Supplementary Material: Sparse Additive Off-Policy Evaluation for Reinforcement Learning with Potentially Limited Number of Trajectories

Tuoyi Zhao[1], Chengchun Shi[2], Zhengling Qi[3] and Lan Wang[4]

[1] *Yale University*

[2] *London School of Economics and Political Science*

[3] *George Washington University*

[4] *University of Miami*

S1 Discussions on Sparsity Assumptions

S1.1 Sparse Q Function and Sparse Completeness

Recall the sparse completeness. Let $\mathcal{G}_{\mathcal{K}} = \{f : \mathcal{X}^d \times \mathcal{A} \to \mathbb{R} |\ f(x, a) = f(x_{\mathcal{K}}, a)\}$. Given the target policy π, define the following operator,

$$\Gamma^\pi f(x, a) = r(x, a) + \gamma \mathbb{E}^\pi[f(x', a')|x, a].$$

Assumption 1. sparse completeness. Given the target policy π, there exists an active set $\mathcal{K} \subseteq [d], |\mathcal{K}| \leq d_0$ such that $\mathcal{G}_{\mathcal{K}}$ is closed under the operator Γ^π.

The sparse completeness of target policy π guarantees the $\mathcal{G}_{\mathcal{K}}$ can capture the operator Γ^π. Note that $f \equiv 0$ belongs to $\mathcal{G}_\mathcal{K}$ for all $\mathcal{K}$, Sparse completeness under the target policy π guarantees that $\mathcal{G}_K$ is closed under the operator Γ^π. $r \in \mathcal{G}_\mathcal{K}$. As such, $\mathbb{E}^\pi[f(x', a')|x, a] \in \mathcal{G}_\mathcal{K}$ if $f \in \mathcal{G}_\mathcal{K}$. Denote $r^{(t)}(x, a) = \mathbb{E}^\pi (R_{0,t} \mid X_{0,0} = x, A_{0,0} = a)$, $r^{(0)}(x, a) = r(x, a)$. And hence,

$$Q^\pi(x, a) = \sum_{t \geq 0} \gamma^t r^{(t)}(x, a). \tag{S1.1}$$

To prove $Q \in \mathcal{G}_\mathcal{K}$, we only need to prove

$$r^{(t)}(x, a) = r^{(t)}(x_\mathcal{K}, a) \tag{S1.2}$$

By CMIA,

$$\begin{aligned} r^{(t)}(x, a) &= \mathbb{E}^\pi (R_{0,t} \mid X_{0,0} = x, A_{0,0} = a) \\ &= \mathbb{E}^\pi[\mathbb{E}^\pi (R_{0,t} \mid X_{0,1} = x', A_{0,1} = a') \,|X_{0,0} = x, A_{0,0} = a] \\ &= \mathbb{E}^\pi[r^{(t-1)}(x', a')|X_{0,0} = x, A_{0,0} = a] \end{aligned} \tag{S1.3}$$

Note that $r^{(0)}(x, a) = r(x, a) \in \mathcal{G}_\mathcal{K}$. If $r^{(t-1)}(x, a) \in \mathcal{G}_\mathcal{K}$, $r^{(t)}(x, a) = \mathbb{E}^\pi[r^{(t-1)}(x', a')|X_{0,0} = x, A_{0,0} = a] \in \mathcal{G}_\mathcal{K}$. $Q \in \mathcal{G}_\mathcal{K}$ is proved by induction. We proved the sparse completeness implied Assumption 1 directly.

S1.2 Sufficient State

Definition 1. Sufficient State. Ma et al. (2026) We say the sufficient state $\mathcal{X}_\mathcal{K}$ in an MDP if $R_t \perp\!\!\!\perp X_{t,\mathcal{K}^c} \mid (X_{t,\mathcal{K}}, A_t)$ and $X_{t+1,\mathcal{K}} \perp\!\!\!\perp X_{t,\mathcal{K}^c} \mid$

$(X_{t,\mathcal{K}}, A_t)$, for all $t \geq 0$.

If we further consider an online learning problem, then, based on the definition of a sufficient state, Ma et al. (2026) prove the following the proposition,

Proposition 1. *If $\mathcal{X}_\mathcal{K}$ is a sufficient state, then there exists an optimal policy that depends only on $\mathcal{X}_\mathcal{K}$.*

This proposition implies that the optimal policy π^* only depend on the sufficient state $\mathcal{X}_\mathcal{K}$, and allows us to only focus on policies that are functions of sufficient states.

S1.3 Proof of Proposition 1

Proof. For an index set $\mathcal{K} \subseteq [d]$, define

$$\mathcal{G}_\mathcal{K} = \left\{ f : \mathcal{X}^d \times \mathcal{A} \to \mathbb{R} : f(x,a) = \widetilde{f}(x_\mathcal{K}, a) \text{ for some bounded function } \widetilde{f} \right\}.$$

Thus, $\mathcal{G}_\mathcal{K}$ is the class of bounded functions that depend on the state only through $x_\mathcal{K}$.

For any bounded function f, define the target-policy transition operator

$$(\mathcal{P}^\pi f)(x,a) = \int_{\mathcal{X}^d} \sum_{a' \in \mathcal{A}} f(x', a') \pi(a' \mid x') P(dx' \mid x, a),$$

where $P(\cdot \mid x, a)$ denotes the transition kernel. The corresponding Bellman operator is

$$(\Gamma^\pi f)(x, a) = r(x, a) + \gamma(\mathcal{P}^\pi f)(x, a).$$

We show that, under either sufficient condition, Γ^π maps $\mathcal{G}_\mathcal{K}$ into itself.

First, suppose that Sufficient Condition 1 holds. Let $x, \bar{x} \in \mathcal{X}^d$ satisfy $x_\mathcal{K} = \bar{x}_\mathcal{K}$. Because the full transition kernel depends on the current state only through $x_\mathcal{K}$, we have $P(\cdot \mid x, a) = P(\cdot \mid \bar{x}, a)$. Therefore, for any bounded function f,

$$\begin{aligned}
(\mathcal{P}^\pi f)(x, a) &= \int_{\mathcal{X}^d} \sum_{a' \in \mathcal{A}} f(x', a')\pi(a' \mid x')P(dx' \mid x, a) \\
&= \int_{\mathcal{X}^d} \sum_{a' \in \mathcal{A}} f(x', a')\pi(a' \mid x')P(dx' \mid \bar{x}, a) \\
&= (\mathcal{P}^\pi f)(\bar{x}, a).
\end{aligned}$$

Thus, $(\mathcal{P}^\pi f)(x, a)$ depends on the current state only through $x_\mathcal{K}$. Since $r(x, a) = r(x_\mathcal{K}, a)$, it follows that $\Gamma^\pi f \in \mathcal{G}_\mathcal{K}$ whenever $f \in \mathcal{G}_\mathcal{K}$.

