exists some constant $C > 0$ such that $a_{N,T} \le C b_{N,T}$ for any N and T.

B.1 Proof of Lemma 1

We prove Lemma 1 in this section. Define

$$\eta^{\pi}_{(T+1):\infty} = \mathbb{E}\Big\{\mathbb{E}^{\pi}\Big(\sum_{t\geq 0}\gamma^t R_{t+T+1}|S_{T+1}\Big)\Big\},$$

for any policy π. Since the rewards are uniformly bounded, $\sum_{t>M}\gamma^t|R_t|$ is bounded by $O(\gamma^M)$ where the big-O term is uniform in π. As M diverges to infinity, $\eta^{\pi}_{(T+1):\infty}$ can be uniformly approximated by

$$\eta^{\pi}_{(T+1):(T+M)} = \mathbb{E}\Big\{\mathbb{E}^{\pi}\Big(\sum_{t=0}^{M-1}\gamma^t R_{t+T+1}|S_{T+1}\Big)\Big\}, \tag{B.1}$$

with arbitrary precision, for any π.

Similarly, $\mathbb{E}V_T^{\pi}(S_{t+1})$ can be uniformly approximated by

$$\eta^{\pi,T}_{(T+1):(T+M)} = \mathbb{E}\Big\{\mathbb{E}^{\pi,F_T,r_T}\Big(\sum_{t=0}^{M-1}\gamma^t R_{t+T+1}|S_{T+1}\Big)\Big\}, \tag{B.2}$$

with arbitrary precision, where the second expectation is taken by assuming that the transition and reward functions equal to F_T and r_T, respectively.

For any sufficiently small constant $\varepsilon > 0$, it follows from the assumption in formula (4) in the main paper that

$$\begin{aligned} \sup_{T\leq t\leq T+M}\sup_{a,s,\mathbb{S}}|\mathbb{P}(F_t(s,a,\varepsilon_1)\in\mathbb{S}) - \mathbb{P}(F_{t+1}(s,a,\varepsilon_1)\in\mathbb{S})| \leq \varepsilon, \\ \sup_{T\leq t\leq T+M}\sup_{a,s}|r_t(a,s) - r_{t+1}(a,s)| \leq \varepsilon, \end{aligned} \tag{B.3}$$

for sufficiently large NT. In the following, we aim to show that the difference between (B.1) and (B.2) is $O(\varepsilon)$, for any M. Since ε can be made arbitrarily small and M can be made arbitrarily large, the proof is hence completed.

First, it follows from (B.3) that

$$\sup_{\mathbb{S}}|\mathbb{P}(F_T(s,a,\varepsilon_1)\in\mathbb{S}) - \mathbb{P}(F_{T+j}(s,a,\varepsilon_1)\in\mathbb{S})| \leq j\varepsilon \text{ and } |r_T(a,s) - r_{T+j}(a,s)| \leq j\varepsilon, \tag{B.4}$$

for any $1 \leq j \leq M$ and any (a,s).

Second, for any $0 \leq j \leq M$, we define $\eta^{\pi,T,j}_{(T+1):(T+M)}$ as the cumulative discounted reward under π, assuming that the transition and reward functions equal F_T and r_T up to time point $T + j$, and equal $\{F_{T+k}\}_{k>j}$ and $\{r_{T+k}\}_{k>j}$ from $T+j+1$ to $T+M$. By definition, $\eta^{\pi,T,j}_{(T+1):(T+M)} = \eta^{\pi,T}_{(T+1):(T+M)}$ when $j = M$ and $\eta^{\pi,T,j}_{(T+1):(T+M)} = \eta^{\pi}_{(T+1):(T+M)}$ when $j = 0$. To bound the difference between (B.1) and (B.2), it suffices to bound

$$\sum_{j=0}^{M-1} |\eta^{\pi,T,j+1}_{(T+1):(T+M)} - \eta^{\pi,T,j}_{(T+1):(T+M)}|. \tag{B.5}$$

Third, for any $j < M$, the difference $|\eta^{\pi,T,j+1}_{(T+1):(T+M)} - \eta^{\pi,T,j}_{(T+1):(T+M)}|$ can be upper bounded by

$$\gamma^j |\mathbb{E}\{\mathbb{E}^{\pi,F_T,r_T}(\mathbb{E}^{\pi,r_T} R_{T+j+1}|S_{T+j+1})|S_{T+1}\} - \mathbb{E}\{\mathbb{E}^{\pi,F_T,r_T}(\mathbb{E}^{\pi} R_{T+j+1}|S_{T+j+1})|S_{T+1}\}|$$
$$+\gamma^{j+1} |\mathbb{E}\{\mathbb{E}^{\pi,F_T,r_T}(\mathbb{E}^{\pi,r_T} V^\pi(S_{T+j+2})|S_{T+j+1})|S_{T+1}\} - \mathbb{E}\{\mathbb{E}^{\pi,F_T,r_T}(\mathbb{E}^{\pi} V^\pi(S_{T+j+2})|S_{T+j+1})|S_{T+1}\}|.$$

Since the reward is uniformly bounded, so is the value function V^π. By (B.4), the first line is bounded by $\gamma^j \varepsilon(j+1)$, and the second line is bounded by $\gamma^{j+1}\varepsilon(j+1)C$ where C denotes the upper bound for the value function. As such, (B.5) is upper bounded by

$$(C+1)\sum_j \gamma^j \varepsilon(j+1) = O(\varepsilon).$$

The proof is hence completed.

B.2 Proof of Theorem 1

We begin by introducing the following auxiliary lemmas. Specifically, Lemma B.1 derives the uniform rate of convergence of the estimated Q-function. Lemma B.2 provides a uniform upper error bound on $|\widehat{W}_{[T_1,T_2]} - W_{[T_1,T_2]}|$. Without loss of generality, assume $T_0 = 0$. Their proofs are provided in Sections B.3 and B.4, respectively.

Lemma B.1. *Under the null hypothesis, there exists some constant $\varepsilon_0 > 0$ such that*

$$\sup_{a,s,T_2-T_1\geq \epsilon T} |\widehat{Q}_{[T_1,T_2]}(a,s) - Q^{opt}(a,s)| = O\{(NT)^{-\varepsilon_0}\},$$

with probability at least $1 - O(N^{-1}T^{-1})$, where ϵ corresponds to the boundary removal parameter defined in Section 4.1.

Lemma B.2. *Under the null hypothesis, there exists some constant $\bar{c} > 0$ such that $\|W_{[T_1,T_2]}^{-1}\|_2 \le \bar{c}$ and that $\sup_{T_2-T_1\ge \epsilon T} |\widehat{W}_{[T_1,T_2]} - W_{[T_1,T_2]}| = O\{(\epsilon NT)^{-1/2}\sqrt{L\log(NT)}\}$ with probability at least $1 - O(N^{-1}T^{-1})$. Here, $\|W_{[T_1,T_2]}^{-1}\|_2$ corresponds to the matrix operator norm of $W_{[T_1,T_2]}^{-1}$.*

B.2.1 ℓ_1Type Test

We begin with an outline of the proof of Theorem 1. The proof is divided into four steps.

In Step 1, we show there exist some constants $c, C > 0$ such that

$$\mathbb{P}(|\sqrt{NT}(\text{TS}_1 - \text{TS}_1^*)| \le C(NT)^{-c}) \to 1, \tag{B.6}$$

where

$$\text{TS}_1^* = \max_{\epsilon T < u < (1-\epsilon)T} \sqrt{\frac{u(T-u)}{T^2}} \left\{ \frac{1}{T} \sum_{t=0}^{T-1} \sum_{a} \int_{s} |\widehat{Q}_{[0,u]}(a,s) - \widehat{Q}_{[u,T]}(a,s)| \pi_t^b(a|s) p_t^b(s) ds \right\},$$

where p_t^b denotes the marginal distribution of S_t under the behavior policy. By definition, TS_1^* corresponds to a version of TS_1 by assuming the marginal distribution of the observed state-action pairs is known to us.

In the second step, we define

$$\widehat{Q}_{[T_1,T_2]}^{b,0}(a,s) = \frac{1}{N(T_2-T_1)} \phi_L^\top(a,s) W_{[T_1,T_2]}^{-1} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \delta_{i,t}(\widehat{\beta}_{[T_1,T_2]}) e_{i,t}, \quad \forall T_1, T_2,$$

a version of $\widehat{Q}_{[T_1,T_2]}^{b}(a,s)$ with $\widehat{W}_{[T_1,T_2]}^{-1}$ replaced by its oracle value, and establish a uniform upper error bound for $\max_{i,t,T_1,T_2} |\widehat{Q}_{[T_1,T_2]}^{b}(A_{i,t}, S_{i,t}) - \widehat{Q}_{[T_1,T_2]}^{b,0}(A_{i,t}, S_{i,t})|$. Specifically, we show that there exists some constant $c > 0$ such that the uniform upper error bound decays to zero at a rate of $O\{(NT)^{-1/2-c}\}$, with probability approaching 1 (WPA1). By triangle inequality, we can show that

$$\mathbb{P}(|\sqrt{NT}(\text{TS}_1^b - \text{TS}_1^{b,0})| \le C(NT)^{-c}) \to 1, \tag{B.7}$$

for some constant $C > 0$, where

$$\text{TS}_1^{b,0} = \max_{\epsilon T < u < (1-\epsilon)T} \sqrt{\frac{u(T-u)}{T^2}} \left\{ \frac{1}{NT} \sum_{t=0}^{T-1} \sum_{i=1}^{N} |\widehat{Q}_{[0,u]}^{b,0}(A_{i,t}, S_{i,t}) - \widehat{Q}_{[u,T]}^{b,0}(A_{i,t}, S_{i,t})| \right\}.$$

Using similar arguments as in the proof of (B.6) in Step 1 (see Pages 7 – 9), we can show that $|\text{TS}_1^{b,0} - \text{TS}_1^{b,*}|$ is upper bounded by $C(NT)^{-c}$ for some $c, C > 0$, WPA1, where

$$\text{TS}_1^{b,*} = \max_{\epsilon T < u < (1-\epsilon)T} \sqrt{\frac{u(T-u)}{T^2}} \left\{ \frac{1}{T} \sum_{t=0}^{T-1} \sum_{a} \int_{s} |\widehat{Q}_{[0,u]}^{b,0}(a,s) - \widehat{Q}_{[u,T]}^{b,0}(a,s)| \pi_t^b(a|s) p_t^b(s) ds \right\}.$$

This together with (B.7) yields that

$$\mathbb{P}(|\sqrt{NT}(\text{TS}_1^b - \text{TS}_1^{b,*})| \le C(NT)^{-c}) \to 1,$$

for some constants $c, C > 0$. It also implies that

$$\mathbb{P}(|\sqrt{NT}(\text{TS}_1^b - \text{TS}_1^{b,*})| \le C(NT)^{-c}|\text{Data}) \xrightarrow{P} 1. \tag{B.8}$$