Next, suppose that Sufficient Condition 2 holds. Let $f \in \mathcal{G}_\mathcal{K}$, so that $f(x', a') = \widetilde{f}(x'_\mathcal{K}, a')$ for some bounded function $\widetilde{f}$. Let $P_\mathcal{K}(\cdot \mid x, a)$ denote the marginal transition kernel of $X'_\mathcal{K}$. Because the target policy depends only on the active coordinates, we have $\pi(a' \mid x') = \pi(a' \mid x'_\mathcal{K})$. Consequently,

$$(\mathcal{P}^\pi f)(x, a) = \int_{\mathcal{X}^{|\mathcal{K}|}} \sum_{a' \in \mathcal{A}} \widetilde{f}(x'_\mathcal{K}, a')\pi(a' \mid x'_\mathcal{K})P_\mathcal{K}(dx'_\mathcal{K} \mid x, a).$$

The assumed reduced transition property implies that $P_{\mathcal{K}}(\cdot \mid x, a) = P_{\mathcal{K}}(\cdot \mid \bar{x}, a)$, whenever $x_{\mathcal{K}} = \bar{x}_{\mathcal{K}}$. Hence,

$$(\mathcal{P}^\pi f)(x, a) = (\mathcal{P}^\pi f)(\bar{x}, a),$$

so that $\mathcal{P}^\pi f \in \mathcal{G}_{\mathcal{K}}$. Together with $r \in \mathcal{G}_{\mathcal{K}}$, this gives $\Gamma^\pi f \in \mathcal{G}_{\mathcal{K}}$.

Thus, under either sufficient condition, $\mathcal{G}_{\mathcal{K}}$ is closed under the Bellman operator Γ^π. To complete the argument, let $f_0 = 0$ and define the Bellman iterates $f_{m+1} = \Gamma^\pi f_m$, $m \geq 0$. Because $f_0 \in \mathcal{G}_{\mathcal{K}}$ and $\mathcal{G}_{\mathcal{K}}$ is closed under Γ^π, we have $f_m \in \mathcal{G}_{\mathcal{K}}$ for every m. Moreover, since $\gamma < 1$,

$$\|\Gamma^\pi f - \Gamma^\pi g\|_\infty \leq \gamma \|f - g\|_\infty.$$

Therefore, the Bellman iterates converge uniformly to the unique bounded fixed point of Γ^π. Since Q^π is a fixed point, the limit must equal Q^π. As $\mathcal{G}_{\mathcal{K}}$ is closed under uniform limits, it follows that $Q^\pi \in \mathcal{G}_{\mathcal{K}}$. Equivalently,

$$Q^\pi(x, a) = Q^\pi(x_{\mathcal{K}}, a) \qquad \text{for every } (x, a) \in \mathcal{X}^d \times \mathcal{A}.$$

Taking $S_a = \mathcal{K}$ for every $a \in \mathcal{A}$ and using $|\mathcal{K}| \ll d$ establishes Assumption 1.

$\square$

S2 Properties of B-spline basis functions

In this section, we summarize several properties of the B-spline basis functions used in our analysis. Let L denote the number of basis functions, and let m denote the order of the B-splines, where m is fixed and chosen according to the smoothness parameter v. Let

$$\boldsymbol{b}(x) = \big(b_1(x), \ldots, b_L(x)\big)^\top$$

be the normalized B-spline basis associated with a knot sequence $\{\tau_j\}$ (de Boor, 1976). We augment the knot sequence at the two boundaries such that

$$\tau_0 = \tau_1 = \cdots = \tau_m < \tau_{m+1} < \cdots < \tau_{L-1} < \tau_L = \tau_{L+1} = \cdots = \tau_{L+m}.$$

We further assume that the distinct knots are equally spaced, namely,

$$\tau_{m+1} - \tau_m = \cdots = \tau_L - \tau_{L-1}.$$

Define

$$d_j = \frac{\tau_{j+m} - \tau_j}{m}$$

for all admissible j. Since m is fixed and the knots are equally spaced, we have $d_j \asymp L^{-1}$ uniformly in j.

Throughout the proofs, we use the rescaled B-spline basis

$$B_j(x_{i,t}) = \sqrt{L}\, b_j(x_{i,t}), \qquad j = 1, \ldots, L, \quad t \geq 1.$$

Lemma 1. *For the knot sequence defined above and any $x \in \mathcal{X}$,*

$$\sum_{j=1}^{L} b_j(x) = 1 \qquad \textit{and} \qquad 0 \leq b_j(x) \leq 1, \quad j = 1, \ldots, L.$$

Moreover, there exists a constant $c_m > 0$*, depending only on the order* m*, such that*

$$\sup_{x \in \mathcal{X}} \|\boldsymbol{b}(x)\|_2 \le c_m.$$

Proof. The partition-of-unity property

$$\sum_{j=1}^{L} b_j(x) = 1$$

follows from Marsden's identity (de Boor, 1976). In addition, the B-spline basis functions are nonnegative, and therefore $0 \le b_j(x) \le 1$ for every j.

Since B-splines of order m have local support, at most m components of $\boldsymbol{b}(x)$ are nonzero at any fixed x. It follows that

$$\|\boldsymbol{b}(x)\|_2^2 = \sum_{j=1}^{L} b_j(x)^2 \le c_m.$$

This bound is purely a property of the B-spline basis and does not depend on the distribution of X. □

Lemma 2. *There exists a constant* $c^* > 0$*, independent of* L*, such that*

$$\sup_{x \in \mathcal{X}} \|\boldsymbol{B}(x)\|_\infty \le \sqrt{L}, \qquad \sup_{x \in \mathcal{X}} \|\boldsymbol{B}(x)\|_2 \le c_m \sqrt{L},$$

and

$$(c^*)^{-1} \le \lambda_{\min} \left\{ \int_{\mathcal{X}} \boldsymbol{B}(x)\boldsymbol{B}(x)^\top \, \mathrm{d}x \right\} \le \lambda_{\max} \left\{ \int_{\mathcal{X}} \boldsymbol{B}(x)\boldsymbol{B}(x)^\top \, \mathrm{d}x \right\} \le c^*. \tag{S2.4}$$

Proof. By Lemma 1,

$$\sup_{x\in\mathcal{X}} \|\boldsymbol{b}(x)\|_\infty \le 1 \qquad \text{and} \qquad \sup_{x\in\mathcal{X}} \|\boldsymbol{b}(x)\|_2 \le c_m.$$

Since $\boldsymbol{B}(x) = \sqrt{L}\boldsymbol{b}(x)$, the first two claims follow immediately.

It remains to establish the eigenvalue bounds. A standard stability property of B-spline bases implies that there exist constants $c_1, c_2 > 0$, depending only on the order m, such that, for every $\boldsymbol{c} = (c_1, \ldots, c_L)^\top \in \mathbb{R}^L$,

$$c_1 \sum_{j=1}^{L} d_j c_j^2 \le \left\| \sum_{j=1}^{L} c_j b_j \right\|_{L_2(\mathcal{X})}^2 \le c_2 \sum_{j=1}^{L} d_j c_j^2; \tag{S2.5}$$

see, for example, de Boor (1976). Because the knots are equally spaced and m is fixed, $d_j \asymp L^{-1}$ uniformly in j. Hence, there exist constants $\underline{c}, \overline{c} > 0$, independent of L, such that

$$\underline{c} L^{-1} \|\boldsymbol{c}\|_2^2 \le \left\| \sum_{j=1}^{L} c_j b_j \right\|_{L_2(\mathcal{X})}^2 \le \overline{c} L^{-1} \|\boldsymbol{c}\|_2^2.$$

Equivalently,

$$\underline{c} L^{-1} \|\boldsymbol{c}\|_2^2 \le \boldsymbol{c}^\top \left\{ \int_{\mathcal{X}} \boldsymbol{b}(x)\boldsymbol{b}(x)^\top \, \mathrm{d}x \right\} \boldsymbol{c} \le \overline{c} L^{-1} \|\boldsymbol{c}\|_2^2.$$

Therefore,

$$\underline{c} L^{-1} \le \lambda_{\min} \left\{ \int_{\mathcal{X}} \boldsymbol{b}(x)\boldsymbol{b}(x)^\top \, \mathrm{d}x \right\} \le \lambda_{\max} \left\{ \int_{\mathcal{X}} \boldsymbol{b}(x)\boldsymbol{b}(x)^\top \, \mathrm{d}x \right\} \le \overline{c} L^{-1}.$$

Finally, since $\boldsymbol{B}(x) = \sqrt{L}\boldsymbol{b}(x)$,

$$\int_{\mathcal{X}} \boldsymbol{B}(x)\boldsymbol{B}(x)^\top \, \mathrm{d}x = L \int_{\mathcal{X}} \boldsymbol{b}(x)\boldsymbol{b}(x)^\top \, \mathrm{d}x.$$

The desired result follows by taking $c^* = \max\{\underline{c}^{-1}, \overline{c}\}$. $\square$

Remark 1. In the main paper, we use the augmented basis vector

$$\boldsymbol{\Phi}(x) = \left(1, \boldsymbol{\Phi}_1(x_1)^\top, \ldots, \boldsymbol{\Phi}_d(x_d)^\top\right)^\top.$$

For each $k = 1, \ldots, d$, let

$$\overline{\boldsymbol{B}}_k(x_k) = (B_{k,1}(x_k), \ldots, B_{k,L+1}(x_k))^\top \in \mathbb{R}^{L+1}$$

denote the rescaled B-spline basis constructed using a deterministic sequence of equally spaced knots over the support of x_k. Thus, Lemma 2 applies to $\overline{\boldsymbol{B}}_k$ with $L+1$ in place of L.

Let

$$\boldsymbol{m}_k = \int_{\mathcal{X}} \overline{\boldsymbol{B}}_k(z)\, dG_k(z),$$

and let $\boldsymbol{A}_k \in \mathbb{R}^{(L+1)\times L}$ have orthonormal columns spanning the orthogonal complement of $\boldsymbol{m}_k$, so that

$$\boldsymbol{A}_k^\top \boldsymbol{A}_k = I_L, \qquad \boldsymbol{A}_k^\top \boldsymbol{m}_k = 0.$$

We define the centered B-spline contrast basis by

$$\boldsymbol{\Phi}_k(x_k) = \boldsymbol{A}_k^\top \overline{\boldsymbol{B}}_k(x_k) \in \mathbb{R}^L.$$

By construction,

$$\int_{\mathcal{X}} \boldsymbol{\Phi}_k(z)\, dG_k(z) = 0.$$

In particular, no nonzero linear combination of the components of $\mathbf{\Phi}_k$ is a constant function. Therefore, the explicit intercept and the spline coefficients are identified under this parameterization.

This construction preserves the standard spline approximation rate. Indeed, centering any spline approximation formed from $\overline{\boldsymbol{B}}_k$ produces an element of the linear span of $\mathbf{\Phi}_k$. Since the target component functions are centered under G_k, the uniform approximation error remains of order L^{-m}, up to a constant factor.

Moreover, since $\boldsymbol{A}_k$ has orthonormal columns, Lemma 2 gives

$$\sup_{x_k \in \mathcal{X}} \|\mathbf{\Phi}_k(x_k)\|_2 \le \sup_{x_k \in \mathcal{X}} \|\overline{\boldsymbol{B}}_k(x_k)\|_2 \le c_m \sqrt{L+1} \lesssim \sqrt{L}.$$

Consequently,

$$\sup_{x \in \mathcal{X}^d} \|\mathbf{\Phi}(x)\|_\infty \lesssim \sqrt{L}, \qquad \max_{1 \le k \le d} \sup_{x \in \mathcal{X}^d} \|\mathbf{\Phi}_k(x_k)\|_2 \lesssim \sqrt{L}.$$

In addition,

$$\int_{\mathcal{X}} \mathbf{\Phi}_k(z)\mathbf{\Phi}_k(z)^\top \, dz = \boldsymbol{A}_k^\top \left\{ \int_{\mathcal{X}} \overline{\boldsymbol{B}}_k(z)\overline{\boldsymbol{B}}_k(z)^\top \, dz \right\} \boldsymbol{A}_k.$$

It follows from Lemma 2 that there exists a constant $c^* > 0$, independent of L, such that

$$(c^*)^{-1} \le \lambda_{\min} \left\{ \int_{\mathcal{X}} \mathbf{\Phi}_k(z)\mathbf{\Phi}_k(z)^\top \, dz \right\} \le \lambda_{\max} \left\{ \int_{\mathcal{X}} \mathbf{\Phi}_k(z)\mathbf{\Phi}_k(z)^\top \, dz \right\} \le c^*.$$

S3 Additional technical details

Lemma 3. *Under assumption (A2) and (A3),* $\boldsymbol{\lambda}^{\boldsymbol{r}}_{\boldsymbol{min}}(\boldsymbol{\Sigma}) \geq \bar{c}/2$.