In the third step, we define TS_1^{**} to be a version of TS_1^* with $\widehat{Q}_{[T_1,T_2]}$ replaced by the leading term in Assumption (A1). Similarly, we define $\text{TS}_1^{b,**}$ to be a version of $\text{TS}_1^{b,*}$ with $\delta_{i,t}(\widehat{\beta}_{[T_1,T_2]})$ replaced by the oracle value $\delta_{i,t}^*$. We will show that

$$\begin{aligned} &\mathbb{P}(|\sqrt{NT}(\text{TS}_1^* - \text{TS}_1^{**})| \le C(NT)^{-c}) \to 1, \\ &\mathbb{P}(|\sqrt{NT}(\text{TS}_1^{b,*} - \text{TS}_1^{b,**})| \le C(NT)^{-c}|\text{Data}) \xrightarrow{P} 1. \end{aligned} \tag{B.9}$$

Combining the results in (B.6)-(B.9), we have shown that

$$\begin{aligned} &\mathbb{P}(|\sqrt{NT}(\text{TS}_1 - \text{TS}_1^{**})| \le C(NT)^{-c}) \to 1, \\ &\mathbb{P}(|\sqrt{NT}(\text{TS}_1^b - \text{TS}_1^{b,**})| \le C(NT)^{-c}|\text{Data}) \xrightarrow{P} 1. \end{aligned} \tag{B.10}$$

In the last step, we aim to show the proposed test controls the type-I error. A key step in our proof is to bound the Kolmogorov distance between TS_1^{**} and $\text{TS}_1^{b,**}$. This together with (B.10) yields the validity of the proposed test. Notice that TS_1^{**} can be viewed as a function of the set of mean zero random vectors

$$\left\{ Z_u \equiv \frac{W_{[0,u]}^{-1}}{Nu} \sum_{i=1}^{N} \sum_{t=0}^{u-1} \phi_L(A_{i,t}, S_{i,t}) \delta_{i,t}^* - \frac{W_{[u,T]}^{-1}}{N(T-u)} \sum_{i=1}^{N} \sum_{t=u}^{T-1} \phi_L(A_{i,t}, S_{i,t}) \delta_{i,t}^* : u \right\}. \tag{B.11}$$

Similarly, $\text{TS}_1^{b,**}$ can be represented as a function of the bootstrapped samples

$$\left\{ Z_u^b \equiv \frac{W_{[0,u]}^{-1}}{Nu} \sum_{i=1}^{N} \sum_{t=0}^{u-1} \phi_L(A_{i,t}, S_{i,t}) \delta_{i,t}^* e_{i,t} - \frac{W_{[u,T]}^{-1}}{N(T-u)} \sum_{i=1}^{N} \sum_{t=u}^{T-1} \phi_L(A_{i,t}, S_{i,t}) \delta_{i,t}^* e_{i,t} : u \right\} \tag{B.12}$$

When T and L are fixed, the classical continuous mapping theorem can be applied to establish the weak convergence results. However, in our setting, L needs to diverge with the number of observations to alleviate the model misspecification error. We also allow T to approach infinity. Hence, classical weak convergence results cannot be applied. Toward that end, we establish a nonasymptotic error bound for the Kolmogorov distance as a function of N, T and L, and show that this bound decays to zero under the given conditions. The proof is based on the high-dimensional martingale central limit theorem developed by Belloni and Oliveira (2018).

We next detail the proof for each step.

Step 1. For each u, we aim to develop a concentration inequality to bound the difference

$$\left|\sqrt{\frac{u(T-u)}{T^2}}\left\{\frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\sum_{a}\int_{s}[|\widehat{Q}_{[0,u]}(A_{i,t},S_{i,t})-\widehat{Q}_{[u,T]}(A_{i,t},S_{i,t})| \right.\right. \tag{B.13}$$
$$\left.\left. - |\widehat{Q}_{[0,u]}(a,s)-\widehat{Q}_{[u,T]}(a,s)|]\pi_t^b(a|s)p_t^b(s)ds\right\}\right|.$$

In the proof of Lemma B.1, we have shown that $\sup_{T_1,T_2}\|\widehat{\beta}_{[T_1,T_2]}-\beta^*_{[T_1,T_2]}\|_2 = O(L^{-c})$ for some constant $c > 1/2$, with probability at least $1-O(N^{-1}T^{-1})$.

Define the set $\mathcal{B}_{[T_1,T_2]}(C) = \{\beta \in \mathbb{R}^L : \|\beta-\beta^*_{[T_1,T_2]}\|_2 \leq CL^{-c}\}$. It follows that there exists some sufficiently large constant $C > 0$ such that $\widehat{\beta}_{[T_1,T_2]} \in \mathcal{B}_{[T_1,T_2]}(C)$ with probability at least $1-O(N^{-1}T^{-1})$. (B.13) can thus be upper bounded by

$$\sup_{\substack{\beta_1\in\mathcal{B}_{[0,u]}(C)\\ \beta_2\in\mathcal{B}_{[u,T]}(C)}}\left|\frac{1}{2NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\{|\phi_L^\top(A_{i,t},S_{i,t})(\beta_1-\beta_2)|-\mathbb{E}|\phi_L^\top(A_{i,t},S_{i,t})(\beta_1-\beta_2)|\}\right|. \tag{B.14}$$

The upper bound for (B.14) can be established using similar arguments as in the proof of Lemma B.2. To save space, we only provide a sketch of the proof here. Please refer to the proof of Lemma B.2 for details.

Notice that the suprema in (B.14) are taken with respect to infinitely many βs. As such, standard concentration inequalities are not applicable to bound (B.14). Toward that end, we first take an ε-net of $\mathcal{B}_{[0,u]}(C)$ and $\mathcal{B}_{[u,T]}(C)$ for some sufficiently small $\varepsilon > 0$, denote by $\mathcal{B}^*_{[0,u]}(C)$ and $\mathcal{B}^*_{[u,T]}(C)$, respectively, such that for any $\beta \in \mathcal{B}_{[0,u]}(C)$ (or $\mathcal{B}_{[u,T]}(C)$), there exists some $\beta^* \in \mathcal{B}^*_{[0,u]}(C)$ (or $\mathcal{B}^*_{[u,T]}(C)$) that satisfies $\|\beta-\beta^*\|_2 \leq \varepsilon$. The purpose of introducing some an ε-net is to approximate these sets by collections of finitely many βs so that concentration inequalities are applicable to establish the upper bound. Set $\varepsilon = C(NT)^{-2}L^{-c}$. It follows from Lemma 2.2 of Mendelson et al. (2008) that there exist some $\mathcal{B}^*_{[0,u]}(C)$ and $\mathcal{B}^*_{[u,T]}(C)$ with number of elements upper bounded by $5^L(NT)^{2L}$.

Under (A4), we have $\sup_{a,s} \|\phi_L(a,s)\|_2 = O(\sqrt{L})$. Thanks to this uniform bound, the quantity within the absolute value symbol in (B.14) is a Lipschitz continuous function of (β_1, β_2), with the Lipschitz constant upper bounded by $O(\sqrt{L})$. As such, (B.14) can be approximated by

$$\sup_{\substack{\beta_1 \in \mathcal{B}^*_{[0,u]}(C) \\ \beta_2 \in \mathcal{B}^*_{[u,T]}(C)}} \left| \underbrace{\frac{1}{2NT} \sum_{t=0}^{T-1} \sum_{i=1}^{N} \sum_{a} \{ |\phi_L^\top(A_{i,t}, S_{i,t})(\beta_1 - \beta_2)| - \mathbb{E}|\phi_L^\top(A_{i,t}, S_{i,t})(\beta_1 - \beta_2)| \}}_{I(\beta_1,\beta_2) \ \ \text{(without absolute value)}} \right|, \tag{B.15}$$

with the approximation error given by $O(C\sqrt{L}N^{-2}T^{-2}L^{-c})$ where the big-O term is uniform in u.

It remains to develop a concentration inequality for (B.15). Since the number of elements in $\mathcal{B}^*_{[0,u]}(C)$ and $\mathcal{B}^*_{[u,T]}(C)$ are bounded, we could develop a tail inequality for the quantity within the absolute value symbol in (B.15) for each combination of β_1 and β_2, and then apply Bonferroni inequality to establish a uniform upper error bound. More specifically, for each pair (β_1, β_2), let

$$I^*(\beta_1, \beta_2) = \frac{1}{2N\sqrt{T}} \sum_{t=0}^{T-1} \sum_{i=1}^{N} [\mathbb{E}\{ |\phi_L^\top(A_{i,t}, S_{i,t})(\beta_1 - \beta_2)| | S_{i,t-1}\} - \mathbb{E}|\phi_L^\top(A_{i,t}, S_{i,t})(\beta_1 - \beta_2)|],$$

with the convention that $S_{i,-1} = \emptyset$. Notice that $I(\beta_1, \beta_2) - I^*(\beta_1, \beta_2)$ forms a mean-zero martingale under the Markov assumption, we can first apply the martingale concentration inequality (see e.g., Tropp, 2011) to show that

$$|I(\beta_1, \beta_2) - I^*(\beta_1, \beta_2)| = O(L^{-\varepsilon_0}\sqrt{N^{-1}T^{-1}\log(NT)}), \tag{B.16}$$

for some $\varepsilon_0 > 0$, with probability at least $1 - O\{(NT)^{-CL}\}$ for some sufficiently large constant $C > 0$. Here, the upper bound $O(L^{-\varepsilon_0}\sqrt{N^{-1}T^{-1}\log(NT)})$ decays faster than the parametric rate, due to the fact that the variance of the summand decays to zero under the condition that

$\sup_{a,s,u}|\phi_L^\top(a,s)(\beta^*_{[0,u]}-\beta^*_{[u,T]})| = O(N^{-c_6}T^{-c_6})$ for some $c_6 > 1/2$, imposed in the statement of Theorem 1. Specifically, notice that $\mathrm{Var}\{\phi_L^\top(A_t,S_t)(\beta_1-\beta_2)|S_{t-1}\}$ is upper bounded by

$$\mathbb{E}[\{\phi_L^\top(A_t,S_t)(\beta_1-\beta_2)\}^2|S_{t-1}] \le 3\mathbb{E}[\{\phi_L^\top(A_t,S_t)(\beta^*_{[0,u]}-\beta^*_{[u,T]})\}^2|S_{t-1}]$$
$$+3\mathbb{E}[\{\phi_L^\top(A_t,S_t)(\beta_1-\beta^*_{[0,u]})\}^2|S_{t-1}] + 3\mathbb{E}[\{\phi_L^\top(A_t,S_t)(\beta_2-\beta^*_{[u,T]})\}^2|S_{t-1}] \le C(NT)^{-2c_6}$$
$$+C\max_{a,a',s}\lambda_{\max}\left\{\int_{s'}\phi_L(a',s')\phi_L^\top(a',s')p(s'|a,s)ds'\right\}\max(\|\beta_1-\beta^*_{[0,u]}\|_2^2, \|\beta_2-\beta^*_{[u,T]}\|_2^2)$$
$$= O(NT)^{-2c_6} + O\{\max(\|\beta_1-\beta^*_{[0,u]}\|_2^2, \|\beta_2-\beta^*_{[u,T]}\|_2^2)\} = O(NT)^{-2c_6} + O(L^{-2c}),$$

for some constant $C > 0$, where the first inequality is due to Cauchy-Schwarz inequality, the second inequality is due to the condition on $\sup_{a,s,u}|\phi_L^\top(a,s)(\beta^*_{[0,u]}-\beta^*_{[u,T]})|$, and the first equality is due to (A4) and the fact that p is uniformly bounded (see (A3) and the definition of the p-smoothness function class in Section A).