Proof. Denote the marginal distribution of $X_{0,t}$ by μ_t and the average distribution $\bar{\mu} = \sum_{t=0}^{T-1} \mu_t/T$. By the definition of $\boldsymbol{\Sigma}$, we have

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathbb{E}\left\{ \frac{1}{T} \sum_{t=0}^{T-1} (\boldsymbol{\xi}_{0,t}(\boldsymbol{\xi}_{0,t} - \gamma U_{0,t+1})^T) \right\} \\
&= \int_{x \in \mathcal{X}^d} \sum_{a \in \mathcal{A}} \boldsymbol{\xi}(x,a)(\boldsymbol{\xi}(x,a) - \gamma \mathbb{E}[U(X_{0,1})|X_{0,0} = x, A_{0,0} = a])^T b(a|x)\bar{\mu}(x) dx \\
&= \int_{x \in \mathcal{X}^d} \sum_{a \in \mathcal{A}} \boldsymbol{\xi}(x,a)(\boldsymbol{\xi}(x,a) - \gamma \boldsymbol{u}(x,a))^T b(a|x)\bar{\mu}(x) dx \\
&= \int_{x \in \mathcal{X}^d} \sum_{a \in \mathcal{A}} [\boldsymbol{\xi}(x,a)\boldsymbol{\xi}(x,a)^T - \gamma \boldsymbol{\xi}(x,a)\boldsymbol{u}(x,a)^T] b(a|x)\bar{\mu}(x) dx
\end{aligned} \tag{S3.6}$$

For $\forall \boldsymbol{a} \in \mathbb{R}^{2(1+dL)}$, denote

$$\begin{aligned}
\eta_1(\boldsymbol{a}) &= \int_{x \in \mathcal{X}^d} \sum_{a \in \mathcal{A}} \{\boldsymbol{a}^T \boldsymbol{\xi}(x,a)\}^2 b(a|x)\bar{\mu}(x) dx \\
\eta_2(\boldsymbol{a}) &= \int_{x \in \mathcal{X}^d} \sum_{a \in \mathcal{A}} \{\boldsymbol{a}^T u(x,a)\}^2 b(a|x)\bar{\mu}(x) dx
\end{aligned} \tag{S3.7}$$

It follows from Cauchy-Schwarz inequality that

$$\int_{x \in \mathcal{X}^d} \sum_{a \in \mathcal{A}} [\boldsymbol{a}^T \boldsymbol{\xi}(x,a)\boldsymbol{u}(x,a)^T \boldsymbol{a}] b(a|x)\bar{\mu}(x) dx \leq \eta_1(\boldsymbol{a})^{1/2} \eta_2(\boldsymbol{a})^{1/2} \tag{S3.8}$$

Therefore

$$\boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a} \geq \eta_1(\boldsymbol{a}) - \gamma \eta_1(\boldsymbol{a})^{1/2} \eta_2(\boldsymbol{a})^{1/2} = \eta_1(\boldsymbol{a})^{1/2} \left(\frac{\eta_1(\boldsymbol{a}) - \gamma^2 \eta_2(\boldsymbol{a})}{\eta_1(\boldsymbol{a})^{1/2} + \gamma \eta_2(\boldsymbol{a})^{1/2}}\right) \tag{S3.9}$$

Under the assumption (A3), if $\boldsymbol{a} \in \mathbb{C}$, $\eta_1(\boldsymbol{a}) - \gamma^2 \eta_2(\boldsymbol{a}) \geq \bar{c} \|\boldsymbol{a}\|_2^2 \geq 0$, $\eta_1(\boldsymbol{a})^{1/2} \geq \gamma \eta_2(\boldsymbol{a})^{1/2}$. We have

$$\boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a} \geq \frac{\bar{c}}{2} \|\boldsymbol{a}\|_2^2 \quad \forall \boldsymbol{a} \in \mathbb{C} \tag{S3.10}$$

□

Lemma 4. *There exists a good event $\mathcal{G}_1$ satisfying*

$$P(\mathcal{G}_1) \geq 1 - \exp\{-c_2 \log(nT \vee d)\}$$

such that, on $\mathcal{G}_1$, $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)_\oplus$ given by (3.5) *is feasible for the optimization problem* (3.11) *with*

$$\lambda = c_1 \sqrt{L} \frac{\log(nT \vee d)}{\sqrt{nT}},$$

where c_1 and c_2 are positive constants.

Proof. Throughout this proof, $C_1, C_2, \ldots$ denote positive constants that are local to the present proof and are re-indexed from C_1. They are distinct from the lowercase constants $c_1, c_2, \ldots$ appearing in the statements of the results.

Define

$$\varepsilon_{i,t} = R_{i,t} + \gamma \sum_{a \in \mathcal{A}} Q^{\pi}(X_{i,t+1}, a)\pi(a \mid X_{i,t+1}) - Q^{\pi}(X_{i,t}, A_{i,t}), \tag{S3.11}$$

and

$$\begin{aligned} e_{i,t} = \gamma \sum_{a \in \mathcal{A}} &\left\{ Q^{\pi}(X_{i,t+1}, a) - \alpha_a^* - \sum_{k=1}^{d} \Phi_k^T(X_{k,i,t+1}) \boldsymbol{\beta}_{k,a}^* \right\} \pi(a \mid X_{i,t+1}) \\ &- \left\{ Q^{\pi}(X_{i,t}, A_{i,t}) - \alpha_{A_{i,t}}^* - \sum_{k=1}^{d} \Phi_k^T(X_{k,i,t}) \boldsymbol{\beta}_{k,A_{i,t}}^* \right\}. \end{aligned} \tag{S3.12}$$

Under the smoothness condition on $q_{k,a}^{\pi}(\cdot)$ and the sparsity assumption, $Q^{\pi}(x,a)$ is uniformly bounded by a constant C_1. It follows that

$$|\varepsilon_{i,t}| \leq c_r + (1+\gamma)C_1 \leq c_r + 2C_1 = c_{\varepsilon}, \qquad \forall i, t. \tag{S3.13}$$

Combined with (3.5), the definition of $e_{i,t}$ implies that

$$|e_{i,t}| \leq (1+\gamma) d_0 C_2 L^{-m} \leq 2 d_0 C_2 L^{-m}, \qquad \forall i, t. \tag{S3.14}$$

Our goal is to prove that, with high probability,

$$\max_k \left\| \frac{1}{nT} \sum_{i,t} \boldsymbol{\xi}_{G_k,i,t} (\varepsilon_{i,t} - e_{i,t}) \right\|_2 \leq \lambda. \tag{S3.15}$$

For any $k = 1, \ldots, d$, by (S3.14),

$$\begin{aligned} \left\| \frac{1}{nT} \sum_{i,t} \boldsymbol{\xi}_{G_k,i,t} e_{i,t} \right\|_2 &\leq \frac{1}{nT} \sum_{i,t} \left\| \boldsymbol{\xi}_{G_k,i,t} e_{i,t} \right\|_2 \\ &\leq 2 d_0 C_2 L^{-m} \frac{1}{nT} \sum_{i,t} \left\| \boldsymbol{\xi}_{G_k,i,t} \right\|_2. \end{aligned} \tag{S3.16}$$

Moreover,

$$\left\|\boldsymbol{\xi}_{G_k,i,t}\right\|_2 \leq \max_k \sup_{x\in\mathcal{X}} \|\Phi_{G_k}(x)\|_2 . \tag{S3.17}$$