Next, under (A5), the transition functions $\{F_t\}_t$ satisfies the conditions in the statement of Theorem 3.1 in Alquier et al. (2019). In addition, under (A4) and the condition that $\sup_{a,s,u}|\phi_L^\top(a,s)(\beta^*_{[0,u]}-\beta^*_{[u,T]})| = O\{(NT)^{-c_6}\}$ for some $c_6 > 1/2$, each summand in the definition of $I^*(\beta_1,\beta_2)$ is upper bounded by $O(L^{-c})$ for some $c > 1/2$. We can apply the concentration inequality for non-stationary Markov chains developed therein to show that $|I^*(\beta_1,\beta_2)| = O(L^{-\varepsilon_0}\sqrt{N^{-1}T^{-1}\log(NT)})$, with probability at least $1-O\{(NT)^{-CL}\}$ for some sufficiently large constant $C > 0$. Similarly, the upper bound $O(L^{-\varepsilon_0}\sqrt{N^{-1}T^{-1}\log(NT)})$ decays to zero at a rate faster than the parametric rate due to that each summand $\mathbb{E}\{|\phi_L^\top(A_{i,t},S_{i,t})(\beta_1-\beta_2)||S_{i,t-1}\} - \mathbb{E}|\phi_L^\top(A_{i,t},S_{i,t})(\beta_1-\beta_2)|$ is bounded by $O(L^{-\min(c,c_6)})$. This together with the upper bound for $|I(\beta_1,\beta_2)-I^*(\beta_1,\beta_2)|$ in (B.16), Bonferroni inequality and the condition that L is proportional to $(NT)^{c_5}$ yields the desired uniform upper bound for (B.15). This completes Step 1 of the proof.

Step 2. By definition $\widehat{Q}^{b,0}_{[T_1,T_2]}(A_{i,t},S_{i,t}) - \widehat{Q}^{b}_{[T_1,T_2]}(A_{i,t},S_{i,t})$ is equal to the sum of

$$\frac{1}{N(T_2-T_1)}\phi_L^\top(A_{i,t},S_{i,t})(\widehat{W}^{-1}_{[T_1,T_2]} - W^{-1}_{[T_1,T_2]})\sum_{i'=1}^{N}\sum_{t'=T_1}^{T_2-1}\phi_L(A_{i',t'},S_{i,t})\delta_{i',t'}(\widehat{\beta}_{[T_1,T_2]})e_{i',t'} \quad \text{(B.17)}$$

and

$$\frac{1}{N(T_2-T_1)}\phi_L^\top(A_{i,t},S_{i,t})W_{[T_1,T_2]}^{-1}\sum_{i'=1}^{N}\sum_{t'=T_1}^{T_2-1}\phi_L(A_{i',t'},S_{i',t'})(\delta_{i',t'}(\widehat{\beta}_{[T_1,T_2]})-\delta^*_{i',t'})e_{i',t'}. \quad \text{(B.18)}$$

Consider the first term. In Lemma B.2, we establish a uniform upper error bound for $\|\widehat{W}_{[T_1,T_2]}-W_{[T_1,T_2]}\|_2$ and show that $\|W_{[T_1,T_2]}^{-1}\|_2$ is upper bounded by some constant. Using similar arguments in Part 3 of the proof of Lemma 3 in Shi et al. (2021), we can show that $\|\widehat{W}_{[T_1,T_2]}^{-1}-W_{[T_1,T_2]}^{-1}\|_2$ is of the same order of magnitude as $\|\widehat{W}_{[T_1,T_2]}-W_{[T_1,T_2]}\|_2$. The boundedness assumption of R_t implies that the Q-function is bounded. This together with Lemma B.1 implies the estimated Q-function is bounded as well, and so is $\delta_{i',t'}(\widehat{\beta}_{[T_1,T_2]})$. By (A4), the conditional variance of (B.17) given the data is upper bounded by

$$\frac{CL^2}{\epsilon^2N^2T^2}\lambda_{\max}\left\{\frac{1}{N(T_2-T_1)}\sum_{i=1}^{N}\sum_{t=T_1}^{T_2-1}\phi_L(A_{i,t},S_{i,t})\phi_L^\top(A_{i,t},S_{i,t})\right\}.$$

Similar to Lemma B.2, we can show that the maximum eigenvalue of the matrix inside the curly brackets converges to $\lambda_{\max}\{(T_2-T_1)^{-1}\sum_{t=T_1}^{T_2-1}\mathbb{E}\phi_L(A_t,S_t)\phi_L^\top(A_t,S_t)\}$, which is bounded by some finite constant under (A4). Under the given conditions on ϵ and L, the conditional variance of (B.17) given the data is of the order $(NT)^{-2c-1}$, for some constant $c>0$, WPA1. Notice that the probability of a standard normal random variable exceeding z is bounded by $\exp(-z^2/2)$ for any $z>1$; see e.g., the inequality for the Gaussian Mill's ratio (Birnbaum, 1942). Since (B.17) is a mean-zero Gaussian random variable given the data, it is of the order $O\{(NT)^{-c-1/2}\sqrt{\log(NT)}\}$, with probability at least $1-O\{(NT)^{-C}\}$ for any sufficiently large constant $C>0$. This together with Bonferroni inequality yields the desired uniform upper error bound for the first term. As for the second term, notice that by Lemma B.1, the difference $\delta_{i,t}(\widehat{\beta}_{[T_1,T_2]})-\delta^*_{i,t}$ decays at a rate of $(NT)^{-c}$ for some constant $c>0$, uniformly in i,t,T_1,T_2, WPA1. Based on this result, we can similarly derive the upper error bound for the second term. This completes the proof.

Step 3. By triangle inequality and (A1), the difference $|\text{TS}_1^{b,*}-\text{TS}_1^{b,**}|$ can be upper bounded by

$$\max_{\epsilon T<u<(1-\epsilon)T} \frac{1}{2T}\sum_{t=0}^{T-1}\sum_a \int_s |\phi^\top(a,s)(b_{[0,u]}-b_{[u,T]}+O(N^{-c_1}T^{-c_1}))|\pi_t^b(a|s)p_t^b(s)ds$$
$$\leq \max_{\epsilon T<u<(1-\epsilon)T} \frac{1}{2T}\sum_{t=0}^{T-1}\sum_a \int_s |\phi^\top(a,s)(b_{[u,T]}-b_{[0,u]})|\pi_t^b(a|s)p_t^b(s)ds$$
$$+O(N^{-c_1}T^{-c_1}) \max_{\epsilon T<u<(1-\epsilon)T} \sup_{\nu\in\mathbb{R}^L:\|\nu\|_2=1} \frac{1}{2T}\sum_{t=0}^{T-1}\sum_a \int_s |\phi^\top(a,s)\nu|\pi_t^b(a|s)p_t^b(s)ds,$$

WPA1. The first term on the RHS is $O\{(NT)^{-c_6}\}$ under the given conditions in the statement of Theorem 1. Under (A2), $\{p_t\}_t$ is uniformly bounded. So is $\{p_t^b\}_t$. It follows from (A4) that the second term is $O(N^{-c_1}T^{-c_1})$. This yields the first assertion in (B.9). The second assertion can be similarly proven based on the result $\sup_{T_1,T_2}\|\widehat{\beta}_{[T_1,T_2]}-\beta^*_{[T_1,T_2]}\|_2 = O(L^{-c})$ for some constant $c>1/2$, WPA1, as shown in Lemma B.1.

Step 4. As we have commented, the proof is based on the high-dimensional martingale central limit theorem developed by Belloni and Oliveira (2018). Let Z and Z^b denote the high-dimensional random vectors formed by stacking the random vectors in the set (B.11) and (B.12), respectively. It can be represented as $\sum_{i,t} Z_{i,t}$ where each $Z_{i,t}$ depends on the data tuple $(S_{i,t},A_{i,t},R_{i,t},S_{i,t+1})$. We first observe that it corresponds to a sum of high-dimensional martingale difference. Specifically, for any integer $1\leq g\leq NT$, let $i(g)$ and $t(g)$ be the quotient and the remainder of $g+T-1$ divided by T that satisfy

$$g=\{i(g)-1\}T+t(g)+1 \;\text{ and }\; 0\leq t(g)<T.$$

Let $\mathcal{F}^{(0)}=\{S_{1,0},A_{1,0}\}$. Then we recursively define $\{\mathcal{F}^{(g)}\}_{1\leq g\leq NT}$ as follows:

$$\mathcal{F}^{(g)}=\left\{\begin{array}{ll} \mathcal{F}^{(g-1)}\cup\{R_{i(g),t(g)},S_{i(g),t(g)+1},A_{i(g),t(g)+1}\}, & \text{if } t(g)<T-1; \\ \mathcal{F}^{(g-1)}\cup\{R_{i(g),T-1},S_{i(g),T},S_{i(g)+1,0},A_{i(g)+1,0}\}, & \text{otherwise.}\end{array}\right.$$

This allows us to rewrite Z as $\sum_{g=1}^{NT}Z^{(g)}=\sum_{g=1}^{NT}Z_{i(g),t(g)}$. Similarly, we can rewrite Z^b as $\sum_{g=1}^{NT}Z^{b,(g)}$. Under the Markov assumption (MA) and conditional mean independence assumption (CMIA), it forms a sum of martingale difference sequence with respect to the filtration $\{\sigma(\mathcal{F}^{(g)})\}_{g\geq 0}$ where $\sigma(\mathcal{F})$ denotes the σ-algebra generated by $\mathcal{F}$. Similarly, we can represent $Z_u = \sum_{g=1}^{NT} Z_u^{(g)}$. The test statistic can be represented as

$$\text{TS}_1^{**} = \max_{\epsilon T < u < (1-\epsilon)T} \underbrace{\sqrt{\frac{u(T-u)}{T^2}} \frac{1}{T} \sum_{t=0}^{T-1} \sum_a \int_s |\phi_L^\top(a,s) Z_u| p_t^b(s) \pi_t^b(a|s) ds}_{\psi_u} .$$

Due to the existence of the max operator and the absolute value function, TS_1^{**} is a non-smooth function of Z. We next approximate the maximum and absolute value functions using a smooth surrogate. Let θ be a sufficiently large real number. Consider the following smooth approximation of the maximum function, $F_\theta(\{\psi_u\}_u)$, defined as

$$\frac{1}{\theta} \log(\sum_u \exp(\theta \psi_u)).$$

It can be shown that

$$\max_{\epsilon T < u < (1-\epsilon)T} \psi_u \le F_\theta(\{\psi_u\}_{\epsilon T < u < (1-\epsilon)T}) \le \max_{\epsilon T < u < (1-\epsilon)T} \psi_u + \frac{\log T}{\theta}. \tag{B.19}$$

See e.g., Equation (37) of Chernozhukov et al. (2014).

Similarly, we can approximate the absolute value function $|x| = \max(x,0) + \max(-x,0)$ by $\theta^{-1}\{\log(1+\exp(\theta x)) + \log(1+\exp(-\theta x))\}$. Define the corresponding smooth function $f_\theta(Z_u)$ as

$$\sqrt{\frac{u(T-u)}{T^2}} \frac{1}{T\theta} \sum_{t=0}^{T-1} \sum_a \int_s \{\log(1+\exp(\theta \phi_L^\top(a,s) Z_u)) + \log(1+\exp(-\theta \phi_L^\top(a,s) Z_u))\} p_t^b(s) \pi_t^b(a|s) ds.$$

Similarly, we have $\psi_u \le f_\theta(Z_u) \le \psi_u + \theta^{-1} \log 2$. This together with (B.19) yields

$$\text{TS}_1^{**} \le F_\theta(\{f_\theta(Z_u)\}_u) \le F_\theta(\{\psi_u + \theta^{-1}\log 2\}_u) \le \text{TS}_1^{**} + \frac{\log(2T)}{\theta}. \tag{B.20}$$

For a given z, consider the probability $\mathbb{P}(\sqrt{NT}\text{TS}_1^{**} \le z)$. According to Section B.1 of Belloni and Oliveira (2018), for any $\delta > 0$, there exists a thrice differentiable function h that satisfies $|h'| \le \delta^{-1}$, $|h''| \le \delta^{-2} K$ and $|h'''| \le \delta^{-3} K$ for some universal constant $K > 0$ such that

$$\mathbb{I}(x \le z/\sqrt{NT} + \delta) \le h(x) \le \mathbb{I}(x \le z/\sqrt{NT} + 4\delta). \tag{B.21}$$

Define a composite function $m(\{Z_u\}_u) = h \circ F_\theta(\{f_\theta(Z_u)\}_u)$. Combining (B.21) with (B.20) yields that

$$\mathbb{P}(\sqrt{NT}\text{TS}_1^{**} \le z) \le \mathbb{E}m(\{Z_u\}_u) \le \mathbb{P}(\sqrt{NT}\text{TS}_1^{**} \le z + 4\sqrt{NT}\delta). \tag{B.22}$$

Similarly, we have

$$\mathbb{P}(\sqrt{NT}\text{TS}_1^{b,**} \le z|\text{Data}) \le \mathbb{E}[m(\{Z_u^b\}_u)|\text{Data}] \le \mathbb{P}(\sqrt{NT}\text{TS}_1^{b,**} \le z + 4\sqrt{NT}\delta|\text{Data}),$$

where Z_u^b is defined in (B.12). This together with (B.22) yields that

$$\begin{aligned}\max\{\max_z |\mathbb{P}(\sqrt{NT}\text{TS}_1^{b,**} \le z|\text{Data}) - \mathbb{P}(\sqrt{NT}\text{TS}_1^{**} \le z - 4\sqrt{NT}\delta)|,\\ \max_z |\mathbb{P}(\sqrt{NT}\text{TS}_1^{b,**} \le z|\text{Data}) - \mathbb{P}(\sqrt{NT}\text{TS}_1^{**} \le z + 4\sqrt{NT}\delta)|\}\\ \le |\mathbb{E}m(\{Z_u\}) - \mathbb{E}[m(\{Z_u^b\}_u)|\text{Data}]|.\end{aligned} \tag{B.23}$$

We next apply Corollary 2.1 of Belloni and Oliveira (2018) to establish an upper bound for $|\mathbb{E}m(\{Z_u\}_u) - \mathbb{E}[m(\{Z_u^b\}_u)|\text{Data}]|$. Similar to Lemma 4.3 of Chernozhukov et al. (2014), we can show that $c_0 \equiv \sup_{z,z'} |m(z) - m(z')| \le 1$,

$$\begin{aligned} c_2 &\equiv \sup_z \sum_{j_1,j_2} \left|\frac{\partial^2 m(z)}{\partial z_{j_1}\partial z_{j_2}}\right| \preceq \delta^{-2}L + \delta^{-1}\theta L,\\ c_3 &\equiv \sup_z \sum_{j_1,j_2,j_3} \left|\frac{\partial^3 m(z)}{\partial z_{j_1}\partial z_{j_2}\partial z_{j_3}}\right| \preceq \delta^{-3}L^{3/2} + \delta^{-2}\theta L^{3/2} + \delta^{-1}\theta^2 L^{3/2},\end{aligned}$$

under (A4). In addition, similar to Lemma B.2, we can show that the quadratic variation process $\sum_g \mathbb{E}\{Z^{(g)}(Z^{(g)})^\top|\mathcal{F}^{(g-1)}\}$ will converge to some deterministic matrix with elementwise maximum norm bounded by $C\sqrt{\log(NT)}/(\epsilon NT)^{3/2}$ for some constant $C > 0$, with probability at least $1 - O(N^{-2}T^{-2})$. So is the conditional covariance matrix of Z^b given the data, e.g.,

$$\sum_g \mathbb{E}[Z^{b,(g)}(b, Z^{(g)})^\top|\{S_{i,t}, A_{i,t}, R_{i,t}\}_{1\le i\le N, 0\le t\le T}].$$

Moreover, under (A4), the third absolute moment of each element in $Z^{(g)}$ is bounded by $O(N^{-3}T^{-3}\sqrt{L})$. Let $\delta = \theta^{-1}$. It follows from Corollary 2.1 of Belloni and Oliveira (2018) that

$$|\mathbb{E}m(\{Z_u\}) - \mathbb{E}[m(\{Z_u^b\}_u)|\text{Data}]| \preceq \frac{\theta^2 L\sqrt{\log(NT)}}{(\epsilon NT)^{3/2}} + \frac{\theta^3 L^2}{\epsilon^3 (NT)^2}, \tag{B.24}$$

WPA1. Notice that $\sqrt{NT}\text{TS}_1^{**}$ has a bounded probability density function. It follows that

$$\sup_z |\mathbb{P}(\sqrt{NT}\text{TS}_1^{**} \le z - 4\sqrt{NT}\delta) - \mathbb{P}(\sqrt{NT}\text{TS}_1^{**} \le z + 4\sqrt{NT}\delta)| \preceq \sqrt{NT}\delta. \quad \text{(B.25)}$$

By setting $\theta = (NT)^c$ for some constant $1/2 < c < 2/3 - 2c_5/3$, both the RHS of (B.24) and (B.25) decay to zero. In view of (B.23), we have shown that

$$\sup_z |\mathbb{P}(\sqrt{NT}\text{TS}_1^{b,**} \le z|\text{Data}) - \mathbb{P}(\sqrt{NT}\text{TS}_1^{**} \le z)| \xrightarrow{p} 0.$$

Similarly, based on (B.10), we can show that

$$\sup_z |\mathbb{P}(\sqrt{NT}\text{TS}_1^{b} \le z|\text{Data}) - \mathbb{P}(\sqrt{NT}\text{TS}_1 \le z)| \xrightarrow{p} 0.$$

The proof is hence completed.

B.2.2 Maximum-Type Tests

For any u, a and s, define the variance estimator $\widehat{\sigma}_u^2(a,s)$ by

$$\frac{\phi_L^\top(a,s)\widehat{W}^{-1}_{[T-\kappa,T-u]}}{N^2(\kappa-u)^2}\left[\sum_{i=1}^{N}\sum_{t=T-\kappa}^{T-u-1}\phi_L(A_{i,t},S_{i,t})\phi_L^\top(A_{i,t},S_{i,t})\delta^2_{i,t}(\widehat{\beta}_{[T-\kappa,T-u]})\right]\{\widehat{W}^{-1}_{[T-\kappa,T-u]}\}^\top\phi_L^\top(a,s)$$
$$+\frac{1}{N^2u^2}\phi_L^\top(a,s)\widehat{W}^{-1}_{[T-u,T]}\left[\sum_{i=1}^{N}\sum_{t=T-u}^{T-1}\phi_L(A_{i,t},S_{i,t})\phi_L^\top(A_{i,t},S_{i,t})\delta^2_{i,t}(\widehat{\beta}_{[T-u,T]})\right]\{\widehat{W}^{-1}_{[T-u,T]}\}^\top\phi_L^\top(a,s).$$

We next show that both the unnormalised and normalised maximum-type tests have good size property. The proof is very similar to that in Section B.2.1. We provide a sketch of the proof and outline some major key steps only.

Proof for the unnormalised test: The first step in the proof is to show that $\sqrt{NT}(\text{TS}_\infty - \text{TS}_\infty^*) = o_p(1)$, where TS_∞^* is a version of TS_∞ with $\widehat{Q}_{[T_1,T_2]}$ replaced by the leading term according to (A1). By Assumptions (A1), (A4) and the condition that $(NT)^{2c_1-1} \gg L$, this can be proven using similar arguments to Step 3 of the proof in Section B.2.1.

The second step in the proof is to show that $\sqrt{NT}(\text{TS}_\infty^b - \text{TS}_\infty^{b,*}) = o_p(1)$ where $\text{TS}_\infty^{b,*}$ is a version of TS_∞^b with $\widehat{W}_{[T_1,T_2]}$ and $\delta_{i,t}(\widehat{\beta}_{[T_1,T_2]})$ replaced by their oracle values $W_{[T_1,T_2]}$ and $\delta_{i,t}^*$, respectively, for any T_1 and T_2. This can be proven using similar arguments to Steps 2 and 3 of the proof in Section B.2.1.

Notice that in the test statistic, the maximum is taken over all state-action pairs. Recall that the state space is $[0,1]^d$. Consider an ε-net of $[0,1]^d$ with $\varepsilon = \sqrt{d}/(NT)^4$. Let TS_∞^{**} and $\text{TS}_\infty^{b,**}$ be versions of TS_∞^{*} and $\text{TS}_\infty^{b,*}$ where the maximum is taken over the ε-net. Under (A10), using similar arguments to Step 1 of the proof in Section B.2.1, we can show that $\sqrt{NT}(\text{TS}_\infty^{*} - \text{TS}_\infty^{**}) = o_p(1)$ and $\sqrt{NT}(\text{TS}_\infty^{b,*} - \text{TS}_\infty^{b,**}) = o_p(1)$. This corresponds to the second step of the proof.

Finally, the last step in the proof is to show

$$\sup_z |\mathbb{P}(\sqrt{NT}\text{TS}_\infty^{b,**} \le z|\text{Data}) - \mathbb{P}(\sqrt{NT}\text{TS}_\infty^{**} \le z)| \xrightarrow{p} 0.$$

Similar to Step 4 of the proof in Section B.2.1, this step can be proven based on the high-dimensional martingale central limit theorem developed by Belloni and Oliveira (2018). Together with the first two steps, we obtain that

$$\sup_z |\mathbb{P}(\sqrt{NT}\text{TS}_\infty^{b} \le z|\text{Data}) - \mathbb{P}(\sqrt{NT}\text{TS}_\infty \le z)| \xrightarrow{p} 0.$$

The proof is hence completed.