Together with the bound in Remark 1, we have

$$\left\|\frac{1}{nT}\sum_{i,t}\boldsymbol{\xi}_{G_k,i,t}e_{i,t}\right\|_2 \leq 2d_0c_mC_2L^{1/2-m}. \tag{S3.18}$$

By the triangle inequality, we have

$$\begin{aligned}
&P\left(\frac{1}{nT}\left\|\sum_{i,t}\boldsymbol{\xi}_{G_k,i,t}(\varepsilon_{i,t}-e_{i,t})\right\|_2 \geq \lambda\right)\\
&\quad\leq P\left(\frac{1}{nT}\left\|\sum_{i,t}\boldsymbol{\xi}_{G_k,i,t}\varepsilon_{i,t}\right\|_2 \geq \lambda - \frac{1}{nT}\left\|\sum_{i,t}\boldsymbol{\xi}_{G_k,i,t}e_{i,t}\right\|_2\right)\\
&\quad\leq P\left(\frac{1}{nT}\left\|\sum_{i,t}\boldsymbol{\xi}_{G_k,i,t}\varepsilon_{i,t}\right\|_2 \geq \lambda - 2d_0c_mC_2L^{1/2-m}\right).
\end{aligned} \tag{S3.19}$$

Denote

$$\widetilde{\lambda} = \lambda - 2d_0c_mC_2L^{1/2-m}. \tag{S3.20}$$

For any integer $1 \leq g \leq nT$, let $i(g)$ and $t(g)$ be defined by

$$i(g) = \left\lfloor\frac{g-1}{T}\right\rfloor + 1, \qquad t(g) = g - \{i(g)-1\}T - 1. \tag{S3.21}$$

Then

$$g = \{i(g)-1\}T + t(g) + 1, \tag{S3.22}$$

where $1 \leq i(g) \leq n$ and $0 \leq t(g) \leq T-1$.

We define the following filtration $\{\boldsymbol{\sigma}(\mathcal{F}^{(g)})\}_{g=0}^{nT}$ iteratively. Begin with

$$\mathcal{F}^{(0)} = \{X_{1,0}, A_{1,0}\}.$$

For $1 \le g \le nT$, define

$$\mathcal{F}^{(g)} = \mathcal{F}^{(g-1)} \cup \left\{ R_{i(g),t(g)}, X_{i(g),t(g)+1}, A_{i(g),t(g)+1} \right\}, \qquad \text{if } t(g) < T-1,$$

$$\mathcal{F}^{(g)} = \mathcal{F}^{(g-1)} \cup \left\{ R_{i(g),T-1}, X_{i(g),T}, X_{i(g)+1,0}, A_{i(g)+1,0} \right\}, \quad \text{if } t(g) = T-1 \text{ and } i(g) < n,$$

$$\mathcal{F}^{(g)} = \mathcal{F}^{(g-1)} \cup \left\{ R_{n,T-1}, X_{n,T} \right\}, \qquad \text{if } t(g) = T-1 \text{ and } i(g) = n.$$

By the Bellman equation, together with MA and CMIA,

$$\mathbb{E}\left\{ \varepsilon_{i(g),t(g)} \mid \boldsymbol{\sigma}(\mathcal{F}^{(g-1)}) \right\} = \mathbb{E}\left\{ \varepsilon_{i(g),t(g)} \mid X_{i(g),t(g)}, A_{i(g),t(g)} \right\} = 0. \qquad \text{(S3.23)}$$

Notice that $\boldsymbol{\xi}_{G_k,i,t}$ is a deterministic function of $X_{i,t}$ and $A_{i,t}$. Therefore,

$$\mathbb{E}\left\{ \boldsymbol{\xi}_{G_k,i(g),t(g)} \varepsilon_{i(g),t(g)} \mid \boldsymbol{\sigma}(\mathcal{F}^{(g-1)}) \right\} = 0. \qquad \text{(S3.24)}$$

This indicates that $\left\{ \boldsymbol{\xi}_{G_k,i(g),t(g)} \varepsilon_{i(g),t(g)} \right\}_{g=1}^{nT}$ is a vector-valued martingale difference sequence with respect to the filtration $\{\boldsymbol{\sigma}(\mathcal{F}^{(g)})\}_{g=0}^{nT}$.

The martingale difference sequence $\boldsymbol{\xi}_{G_k}^{(g)} \varepsilon^{(g)}$ satisfies the following almost-sure bound:

$$\left\| \boldsymbol{\xi}_{G_k}^{(g)} \varepsilon^{(g)} \right\|_2 \le c_\varepsilon \max_k \sup_{x \in \mathcal{X}} \|\Phi_k(x)\|_2 \le c_\varepsilon c_m \sqrt{L}. \qquad \text{(S3.25)}$$

Moreover,

$$\begin{aligned} &\left\| \sum_{g=1}^{nT} \mathbb{E}\left\{ \left(\boldsymbol{\xi}_{G_k}^{(g)} \varepsilon^{(g)} \right)^T \boldsymbol{\xi}_{G_k}^{(g)} \varepsilon^{(g)} \mid \boldsymbol{\sigma}(\mathcal{F}^{(g-1)}) \right\} \right\|_2 \\ &\qquad \le c_\varepsilon^2 \sum_{g=1}^{nT} \left\| \left(\boldsymbol{\xi}_{G_k}^{(g)} \right)^T \boldsymbol{\xi}_{G_k}^{(g)} \right\|_2 \\ &\qquad \le nT c_\varepsilon^2 c_m^2 L, \end{aligned} \qquad \text{(S3.26)}$$

and

$$\begin{aligned}
&\left\|\sum_{g=1}^{nT}\mathbb{E}\left\{\boldsymbol{\xi}_{G_k}^{(g)}\varepsilon^{(g)}\left(\boldsymbol{\xi}_{G_k}^{(g)}\varepsilon^{(g)}\right)^T \mid \boldsymbol{\sigma}(\mathcal{F}^{(g-1)})\right\}\right\|_2 \\
&\quad\le c_\varepsilon^2\sum_{g=1}^{nT}\left\|\boldsymbol{\xi}_{G_k}^{(g)}\left(\boldsymbol{\xi}_{G_k}^{(g)}\right)^T\right\|_2 \\
&\quad\le nTc_\varepsilon^2c_m^2L. \qquad \text{(S3.27)}
\end{aligned}$$

By the matrix Freedman inequality (Tropp, 2011, Corollary 1.3), we have

$$P\left(\left\|\sum_{i,t}\boldsymbol{\xi}_{G_k,i,t}\varepsilon_{i,t}\right\|_2 \ge nT\widetilde{\lambda}\right) \le (L+1)\exp\left\{-\frac{nT\widetilde{\lambda}^2/2}{c_\varepsilon^2c_m^2L + c_\varepsilon c_m\sqrt{L}\widetilde{\lambda}/3}\right\}. \qquad \text{(S3.28)}$$