Proof for the normalised test: Using similar arguments to Step 2 of the proof in Section B.2.1, we can show that WPA1, (i) $\max_{T_2 - T_1 \ge \epsilon T} \|\widehat{W}_{[T_1,T_2]}^{-1}\|_2 = O(1)$;

$$\text{(ii)} \min_{T_2 - T_1 \ge \epsilon T} \lambda_{\min} \left\{ \frac{1}{N(T_2 - T_2)} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2 - 1} \phi_L(A_{i,t}, S_{i,t}) \phi_L^\top(A_{i,t}, S_{i,t}) \right\} > C,$$

for some constant $C > 0$. In addition, by Lemma B.1, the difference $\delta_{i,t}(\widehat{\beta}_{[T_1,T_2]}) - \delta_{i,t}^*$ decays at a rate of $(NT)^{-c}$ for some constant $c > 0$, uniformly in i, t, T_1, T_2, WPA1. This together with (i) and (ii) implies that there exists some constant $C > 0$ such that $\widehat{\sigma}_u^2(a,s) > C\|\phi_L(a,s)\|_2^2$ for any u, a, s, WPA1. This allows us to show $\sqrt{NT}(\text{TS}_{n,\infty} - \text{TS}_{n,\infty}^*) = o_p(1)$ under a weaker condition that does not require $(NT)^{2c_1 - 1} \gg L$, where $\text{TS}_{n,\infty}^*$ is a version of $\text{TS}_{n,\infty}$ with $\widehat{Q}_{[T_1,T_2]}$ replaced by the leading term according to (A1). The rest of the proof can be established in a similar manner as the proof for the unnormalised test.

B.3 Proof of Lemma B.1

We aim to establish the uniform rate of convergence of $\{\sup_{a,s}|\phi_L(a,s)^\top\widehat{\beta}_{[T_1,T_2]} - Q^{opt}(a,s)| : T_2 - T_1 \geq \epsilon T\}$. Under (A2), the optimal Q-function Q^{opt} is p-smooth (see e.g., the proof of Lemma 1 in Shi et al., 2021). This together with (A3) implies that there exists some β^* such that the bias $\sup_{a,s}|Q^{opt}(a,s) - \phi_L^\top(a,s)\beta^*| = O(L^{-c_2})$, under the null hypothesis. By the definition of $\beta^*_{[T_1,T_2]}$, we have

$$\beta^*_{[T_1,T_2]} - \beta^* = [\mathbb{E}\phi_L(A_t,S_t)\{\phi_L(A_t,S_t) - \gamma\phi_L(\pi^{opt}(S_{t+1}),S_{t+1})\}^\top]^{-1}$$
$$\times[\mathbb{E}\phi_L(A_t,S_t)\{R_t + \gamma\phi_L^\top(\pi^{opt}(S_{t+1}),S_{t+1})\beta^* - \phi_L(A_t,S_t)\beta^*\}].$$

Since $\|W^{-1}_{[T_1,T_2]}\|_2 \leq \bar{c}$, $\|\beta^*_{[T_1,T_2]} - \beta^*\|_2$ can be upper bounded by

$$\bar{c}\|\mathbb{E}\phi_L(A_t,S_t)\{R_t + \gamma\phi_L^\top(\pi^{opt}(S_{t+1}),S_{t+1})\beta^* - \phi_L(A_t,S_t)\beta^*\}\|_2 \leq \bar{c}\sup_{\nu\in\mathbb{R}^L}\mathbb{E}|\nu^\top\phi_L(A_t,S_t)|$$
$$\times|\gamma\phi_L^\top(\pi^{opt}(S_{t+1}),S_{t+1})\beta^* - \gamma Q^{opt}(\pi^{opt}(S_{t+1}),S_{t+1}) - \phi_L(A_t,S_t)\beta^* + Q^{opt}(A_t,S_t)|.$$

The second term on the RHS is of the order $O(L^{-c_2})$. The first term on the RHS can be upper bounded by $\bar{c}\sup_{a\in\mathbb{R}^L}\sqrt{\mathbb{E}|a^\top\phi_L(A_t,S_t)|^2} \leq \bar{c}\sqrt{\lambda_{\max}\mathbb{E}\phi_L(A_t,S_t)\phi_L^\top(A_t,S_t)} = O(1)$, under Condition (A4). As such, we have $\sup_{T_1,T_2}\|\beta^*_{[T_1,T_2]} - \beta^*\|_2 = O(L^{-c_2})$.

Under the conditions that $c_2 > 1/2$ and $\sup_{a,s}\|\phi_L(a,s)\|_2 = O(L^{1/2})$, we have $\sup_{a,s,T_1,T_2}\|\phi(a,s)^\top(\beta^* - \beta^*_{[T_1,T_2]})\|_2 = O(L^{-\epsilon_0})$ for some $\epsilon_0 > 0$. It follows that $\sup_{a,s,T_1,T_2}|Q^{opt}(a,s) - \phi_L^\top(a,s)\beta^*_{[T_1,T_2]}| = O(L^{-\epsilon_0})$. Since L is proportional to $(NT)^{c_5}$ for some $c_5 > 0$, we obtain that $\sup_{a,s,T_1,T_2}|Q^{opt}(a,s) - \phi_L^\top(a,s)\beta^*_{[T_1,T_2]}| = O\{(NT)^{-\varepsilon_0}\}$ for some $\varepsilon_0 > 0$.

As such, it suffices to show $\sup_{a,s}\|\phi_L(a,s)^\top(\widehat{\beta}_{[T_1,T_2]} - \beta^*_{[T_1,T_2]})\|_2 = O\{(NT)^{-\varepsilon_0}\}$, or equivalently, $\sup_{T_1,T_2}\|\widehat{\beta}_{[T_1,T_2]} - \beta^*_{[T_1,T_2]}\|_2 = O(L^{-c})$, with probability at least $1 - O(N^{-1}T^{-1})$, for some $c > 1/2$. Under (A1), it suffices to show both the bias term $\sup_{T_1,T_2}\|b_{[T_1,T_2]}\|_2$ and the standard deviation term $\|N^{-1}(T_2 - T_1)^{-1}W^{-1}_{[T_1,T_2]}\sum_{i=1}^{N}\sum_{t=T_1}^{T_2-1}\phi_L(A_{i,t},S_{i,t})\delta^*_{i,t}\|_2$ are $O(L^{-c})$.

Under the null hypothesis, the bias term satisfies

$$\|b_{[T_1,T_2]}\|_2 = \sup_{T_1,T_2}\left\|\frac{1}{T_2-T_1}W_{[T_1,T_2]}^{-1}\sum_{t=T_1}^{T_2-1}\mathbb{E}\phi_L(A_t,S_t)\phi_L^\top(A_t,S_t)(\beta^*_{[T_1,T_2]}-\beta^*)\right\|_2$$
$$+\sup_{T_1,T_2}\left\|\frac{1}{T_2-T_1}W_{[T_1,T_2]}^{-1}\sum_{t=T_1}^{T_2-1}\mathbb{E}\phi_L(A_t,S_t)\{Q^{opt}(A_t,S_t)-\phi_L^\top(A_t,S_t)\beta^*\}\right\|_2.$$

By Cauchy-Schwarz inequality, the first term on the RHS can be upper bounded by

$$\sup_{t,T_1,T_2}\|W_{[T_1,T_2]}^{-1}\|_2\|\mathbb{E}\phi_L(A_t,S_t)\phi_L(A_t,S_t)^\top\|_2\|\beta^*_{[T_1,T_2]}-\beta^*\|_2 = O(L^{-c_2}).$$

The second term can be shown to be $O(L^{-c_2})$, using similar arguments in bounding $\|\beta^*_{[T_1,T_2]}-\beta^*\|_2$.

The assertion that

$$\sup_{T_1,T_2}\left\|\frac{1}{N(T_2-T_1)}W_{[T_1,T_2]}^{-1}\sum_{i=1}^{N}\sum_{t=T_1}^{T_2-1}\phi_L(A_{i,t},S_{i,t})\delta^*_{i,t}\right\|_2 = O(\sqrt{L(\epsilon NT)^{-1}\log(NT)}) = O(L^{-c}),$$

with probability at least $1-O(N^{-1}T^{-1})$ can be proven using the martingale concentration inequality (see e.g., Corollary 3.1, Tropp, 2011) and the Bonferroni inequality, under the condition that $L = O\{(NT)^{c_5}\}$ for some $c_5 < 1/4$. We omit the details to save space.

B.4 Proof of Lemma B.2

We focus on establishing a uniform upper error bound for $\{|\widehat{W}_{[T_1,T_2]}-W_{[T_1,T_2]}| : T_2-T_1 \geq \epsilon T\}$ in this section. The assertion that $\|W_{[T_1,T_2]}^{-1}\|_2 \leq \bar{c}$ can be proven by Lemma 3 of Shi et al. (2021).

In Lemma B.1, we have established the uniform consistency of the estimated Q-function. Under the margin condition in (A8), it follows that

$$|\phi_L^\top(a,s)\widehat{\beta}_{[T_1,T_2]} - Q^{opt}(a,s) - \phi_L^\top(\pi^{opt}(s),s)\widehat{\beta}_{[T_1,T_2]} + Q^{opt}(\pi^{opt}(s),s)|$$
$$< |Q^{opt}(a,s) - Q^{opt}(\pi^{opt}(s),s)|,$$

for any a and s, with probability at least $1-O(N^{-1}T^{-1})$. As such, we have $\arg\max_a \phi_L^\top(a,s)\widehat{\beta}_{[T_1,T_2]} = \arg\max_a Q^{opt}(a,s)$ and hence $\pi_{\widehat{\beta}_{[T_1,T_2]}} = \pi^{opt}$, with probability at least $1-O(N^{-1}T^{-1})$. It follows that

$$\widehat{W}_{[T_1,T_2]} = \frac{1}{N(T_2-T_1)} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t})\{\phi_L(A_{i,t}, S_{i,t}) - \gamma\phi_L(\pi^{opt}(S_{i,t+1}), S_{i,t+1})\}^\top \quad (B.26)$$

We next provide an upper bound on the difference between the RHS of (B.26) and $W_{[T_1,T_2]}$. Define $\widehat{W}^*_{[T_1,T_2]}$ as

$$\frac{1}{N(T_2-T_1)} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \sum_a \pi^b(a|S_{i,t})\phi_L(a, S_{i,t})[\phi_L(a, S_{i,t}) - \gamma\mathbb{E}\{\phi_L(\pi^{opt}(S_{i,t+1}), S_{i,t+1})|A_{i,t} = a, S_{i,t}\}]^\top.$$

The difference $|\widehat{W}_{[T_1,T_2]} - W_{[T_1,T_2]}|$ can be upper bounded by $|\widehat{W}_{[T_1,T_2]} - \widehat{W}^*_{[T_1,T_2]}| + |\widehat{W}^*_{[T_1,T_2]} - W_{[T_1,T_2]}|$.

Under CMIA, the first term $\widehat{W}_{[T_1,T_2]} - \widehat{W}^*_{[T_1,T_2]}$ corresponds to a sum of martingale difference. Using similar arguments to the proof of Lemma 3 of Shi et al. (2021), we can show that the first term is of the order $O(\sqrt{(\epsilon NT)^{-1}L\log(NT)})$, with probability at least $1 - O\{(NT^{-3})\}$, under the condition that $T_2 - T_1 \geq \epsilon T$. See also, Freedman's inequality for matrix martingales developed by Tropp (2011). It follows from Bonferroni's inequality that $\sup_{T_1,T_2} \|\widehat{W}_{[T_1,T_2]} - \widehat{W}^*_{[T_1,T_2]}\|_2 = O(\sqrt{(\epsilon NT)^{-1}L\log(NT)})$, with probability at least $1 - O\{(NT)^{-1}\}$.

It remains to bound $\|\widehat{W}^*_{[T_1,T_2]} - W_{[T_1,T_2]}\|_2$. Let Γ_0 be an ε-net of the unit sphere in $\mathbb{R}^L$ that satisfies the following: for any $\nu \in \mathbb{R}^L$ with unit ℓ_2 norm, there exists some $\nu_0 \in \Gamma_0$ such that $\|\nu - \nu_0\|_2 \leq \varepsilon$. Set $\varepsilon = (NT)^{-2}$. According to Lemma 2.3 of Mendelson et al. (2008), there exists such an ε-net Γ_0 that belongs to the unit sphere and satisfies $|\Gamma_0| \leq 5^L(NT)^{2L}$.

For any ν_1, ν_2 with unit ℓ_2 norm, define

$$\Psi_t(a, s, \nu_1, \nu_2) = \pi_t^b(a|s)\nu_1^\top\phi_L(a, s)[\phi_L(a, s) - \gamma\mathbb{E}\{\phi_L(\pi^{opt}(S_{t+1}), S_{t+1})|A_t = a, S_t = s\}]^\top\nu_2.$$

The difference $\|\widehat{W}^*_{[T_1,T_2]} - W_{[T_1,T_2]}\|_2$ can be represented as

$$\sup_{\|\nu_1\|_2=\|\nu_2\|_2=1} \left| \frac{1}{N(T_2-T_1)} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \sum_a \{\Psi_t(a, S_{i,t}, \nu_1, \nu_2) - \mathbb{E}\Psi_t(a, S_{i,t}, \nu_1, \nu_2)\} \right|.$$

We first show that $\Psi_t(a, s, \nu_1, \nu_2)$ is a Lipschitz continuous function of ν_1 and ν_2. For any $\nu_1, \nu_2, \nu_3, \nu_4$, the difference $\Psi_t(a, s, \nu_1, \nu_2) - \Psi_t(a, s, \nu_3, \nu_4)$ can be decomposed into the sum of the following two terms:

$$\pi_t^b(a|s)(\nu_1-\nu_3)^\top\phi_L(a,s)[\phi_L(a,s)-\gamma\mathbb{E}\{\phi_L(\pi^{opt}(S_{t+1}),S_{t+1})|A_t=a,S_t=s\}]^\top\nu_2$$
$$+\pi_t^b(a|s_2)\nu_3^\top\phi_L(a,s)[\phi_L(a,s)-\gamma\mathbb{E}\{\phi_L(\pi^{opt}(S_{t+1}),S_{t+1})|A_t=a,S_t=s\}]^\top(\nu_2-\nu_4).$$

The first term is $O(L)$, by the condition $\sup_s \Phi_L(s)=O(\sqrt{L})$ in (A4). Similarly, the second term is $O(L)$ as well. To summarize, we have shown that

$$|\Psi_t(a,s_1,\nu_1,\nu_2)-\Psi_t(a,s_2,\nu_3,\nu_4)|\le cL(\|\nu_1-\nu_3\|_2+\|\nu_2-\nu_4\|_2),$$

for some constant $c>0$.

For any ν_1,ν_2 with unit ℓ_2 norm, there exist $\nu_{1,0},\nu_{2,0}\in\Gamma_0$ that satisfy $\|\nu_1-\nu_{1,0}\|_2\le\varepsilon$ and $\|\nu_2-\nu_{2,0}\|_2\le\varepsilon$. As such, $\Psi_t(a,s_1,\nu_1,\nu_2)-\Psi_t(a,s_2,\nu_1,\nu_2)$ can be upper bounded by

$$\sup_{\nu_{1,0},\nu_{2,0}\in\Gamma_0}\left|\frac{1}{N(T_2-T_1)}\sum_{i=1}^N\sum_{t=T_1}^{T_2-1}\sum_a\{\Psi_t(a,S_{i,t},\nu_{1,0},\nu_{2,0})-\mathbb{E}\Psi_t(a,S_{i,t},\nu_{1,0},\nu_{2,0})\}\right|+\frac{2cL}{(NT)^2}.$$

It remains to establish a uniform upper bound for the first term. We aim to apply the concentration inequality developed by Alquier et al. (2019). However, a direct application of Theorem 3.1 in Alquier et al. (2019) would yield a sub-optimal bound. This is because each summand $\Psi_t(a,S_t,\nu_{1,0},\nu_{2,0})$ is not bounded, since $\|\phi_L\|_2$ is allowed to diverge with the number of observations. To obtain a sharper bound, we further decompose the first term into the sum of the following two terms:

$$\begin{aligned}&\sup_{\nu_{1,0},\nu_{2,0}\in\Gamma_0}\left|\frac{1}{N(T_2-T_1)}\sum_{i=1}^N\sum_{t=T_1}^{T_2-1}\sum_a[\Psi_t(a,S_{i,t},\nu_{1,0},\nu_{2,0})-\mathbb{E}\{\Psi_t(a,S_{i,t},\nu_{1,0},\nu_{2,0})|S_{i,t-1}\}]\right|\\+&\sup_{\nu_{1,0},\nu_{2,0}\in\Gamma_0}\left|\frac{1}{N(T_2-T_1)}\sum_{i=1}^N\sum_{t=T_1}^{T_2-1}\sum_a[\mathbb{E}\{\Psi_t(a,S_{i,t},\nu_{1,0},\nu_{2,0})|S_{i,t-1}\}-\mathbb{E}\Psi_t(a,S_{i,t},\nu_{1,0},\nu_{2,0})]\right|.\end{aligned}\tag{B.27}$$

The first term corresponds to a sum of martingale difference. Using similar arguments in showing $\sup_{T_1,T_2}\|\widehat{W}_{[T_1,T_2]}-\widehat{W}^*_{[T_1,T_2]}\|_2=O(\sqrt{(\epsilon NT)^{-1}L\log(NT)})$, we can show that the first term in (B.27) is of the order $O(\sqrt{(\epsilon NT)^{-1}L\log(NT)})$, with probability at least $1-O(N^{-1}T^{-1})$, where the big-O term is uniform in $\{(T_1,T_2):T_2-T_1\ge\epsilon T\}$.

As for the second term, notice that by definition, $\mathbb{E}\{\Psi_t(a, S_t, \nu_{1,0}, \nu_{2,0})|S_{t-1} = s\}$ equals

$$\int_{s'} \pi_t^b(a|s')\nu_{1,0}^\top \phi_L(a, s')[\phi_L(a, s') - \gamma\mathbb{E}\{\phi_L(\pi^{opt}(S_{t+1}), S_{t+1})|A_t = a, S_t = s'\}]^\top \nu_{2,0} p_{t-1}(s'|a, s)ds'.$$

Under the p-smoothness condition in (A2), $\mathbb{E}\{\Psi_t(a, S_t, \nu_{1,0}, \nu_{2,0})|S_{t-1} = s\}$ is a Lipschitz continuous function of s. However, unlike $\Psi_t(a, s, \nu_{1,0}, \nu_{2,0})$, the integrand

$$|\pi_t^b(a|s')\nu_{1,0}^\top \phi_L(a, s')[\phi_L(a, s') - \gamma\mathbb{E}\{\phi_L(\pi^{opt}(S_{t+1}), S_{t+1})|A_t = a, S_t = s'\}]^\top \nu_{2,0}|$$

is upper bounded by a constant, under the condition $\max_t \lambda_{\max}\{\mathbb{E}\phi_L(A_t, S_t)\phi_L^\top(A_t, S_t)\} = O(1)$ in (A4). See e.g., Equation (E.77) of Shi et al. (2021). As such, the Lipschitz constant is uniformly bounded by some constant. Consequently, the conditions in the statement of Theorem 3.1 in Alquier et al. (2019) are satisfied. We can apply Theorem 3.1 to the mean zero random variable

$$\frac{1}{N(T_2 - T_1)} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \sum_a [\mathbb{E}\{\Psi_t(a, S_{i,t}, \nu_{1,0}, \nu_{2,0})|S_{i,t-1}\} - \mathbb{E}\Psi_t(a, S_{i,t}, \nu_{1,0}, \nu_{2,0})],$$

for each combination of $\nu_{1,0}, \nu_{2,0}, T_1, T_2$, and show that it is of the order $O(\sqrt{\epsilon L(NT)^{-1}\log(NT)})$ with probability at least $1 - O(N^{-CL}T^{-CL})$, for some sufficiently large constant $C > 0$. By Bonferroni's inequality, we can show that

$$\sup_{T_2 - T_1 \geq \epsilon T} \sup_{\nu_{1,0}, \nu_{2,0} \in \Gamma_0} \left| \frac{1}{N(T_2 - T_1)} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \sum_a [\Psi_t(a, S_{i,t}, \nu_{1,0}, \nu_{2,0}) - \mathbb{E}\{\Psi_t(a, S_{i,t}, \nu_{1,0}, \nu_{2,0})|S_{i,t-1}\}] \right|,$$

is upper bounded by $O(\sqrt{L\epsilon^{-1}(NT)^{-1}\log(NT)})$ with probability at least $1 - O(N^{-1}T^{-1})$.

B.5 Proof of Theorem 2

Without loss of generality, assume $T_0 = 0$. We only consider the ℓ_1-type test. The proof for the maximum-type test can be similarly derived. Under the given conditions on T^*, we have

$$\begin{aligned} \text{TS}_1 &\geq \sqrt{\frac{T^*(T - T^*)}{T^2}} \left\{ \frac{1}{NT} \sum_{i,t} |\widehat{Q}_{[0,T^*]}(A_{i,t}, S_{i,t}) - \widehat{Q}_{[T^*,T]}(A_{i,t}, S_{i,t})| \right\} \\ &\geq \sqrt{\epsilon(1-\epsilon)} \left\{ \frac{1}{NT} \sum_{i,t} |\widehat{Q}_{[0,T^*]}(A_{i,t}, S_{i,t}) - \widehat{Q}_{[T^*,T]}(A_{i,t}, S_{i,t})| \right\}. \end{aligned} \tag{B.28}$$

Under (A3), there exist some β_0^* and β_T^* such that

$$\sup_{a,s} |Q_0^{opt}(a,s) - \phi_L^\top(a,s)\beta_0^*| = O(L^{-c_2}) \text{ and } \sup_{a,s} |Q_T^{opt}(a,s) - \phi_L^\top(a,s)\beta_T^*| = O(L^{-c_2}) \quad \text{(B.29)}$$

Using similar arguments in the proof of Lemma B.1, we can show that

$$\max(\|\widehat{\beta}_{[0,T^*]} - \beta_0^*\|_2, \|\widehat{\beta}_{[T^*,T]} - \beta_T^*\|_2) = O(L^{-c_2}) + O(\sqrt{L(\epsilon NT)^{-1}\log(NT)}), \quad \text{(B.30)}$$

WPA1. Using similar arguments in Step 1 of the proof of Theorem 1, we can show that the last line of (B.28) can be well-approximated by

$$\frac{\sqrt{\epsilon(1-\epsilon)}}{T} \sum_{t=0}^{T-1} \sum_a \int_s |\widehat{Q}_{[0,T^*]}(a,s) - \widehat{Q}_{[T^*,T]}(a,s)| \pi_t^b(a|s) p_t^b(s) ds, \quad \text{(B.31)}$$

with the approximation error upper bounded by $\sqrt{\epsilon(1-\epsilon)L(NT)^{-1}\log(NT)}$, WPA1.