Let

$$\widetilde{\lambda} = c_\lambda\sqrt{L}\frac{\log(nT\vee d)}{\sqrt{nT}},$$

where $c_\lambda > 0$ is a sufficiently large constant. By (S3.28),

$$P\left(\left\|\sum_{i,t}\boldsymbol{\xi}_{G_k,i,t}\varepsilon_{i,t}\right\|_2 \ge nT\widetilde{\lambda}\right) \le (L+1)\exp\left[-\frac{c_\lambda^2L\{\log(nT\vee d)\}^2/2}{c_\varepsilon^2c_m^2L + c_\varepsilon c_m c_\lambda L\log(nT\vee d)/(3\sqrt{nT})}\right]. \qquad \text{(S3.29)}$$

Using

$$\frac{x}{a+b} \ge \frac{1}{2}\min\left\{\frac{x}{a}, \frac{x}{b}\right\},$$

the exponent in the preceding display is bounded above by

$$-\min\left\{\frac{c_\lambda^2\{\log(nT\vee d)\}^2}{4c_\varepsilon^2c_m^2}, \frac{3c_\lambda\sqrt{nT}\log(nT\vee d)}{4c_\varepsilon c_m}\right\}.$$

Since $L \asymp (nT)^{1/(2m-1)}$, we have

$$\log(L+1) \le C_3 \log(nT \vee d)$$

for some constant $C_3 > 0$. Therefore, by choosing c_λ sufficiently large, there exists a constant $C_4 > 2$ such that

$$P\left(\left\|\sum_{i,t} \boldsymbol{\xi}_{G_k,i,t}\varepsilon_{i,t}\right\|_2 \ge nT\widetilde{\lambda}\right) \le \exp\left\{-C_4 \log(nT \vee d)\right\}. \tag{S3.30}$$

With $L \asymp (nT)^{1/(2m-1)}$, we have

$$L^{1/2-m} \asymp \frac{1}{\sqrt{nT}}.$$

Therefore,

$$2d_0 c_m C_2 L^{1/2-m} \asymp \frac{d_0}{\sqrt{nT}}.$$

Under the condition $d_0 = o(\sqrt{L})$,

$$2d_0 c_m C_2 L^{1/2-m} = o\left\{\sqrt{L}\frac{\log(nT \vee d)}{\sqrt{nT}}\right\}.$$

Thus, for all sufficiently large nT,

$$2d_0 c_m C_2 L^{1/2-m} \le \sqrt{L}\frac{\log(nT \vee d)}{\sqrt{nT}}. \tag{S3.31}$$

Recalling that

$$\widetilde{\lambda} = \lambda - 2d_0 c_m C_2 L^{1/2-m},$$

we may take

$$\lambda = c_1\sqrt{L}\frac{\log(nT \vee d)}{\sqrt{nT}}, \qquad c_1 = c_\lambda + 1.$$

Finally, taking the finite union of (S3.30) over all $k = 1, \ldots, 2d$, we obtain

$$P\left\{\max_{k=1,\ldots,2d}\left\|\frac{1}{nT}\sum_{i,t}\boldsymbol{\xi}_{G_k,i,t}(\varepsilon_{i,t}-e_{i,t})\right\|_2 \le c_1\sqrt{L}\frac{\log(nT\vee d)}{\sqrt{nT}}\right\} \ge 1-\exp\left\{-c_2\log(nT\vee d)\right\}. \tag{S3.32}$$

Note that $\|\boldsymbol{\xi}_{G_0,i,t}\|_2 = 1$. Using the same argument as above, we have

$$P\left\{\left\|\frac{1}{nT}\sum_{i,t}\boldsymbol{\xi}_{G_0,i,t}(\varepsilon_{i,t}-e_{i,t})\right\|_2 \le c_1\frac{\log(nT\vee d)}{\sqrt{nT}}\right\} \ge 1-\exp\left\{-c_2\log(nT\vee d)\right\}. \tag{S3.33}$$

Therefore, there exists a good event that,

$$P(\mathcal{G}_1) \ge 1-\exp\{-c_2\log(nT\vee d)\}.$$

On the event $\mathcal{G}_1$, $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)_\oplus$ satisfies all the constraints in (3.11) and is therefore feasible. This completes the proof. □

Lemma 5. *There exist positive constants c_3 and c_4 and a good event $\mathcal{G}_2$ satisfying*

$$P(\mathcal{G}_2) \ge 1-\exp\left\{-c_4\log(nT\vee d)\right\}$$

such that, on $\mathcal{G}_2$,

$$\left\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_{\max} \le c_3L\log(nT\vee d)\sqrt{\frac{\log(nT\vee d)}{nT}}.$$

Proof. Throughout this proof, $C_1, C_2, \ldots$ denote positive constants that are local to the present proof and are re-indexed from C_1. They are distinct from the lowercase constants $c_1, c_2, \ldots$ appearing in the statements of the results.

The main argument of this proof follows from the proof of Lemma 3 in Shi et al. (2022). The major difference is the matrix norm considered here.

Let $\Theta(l,k)$ be the (l,k)th entry of

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\boldsymbol{\xi}_{i,t}\left(\boldsymbol{\xi}_{i,t} - \gamma \boldsymbol{U}_{i,t+1}\right)^{\top},$$

and let $\Theta_{i,t}(l,k)$ be the corresponding entry of $\boldsymbol{\xi}_{i,t}(\boldsymbol{\xi}_{i,t} - \gamma \boldsymbol{U}_{i,t+1})^{\top}$. The following argument holds for every pair (l,k) and does not depend on the specific values of (l,k). Thus, we suppress (l,k) and write Θ and $\Theta_{i,t}$ for simplicity.

For $1 \le l, k \le 2 + 2dL$, define $\Theta_i = T^{-1}\sum_{t=0}^{T-1}\Theta_{i,t}$. We have

$$|\Theta_{i,t}| \le \|\boldsymbol{\xi}_{i,t}\|_\infty \left(\|\boldsymbol{\xi}_{i,t}\|_\infty + \|\boldsymbol{U}_{i,t+1}\|_\infty\right) \le 2\|\boldsymbol{\Phi}(X_{i,t})\|_\infty^2 \le 2L.$$

Consequently,

$$|\Theta_{i,t} - \mathbb{E}\Theta_{i,t}| \le 4L. \tag{S3.34}$$

We divide the proof into two scenarios: (i) T is bounded and (ii) T diverges.