By (A4) and Cauchy-Schwarz inequality, we have

$$\sum_a \int_s |\phi_L^\top(a,s)\nu| ds \le \sum_a \sqrt{\int_s |\phi_L^\top(a,s)\nu|^2 ds} \preceq \|\nu\|_2 \lambda_{\max}\left(\int_s \Phi_L(s)\Phi_L^\top(s) ds\right) \preceq \|\nu\|_2.$$

This together with (B.28), (B.30) and (B.31) yields that

$$\text{TS}_1 \ge \frac{\sqrt{\epsilon(1-\epsilon)}}{T} \sum_{t=0}^{T-1} \sum_a \int_s |\phi_L^\top(a,s)(\beta_0^* - \beta_T^*)| \pi_t^b(a|s) p_t^b(s) ds$$
$$+ O(\sqrt{L(NT)^{-1}\log(NT)}) + O(L^{-c_2}),$$

WPA1. Combining this together with (B.29) yields that

$$\text{TS}_1 \ge \frac{\sqrt{\epsilon(1-\epsilon)}}{T} \sum_{t=0}^{T-1} \sum_a \int_s |Q_0^{opt}(a,s) - Q_T^{opt}(a,s)| \pi_t^b(a|s) p_t^b(s) ds$$
$$+ O(\sqrt{L(NT)^{-1}\log(NT)}) + O(L^{-c_2}),$$

WPA1. In addition, using similar arguments in Step 2 of the proof of Theorem 1, we can show that the bootstrapped test statistic TS_1^b is upper bounded by $O(\sqrt{L(NT)^{-1}\log(NT)})$, WPA1. Under the given condition on Δ_1, TS_1 is much larger than the upper αth quantile of TS_1^b, WPA1. As such, the power of the proposed test approaches 1 under the alternative hypothesis. This completes the proof.

B.6 Proof of Theorem 3

We first show the consistency of the estimated Q-function. In FQI, we iteratively update the Q-function according to the formula,

$$Q^{(k+1)} = \arg\min_{Q} \sum_{i,t} \left\{ R_{i,t} + \gamma \max_{a} Q^{(k)}(a, S_{i,t+1}) - Q(A_{i,t}, S_{i,t}) \right\}^2 .$$

At the kth iteration, we define the population-level Q-function

$$Q^{(k+1),*}(a,s) = r(a,s) + \gamma \max_{a'} \int_{s'} Q^{(k)}(a', s') p(s'|a,s) ds'. \tag{B.32}$$

According to the Bellman optimality equation,

$$Q^{opt}(a,s) = r(a,s) + \gamma \max_{a'} \int_{s'} Q^{opt}(a', s') p(s'|a,s) ds'.$$

It follows that

$$\begin{aligned}\sup_{a,s} |Q^{opt}(a,s) - Q^{(k+1)}(a,s)| &\leq \sup_{a,s} |Q^{(k+1),*}(a,s) - Q^{(k+1)}(a,s)| + \sup_{a,s} |Q^{opt}(a,s) - Q^{(k+1),*}(a,s)| \\ &\leq \sup_{a,s} |Q^{(k+1),*}(a,s) - Q^{(k+1)}(a,s)| + \gamma \sup_{a,s} |Q^{opt}(a,s) - Q^{(k)}(a,s)|.\end{aligned}$$

Iteratively applying this inequality for $k = K, K-1, \cdots, 1$, we obtain that

$$\begin{aligned}\sup_{a,s} |Q^{opt}(a,s) - \widehat{Q}_{[T_1,T_2]}(a,s)| \leq \sum_{k} \gamma^{K-k} &\sup_{a,s} |Q^{(k+1),*}(a,s) - Q^{(k+1)}(a,s)| \\ &+ \gamma^{K+1} \sup_{a,s} |Q^{opt}(a,s) - Q^{(0)}(a,s)|.\end{aligned} \tag{B.33}$$

As K diverges to infinity, the second term on the RHS decays to zero. The first term is upper bounded by

$$\frac{1}{1-\gamma} \sup_{a,s,k} |Q^{(k+1),*}(a,s) - Q^{(k+1)}(a,s)|. \tag{B.34}$$

It remains to show this term decays to zero as the sample size approaches to infinity.

Let $\beta^{(k)}$ denote the estimated regression coefficients such that $Q^{(k)}(a,s) = \phi_L^\top(a,s)\beta^{(k)}$. We claim that there exist some constants $C, \bar{C} > 0$ such that

$$\sup_{a,s} \max_{k \in \{0,\cdots,j\}} |Q^{(k)}(a,s)| \leq C \text{ and } \max_{k \in \{0,\cdots,j\}} \|\beta^{(k)}\|_2 \leq \bar{C}, \tag{B.35}$$

with probability at least $1-(j+1)/(NT)$, for sufficiently large NT. The values of C and $\bar{C}$ will be specified later.

We will prove this assertion by induction. Using similar arguments in the proof of Lemma B.2, we can show that

$$\left\| \frac{1}{N(T_2-T_1)} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \phi_L(A_{i,t}, S_{i,t})^\top - \frac{1}{T_2-T_1} \sum_{t=T_1}^{T_2-1} \mathbb{E} \phi_L(A_t, S_t) \phi_L(A_t, S_t)^\top \right\|_2 \le c\sqrt{L(\epsilon NT)^{-1} \log(NT)}, \tag{B.36}$$

for some constant $c > 0$, with probability at least $1 - N^{-1}T^{-1}$. Under (A6), $\lambda_{\min}\{(T_2 - T_1)^{-1} \sum_{t=T_1}^{T_2-1} \mathbb{E}\phi_L(A_t, S_t)\phi_L(A_t, S_t)^\top\}$ is uniformly bounded away from zero. On the event set defined by (B.36), for sufficiently large NT, there exists some $\bar{c} > 0$ such that

$$\lambda_{\min} \left\{ \frac{1}{N(T_2-T_1)} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \phi_L(A_{i,t}, S_{i,t})^\top \right\} \ge \bar{c}. \tag{B.37}$$

When $k = 0$, this posit assertion in (B.35) holds as long as $\sup_{a,s} |Q^{(0)}(a, s)| \le C$ and $\|\beta^{(0)}\|_2 \le \bar{C}$. Suppose the assertion holds for $k = 0, 1, \cdots, J$. We aim to prove this assertion holds for $k = J + 1$. Since the reward is uniformly bounded, so is $\sup_{a,s} |r(a, s)|$. We will choose C to be such that $C \ge 2(1-\gamma)^{-1} \sup_{a,s} |r(a, s)|$. It follows from (B.32) that $\sup_{a,s} |Q^{(k+1),*}(a, s)| \le (1 + \gamma)C/2$. In addition, define

$$\beta^{(k+1),*} = \left\{ \frac{1}{T_2-T_1} \sum_{t=T_1}^{T_2-1} \mathbb{E}\phi_L(A_t, S_t)\phi_L(A_t, S_t)^\top \right\}^{-1} \left\{ \frac{1}{T_2-T_1} \sum_{t=T_1}^{T_2-1} \mathbb{E}\phi_L(A_t, S_t) Q^{(k+1),*}(A_t, S_t) \right\}.$$

By (A4) and (A6), there exist some constants $c, C > 0$ such that $\|\beta^{(k+1),*}\|_2 \le cC$. We choose $\bar{C}$ to be such that $\bar{C} \ge 2cC$. As such, it suffices to show that on the set defined in (B.37), the estimation errors $\sup_{a,s} |Q^{(k+1)}(a, s) - Q^{(k+1),*}(a, s)| \le (1-\gamma)C/2$ and $\|\beta^{(k+1)} - \beta^{(k+1),*}\|_2 \le cC$, with probability at least $1 - (NT)^{-1}$.

Under (A2), there exists some $\beta^{(k+1),**}$ such that $\sup_{a,s} |Q^{(k+1),*}(a, s) - \phi_L^\top(a, s)\beta^{(k+1),**}| \le c_0 L^{-c_2}$ for some constant $c_0 > 0$. Using similar arguments to the proof of Lemma B.1, we can show that $\sup_k \|\beta^{(k+1),**} - \beta^{(k+1),*}\|_2 = O(L^{-c_2})$. Since L is proportional to $(NT)^{c_5}$ for some $c_5 > 0$, for sufficiently large NT, we have $\|\beta^{(k+1),**} - \beta^{(k+1),*}\|_2 \le cC/2$ and hence

$\|\beta^{(k+1),**}\|_2 \le 3cC/2$. Therefore, to show $\|\beta^{(k+1)} - \beta^{(k+1),*}\|_2 \le cC$, it suffices to show $\|\beta^{(k+1)} - \beta^{(k+1),**}\|_2 \le cC/2$.

By definition, we have

$$\beta^{(k+1)} - \beta^{(k+1),**} = \left\{ \frac{1}{N(T_2 - T_1)} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \phi_L(A_{i,t}, S_{i,t})^\top \right\}^{-1}$$

$$\times \left[\frac{1}{N(T_2 - T_1)} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \{ R_{i,t} + \gamma \max_a \phi_L^\top(a, S_{i,t+1}) \beta^{(k)} - \phi_L^\top(A_{i,t}, S_{i,t}) \beta^{(k+1),**} \} \right].$$

On the set defined in (B.37), $\|\beta^{(k+1)} - \beta^{(k+1),**}\|_2$ is upper bounded by

$$\sup_{\substack{\|\beta_0^{(k)}\|_2 \le 2cC, \|\beta_0^{(k+1),**}\|_2 \le 3cC/2 \\ \sup_{a,s} |r(a,s) + \gamma \int_{s'} \max_{a'} \phi_L^\top(a',s') \beta_0^{(j)} p(s'|a,s) ds' - \phi_L^\top(a,s) \beta_0^{(k+1),**}| \le c_0 L^{-c_2}}} \left| \frac{\bar{c}^{-1}}{N(T_2 - T_1)} \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \right.$$

$$\left. \times \{ R_{i,t} + \gamma \max_a \phi_L^\top(a, S_{i,t+1}) \beta_0^{(k)} - \phi_L^\top(A_{i,t}, S_{i,t}) \beta_0^{(k+1),**} \} \right|.$$

Using similar arguments to Step 1 of the proof of Theorem 1, we can show that the upper bound is of the order $O(L^{-c_2}) + O(\sqrt{L(\epsilon NT)^{-1} \log(NT)})$, with probability at least $1 - (NT)^{-1}$. For sufficiently large NT, the assertion $\|\beta^{(k+1)} - \beta^{(k+1),**}\|_2 \le cC/2$ is automatically satisfied. Moreover, by (A4) and the condition that L is proportional to $(NT)^{c_5}$ for some $0 < c_5 < 1/2$, we obtain that $\sup_{a,s} |Q^{(k+1)}(a,s) - Q^{(k+1),*}(a,s)| \preceq L^{1/2 - \bar{C}} + L\sqrt{(\epsilon NT)^{-1} \log(NT)} \ll (1-\gamma)C/2$. The assertion is thus proven.