We first consider the case where T is bounded. By (S3.34),

$$|\Theta_i - \mathbb{E}\Theta_i| \le \frac{1}{T}\sum_{t=0}^{T-1}|\Theta_{i,t} - \mathbb{E}\Theta_{i,t}| \le 4L.$$

Since $\Theta_1, \ldots, \Theta_n$ are independent, Hoeffding's inequality gives

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\{\Theta_i - \mathbb{E}\Theta_i\}\right| > \tau\right) \le 2\exp\left\{-\frac{n\tau^2}{32L^2}\right\}.$$

Taking $\tau = C_1 L\sqrt{\log(nT \vee d)/n}$, we obtain

$$|\Theta - \mathbb{E}\Theta| \le C_1 L\sqrt{\frac{\log(nT \vee d)}{n}}$$

with probability at least $1 - \exp\{-C_2 \log(nT \vee d)\}$. Since T is bounded, the factor $T^{1/2}$ can be absorbed into the constant C_1. Therefore,

$$|\Theta - \mathbb{E}\Theta| \le C_1 L\sqrt{\frac{\log(nT \vee d)}{nT}} \tag{S3.35}$$

with probability at least $1 - \exp\{-C_2 \log(nT \vee d)\}$.

We next consider the case where T diverges. We first derive a tighter high-probability bound for $\Theta_i - \mathbb{E}\Theta_i$ than the deterministic bound of order L. Define

$$\Theta_{0,t}^{*} = \mathbb{E}\left[\Theta_{0,t} \mid X_{0,t}, A_{0,t}\right], \qquad \Theta_{0,t}^{**} = \mathbb{E}\left[\Theta_{0,t}^{*} \mid X_{0,t}\right].$$

We use the decomposition

$$\Theta_{0,t} - \mathbb{E}\Theta_{0,t} = \left(\Theta_{0,t} - \Theta_{0,t}^{*}\right) + \left(\Theta_{0,t}^{*} - \Theta_{0,t}^{**}\right) + \left(\Theta_{0,t}^{**} - \mathbb{E}\Theta_{0,t}\right). \tag{S3.36}$$

Consider the first term in (S3.36). Let

$$\mathcal{F}_t = \boldsymbol{\sigma}\left\{X_{0,0}, A_{0,0}, X_{0,1}, \ldots, X_{0,t}, A_{0,t}, X_{0,t+1}\right\}.$$

Then $\Theta_{0,t} - \Theta^*_{0,t}$ is $\mathcal{F}_t$-measurable and

$$\mathbb{E}\left[\Theta_{0,t} - \Theta^*_{0,t} \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\Theta_{0,t} - \Theta^*_{0,t} \mid X_{0,t}, A_{0,t}\right] \middle| \mathcal{F}_{t-1}\right] = 0.$$

Thus, $\{\Theta_{0,t} - \Theta^*_{0,t}\}_{t\geq 0}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}_{t\geq 0}$. Moreover, by (S3.34), $|\Theta_{0,t} - \Theta^*_{0,t}| \leq 4L$. Azuma's inequality therefore gives

$$\Pr\left(\left|\sum_{t=0}^{T-1}\left(\Theta_{0,t} - \Theta^*_{0,t}\right)\right| > \tau\right) \leq 2\exp\left\{-\frac{\tau^2}{32L^2T}\right\}. \tag{S3.37}$$

Consider the second term in (S3.36). Since the behavior policy is a fixed stationary Markov policy and $X_{0,t}$ is $\mathcal{F}_{t-1}$-measurable,

$$\begin{aligned}\mathbb{E}\left[\Theta^*_{0,t} - \Theta^{**}_{0,t} \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\Theta^*_{0,t} \mid \mathcal{F}_{t-1}\right] - \Theta^{**}_{0,t} \\ &= \mathbb{E}\left[\Theta^*_{0,t} \mid X_{0,t}\right] - \Theta^{**}_{0,t} = 0.\end{aligned}$$

Therefore, $\{\Theta^*_{0,t} - \Theta^{**}_{0,t}\}_{t\geq 0}$ is also a martingale difference sequence with respect to $\{\mathcal{F}_t\}_{t\geq 0}$. Moreover, $|\Theta^*_{0,t} - \Theta^{**}_{0,t}| \leq 4L$. By Azuma's inequality,

$$\Pr\left(\left|\sum_{t=0}^{T-1}\left(\Theta^*_{0,t} - \Theta^{**}_{0,t}\right)\right| > \tau\right) \leq 2\exp\left\{-\frac{\tau^2}{32L^2T}\right\}. \tag{S3.38}$$

We now consider the last term in (S3.36). Let $\widetilde{\Theta}_{0,t} = \Theta^{**}_{0,t} - \mathbb{E}\Theta_{0,t}$. Then $\widetilde{\Theta}_{0,t}$ is a deterministic function of $X_{0,t}$ and $|\widetilde{\Theta}_{0,t}| \leq 4L$. By Assumption (A4), the Markov chain $\{X_{0,t}\}_{t\geq 0}$ is exponentially β-mixing, so that $\beta(q) \leq C_\beta \rho^q$ for some $\rho \in (0,1)$.

Following the proof of Theorem 4.2 in Chen and Christensen (2015), we extend the probability space and construct coupled blocks using Berbee's lemma. Let $M = \lfloor T/q \rfloor$ and define

$$Y_k = \left\{ X_{0,(k-1)q}, \dots, X_{0,kq-1} \right\}, \qquad k = 1, \dots, m.$$

There exist coupled blocks Y_k^* such that Y_k^* and Y_k are identically distributed, $\Pr(Y_k \neq Y_k^*) \leq \beta(q)$, the odd-numbered blocks $Y_1^*, Y_3^*, \dots$ are independent, and the even-numbered blocks $Y_2^*, Y_4^*, \dots$ are independent.

Let $X_{0,t}^*$ denote the coupled process within these blocks and define $\widetilde{\Theta}_{0,t}^* = \widetilde{\Theta}_{0,t}(X_{0,t}^*)$. Let I_o and I_e be the sets of time indices belonging to the odd- and even-numbered complete blocks, respectively, and let $I_r = \{mq, \dots, T-1\}$ be the set of remaining indices. Then $\{0, \dots, T-1\} = I_o \cup I_e \cup I_r$ and $|I_r| < q$.