Under the given conditions on K, the maximum number of iterations, we obtain that both $\widehat{Q}_{[T_1, T_2]}$ and $\widehat{\beta}_{[T_1, T_2]}$ are uniformly bounded WPA1. In addition, using similar arguments, we can show that the estimation error (B.34) decays to zero, WPA1. According to (B.33), we have

$$\max_{K/2 < k \le K} \sup_{a,s} |Q^{opt}(a,s) - Q^{(k+1)}(a,s)| \xrightarrow{p} 0.$$

By (A9), we obtain that $\arg\max_a \phi_L^\top(a,s) \beta^{(k+1)} = \pi^{opt}(s)$, for any $K/2 < k \le K$,WPA1. As such, $\beta^{(k+1)}$ equals

$$\left\{ \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \phi_L(A_{i,t}, S_{i,t})^\top \right\}^{-1} \left[\sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \{ R_{i,t} + \gamma \phi_L^\top(\pi^{opt}(S_{i,t+1}), S_{i,t+1}) \beta^{(k)} \} \right].$$

It follows that

$$(\beta^{(k+1)} - \beta^*_{[T_1,T_2]}) = \left\{ \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \phi_L(A_{i,t}, S_{i,t})^\top \right\}^{-1}$$
$$\times \left[\sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \{ R_{i,t} + \gamma \phi_L^\top (\pi^{opt}(S_{i,t+1}), S_{i,t+1}) \beta^{(k)} \} - \phi_L^\top (A_{i,t}, S_{i,t}) \beta^*_{[T_1,T_2]} \right]. \tag{B.38}$$

We next derive the asymptotic normality of $\widehat{\beta}_{[T_1,T_2]}$. Without loss of generality, suppose K is an even number. Applying (B.38) for $k = K, K-1, \cdots, K/2+1$, we obtain that

$$\widehat{\beta}_{[T_1,T_2]} - \beta^*_{[T_1,T_2]} = \widehat{\mathcal{L}}^{K/2} (\beta^{(K/2+1)} - \beta^*_{[T_1,T_2]}) + \sum_{j=0}^{K/2} \widehat{\mathcal{L}}^j \widehat{\ell},$$

where

$$\widehat{\mathcal{L}} = \gamma \left\{ \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \phi_L(A_{i,t}, S_{i,t})^\top \right\}^{-1} \left\{ \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \phi_L^\top (\pi^{opt}(S_{i,t+1}), S_{i,t+1}) \right\}$$
$$\text{and } \widehat{\ell} = \left\{ \sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \phi_L(A_{i,t}, S_{i,t})^\top \right\}^{-1}$$
$$\times \left[\sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \{ R_{i,t} + \gamma \phi_L^\top (\pi^{opt}(S_{i,t+1}), S_{i,t+1}) \beta^*_{[T_1,T_2]} \} - \phi_L^\top (A_{i,t}, S_{i,t}) \beta^*_{[T_1,T_2]} \right].$$

Similar to Lemma 3 of Shi et al. (2021), under (A6), we can show that $\|\widehat{\mathcal{L}}_2\| \leq \bar{\gamma}$ for some $\bar{\gamma} < 1$, WPA1. Under the given conditions on K, we obtain that

$$\widehat{\beta}_{[T_1,T_2]} - \beta^*_{[T_1,T_2]} = (I - \widehat{\mathcal{L}})^{-1} \widehat{\ell} + O\left(N^{-c_1} T^{-c_1}\right),$$

with probability at least $1 - O(N^{-1}T^{-1})$. Notice that

$$(I - \widehat{\mathcal{L}})^{-1} \widehat{\ell} = \left[\sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \{ \phi_L(A_{i,t}, S_{i,t}) - \gamma \phi_L(\pi^{opt}(S_{i,t+1}), S_{i,t+1}) \}^\top \right]^{-1}$$
$$\times \left[\sum_{i=1}^{N} \sum_{t=T_1}^{T_2-1} \phi_L(A_{i,t}, S_{i,t}) \{ R_{i,t} + \gamma \phi_L^\top (\pi^{opt}(S_{i,t+1}), S_{i,t+1}) \beta^*_{[T_1,T_2]} \} - \phi_L^\top (A_{i,t}, S_{i,t}) \beta^*_{[T_1,T_2]} \right].$$

The assertion follows using similar arguments to the proof of Lemma B.2. We omit the details to save space.

C More on the numerical study

C.1 Smooth Markov Chain Transition

The way we generate a smooth transition function is to first define a piecewise constant function, and smoothly connects the constant functions through a transformation. Specifically, let the piecewise constant function with two segments be $f(s,t) = f_1(s)I\{t \leq T^*\} + f_2(s)I\{t > T^*\}$, where f_1 and f_2 are functions not dependent on t.

We now introduce a smooth transformation $\phi(s) = \frac{\psi(s)}{\psi(s)+\psi(1-s)}$, where $\psi(s) = e^{-1/s}I\{s > 0\}$. Then $g(s; f_1, f_2, s_0, s_1) := f_1(s) + (f_2(s) - f_1(s))\phi\left(\frac{s-s_0}{s_1-s_0}\right)$ is a smooth function from f_1 to f_2 on the interval $[s_0, s_1]$. In addition, the transformed function $\widetilde{f}(s,t) = f_1(s)I\{t \leq T^* - \delta T\} + g(s; f_1, f_2, s_0, s_1)I\{T^* - \delta T < t < T^* + \delta T\} + f_2(s)I\{t > T^*\}$ is smooth in s. Here δ controls the smoothness of the transformation; smaller δ leads to more abrupt change and larger δ leads to smoother change. An example of $\widetilde{f}(s,t)$ is shown in Figure 1.

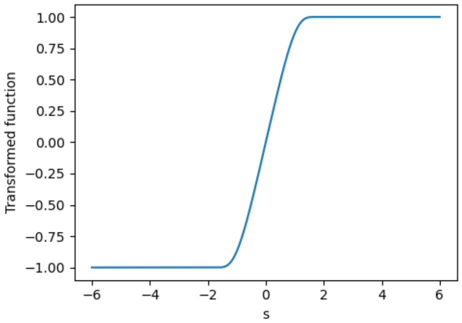

Figure 1: Smooth transformation of piecewise constant function: $f_1(s) = -I\{s \leq -2\}$ and $f_2(s) = I\{s \geq 2\}$.

The four settings of transition dynamics are specified as the following.

(1) Time-homogeneous state transition function and piecewise constant reward function:

$$S_{0,t+1} = 0.5A_{0,t}S_{0,t} + z_{0,t}, t \in [0,T].$$

$$R_{0,t} = \begin{cases} r_1(S_{0,t}, A_{0,t}; t) \equiv -1.5A_{0,t}S_{0,t}, & \text{if } t \in [0, T^*) \\ r_2(S_{0,t}, A_{0,t}; t) \equiv A_{0,t}S_{0,t}, & \text{if } t \in [T^*, T], \end{cases}$$

(2) Time-homogeneous state transition function and smooth reward function:

$$S_{0,t+1} = 0.5A_{0,t}S_{0,t} + z_{0,t}, t \in [0, T].$$

$$R_{0,t} = \begin{cases} r_1(S_{0,t}, A_{0,t}; t), & \text{if } t \in [0, T^* - \delta T), \\ g\left(S_{0,t}; r_1, r_2, T^* - \delta T, T^*\right), & \text{if } t \in [T^* - \delta T, T^*), \\ r_2(S_{0,t}, A_{0,t}; t), & \text{if } t \in [T^*, T]. \end{cases}$$

(3) Piecewise constant state transition and time-homogeneous reward function:

$$S_{0,t+1} = \begin{cases} F_1(S_{0,t}, A_{0,t}; t) \equiv -0.5A_{0,t}S_{0,t} + z_{0,t}, & \text{if } t \in [0, T^*), \\ F_2(S_{0,t}, A_{0,t}; t) \equiv 0.5A_{0,t}S_{0,t} + z_{0,t}, & \text{if } t \in [T^*, T]. \end{cases}$$

$$R_{0,t} = 0.25A_{0,t}S_{0,t}^2 + 4S_{0,t}, t \in [0, T].$$

(4) Smooth state transition and time-homogeneous reward function:

$$S_{0,t+1} = \begin{cases} F_1(S_{0,t}, A_{0,t}; t), & \text{if } t \in [0, T^*), \\ g\left(S_{0,t}; F_1, F_2, T^* - \delta T, T^*\right), & \text{if } t \in [T^* - \delta T, T^*), \\ F_2(S_{0,t}, A_{0,t}; t), & \text{if } t \in [T^*, T]. \end{cases}$$

$$R_{0,t} = 0.25A_{0,t}S_{0,t}^2 + 4S_{0,t}, t \in [0, T].$$

C.2 Simulation Setting of IHS Data

We initiate the state variables as independent normal distributions with $S_{i,0,1} \sim \mathcal{N}(20, 3)$, $S_{i,0,2} \sim \mathcal{N}(20, 2)$, and $S_{i,0,3} \sim \mathcal{N}(7, 1)$, and let them evolve according to

$$\begin{pmatrix} S_{i,t+1,1} \\ S_{i,t+1,2} \\ S_{i,t+1,3} \end{pmatrix} = \boldsymbol{W}_1(A_{i,t})\widetilde{S}_{i,t}I\{t \in [0, T^*)\} + \boldsymbol{W}_2(A_{i,t})\widetilde{S}_{i,t}I\{t \in [T^*, T]\} + \boldsymbol{z}_{i,t},$$

where the transition matrices are

$$\boldsymbol{W}_1(A_{i,t}) = \begin{pmatrix} 10 + 0.6A_{i,t} & 0.4 + 0.3A_{i,t} & 0.1 - 0.1A_{i,t} & -0.04 & 0.1 \\ 11 - 0.4A_{i,t} & 0.05 & 0 & 0.4 & 0 \\ 1.2 - 0.5A_{i,t} & -0.02 & 0 & 0.03 + 0.03A_{i,t} & 0.8 \end{pmatrix},$$
$$\boldsymbol{W}_2(A_{i,t}) = \begin{pmatrix} 10 - 0.6A_{i,t} & 0.4 - 0.3A_{i,t} & 0.1 + 0.1A_{i,t} & 0.04 & -0.1 \\ 11 - 0.4A_{i,t} & 0.05 & 0 & 0.4 & 0 \\ 1.2 + 0.5A_{i,t} & -0.02 & 0 & 0.03 - 0.03A_{i,t} & 0.8 \end{pmatrix},$$

$\widetilde{S}_{i,t} = (1, S_{i,t})^\top$, and $\boldsymbol{z}_{i,t} \sim \mathcal{N}_3(0, \operatorname{diag}(1, 1, 0.2))$ is random noise.

References

Alquier, P., Doukhan, P., and Fan, X. (2019). Exponential inequalities for nonstationary Markov chains. *Dependence Modeling*, 7(1):150–168.

Belloni, A. and Oliveira, R. I. (2018). A high dimensional central limit theorem for martingales, with applications to context tree models. *arXiv preprint arXiv:1809.02741*.

Birnbaum, Z. W. (1942). An inequality for mill's ratio. *The Annals of Mathematical Statistics*, 13(2):245–246.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597.

Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N. (2008). Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289.

Shi, C., Zhang, S., Lu, W., and Song, R. (2021). Statistical inference of the value function for reinforcement learning in infinite horizon settings. *Journal of Royal Statistical Society: Series B*, accepted.

Tropp, J. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270.