By the triangle inequality,

$$\begin{aligned}
&\Pr\left(\left| \sum_{t=0}^{T-1} \widetilde{\Theta}_{0,t} \right| > 4\tau \right) \\
&\quad \leq \Pr\left(\left| \sum_{t \in I_r} \widetilde{\Theta}_{0,t} \right| > \tau \right) + \Pr\left(\left| \sum_{t \in I_o} \widetilde{\Theta}_{0,t}^* \right| > \tau \right) + \Pr\left(\left| \sum_{t \in I_e} \widetilde{\Theta}_{0,t}^* \right| > \tau \right) \\
&\qquad + \Pr\left(\sum_{k=1}^{M} \mathbf{1}\{Y_k \neq Y_k^*\} > 0 \right). \qquad \text{(S3.39)}
\end{aligned}$$

When $\tau > 4qL$, the first probability on the right-hand side of (S3.39) is zero. For the odd-numbered blocks, the block sums

$$\sum_{j=0}^{q-1} \widetilde{\Theta}_{0,(k-1)q+j}^*, \qquad k = 1, 3, \dots,$$

are independent, mean-zero random variables bounded in absolute value by $4qL$. Hence, Hoeffding's inequality gives

$$\Pr\left(\left|\sum_{t\in I_o}\widetilde{\Theta}^*_{0,t}\right| > \tau\right) \le 2\exp\left\{-\frac{C_3\tau^2}{qL^2T}\right\}.$$

The same argument applies to the even-numbered blocks. Moreover, by the union bound and Berbee's coupling,

$$\Pr\left(\sum_{k=1}^{M}\mathbf{1}\{Y_k \neq Y_k^*\} > 0\right) \le m\beta(q) \le \frac{T}{q}C_\beta\rho^q.$$

Combining these results, we obtain

$$\Pr\left(\left|\sum_{t=0}^{T-1}\widetilde{\Theta}_{0,t}\right| > 4\tau\right) \le 4\exp\left\{-\frac{C_4\tau^2}{qL^2T}\right\} + \frac{T}{q}C_\beta\rho^q. \tag{S3.40}$$

Combining (S3.37), (S3.38), and (S3.40), we have

$$\begin{aligned}
&\Pr\left(\left|\sum_{t=0}^{T-1}(\Theta_{0,t} - \mathbb{E}\Theta_{0,t})\right| > 6\tau\right)\\
&\le 4\exp\left\{-\frac{C_5\tau^2}{L^2T}\right\} + 4\exp\left\{-\frac{C_6\tau^2}{qL^2T}\right\} + \frac{T}{q}C_\beta\rho^q\\
&\le 8\exp\left\{-\frac{C_7\tau^2}{qL^2T}\right\} + \frac{T}{q}C_\beta\rho^q.
\end{aligned} \tag{S3.41}$$

Set

$$q = \left\lceil -\frac{5\log(nT\vee d)}{\log\rho}\right\rceil$$

and $\tau = \max\{C_1L\log(nT\vee d)\sqrt{T},\, 5qL\}$. When T is sufficiently large, $\tau = C_1L\log(nT\vee d)\sqrt{T}$. Substituting these choices into (S3.41) and absorbing fixed constants into C_1 gives

$$\Pr\left(\left|\sum_{t=0}^{T-1}\left(\Theta_{0,t}-\mathbb{E}\Theta_{0,t}\right)\right|>C_1 L\log(nT\vee d)\sqrt{T}\right)\leq \exp\left\{-C_8\log(nT\vee d)\right\}. \tag{S3.42}$$

Since $\{R_{i,t}, X_{i,t}, A_{i,t}\}_{t\geq 0}$, $i=1,\ldots,n$, are independent copies of $\{R_{0,t}, X_{0,t}, A_{0,t}\}_{t\geq 0}$, the bound (S3.42) holds for every i. Define

$$Z_i=\sum_{t=0}^{T-1}\left(\Theta_{i,t}-\mathbb{E}\Theta_{i,t}\right), \qquad B=C_1 L\log(nT\vee d)\sqrt{T}.$$

Then $Z_1,\ldots,Z_n$ are independent and mean zero, and $\Pr(|Z_i|>B)\leq \exp\{-C_8\log(nT\vee d)\}$. Define

$$\overline{Z}_i=Z_i\mathbf{1}\{|Z_i|\leq B\}-\mathbb{E}\left[Z_i\mathbf{1}\{|Z_i|\leq B\}\right].$$

The variables $\overline{Z}_1,\ldots,\overline{Z}_n$ are independent, mean zero, and satisfy $|\overline{Z}_i|\leq 2B$. Therefore, Hoeffding's inequality gives

$$\Pr\left(\left|\sum_{i=1}^{n}\overline{Z}_i\right|>x\right)\leq 2\exp\left\{-\frac{x^2}{8nB^2}\right\}. \tag{S3.43}$$

Since $\mathbb{E}Z_i=0$ and $|Z_i|\leq 4LT$,

$$\left|\mathbb{E}\left[Z_i\mathbf{1}\{|Z_i|\leq B\}\right]\right|=\left|\mathbb{E}\left[Z_i\mathbf{1}\{|Z_i|>B\}\right]\right|\leq 4LT\exp\left\{-C_8\log(nT\vee d)\right\}.$$

Consequently, after adjusting the constants, the total truncation bias is smaller than $C_1L\{\log(nT\vee d)\}^{3/2}\sqrt{nT}$. Taking $x=c_3L\{\log(nT\vee d)\}^{3/2}\sqrt{nT}$ in (S3.43), and using

$$\Pr\left(\max_{1\leq i\leq n}|Z_i|>B\right)\leq n\exp\left\{-C_8\log(nT\vee d)\right\},$$

we obtain

$$\Pr\left(\left|\sum_{i=1}^{n}\sum_{t=0}^{T-1}(\Theta_{i,t}-\mathbb{E}\Theta_{i,t})\right| > c_3 L\{\log(nT\vee d)\}^{3/2}\sqrt{nT}\right)$$
$$\leq \exp\left\{-(c_3/C_1)^2\log(nT\vee d)/8\right\}. \tag{S3.44}$$

Dividing both sides of the event in (S3.44) by nT gives

$$\Pr\left(|\Theta-\mathbb{E}\Theta| < c_3 L\log(nT\vee d)\sqrt{\frac{\log(nT\vee d)}{nT}}\right) \geq 1-\exp\left\{-C_9\log(nT\vee d)\right\}. \tag{S3.45}$$

Finally, define

$$\mathcal{G}_2 = \left\{\left\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_{\max} < c_3 L\log(nT\vee d)\sqrt{\frac{\log(nT\vee d)}{nT}}\right\}.$$

Taking a finite union over all entries of the matrix and combining (S3.35) and (S3.45), we obtain

$$P(\mathcal{G}_2) \geq 1-\{2+2dL\}^2\exp\left\{-\min(C_2,C_9)\log(nT\vee d)\right\} \geq 1-\exp\left\{-c_4\log(nT\vee d)\right\}.$$

This completes the proof. □

Lemma 6 (Matrix concentration inequalities)**.** *Let $p_1, p_2 \geq 1$ and write $D = p_1 + p_2$.*

- *Let $\{\boldsymbol{Z}_j\}_{j=1}^m$ be independent, mean-zero random matrices of dimension $p_1 \times p_2$. Suppose that $\|\boldsymbol{Z}_j\|_2 \leq R$ almost surely, and define*

$$\nu^2 = \max\left\{\left\|\sum_{j=1}^m \mathbb{E}\left(\boldsymbol{Z}_j\boldsymbol{Z}_j^\top\right)\right\|_2, \left\|\sum_{j=1}^m \mathbb{E}\left(\boldsymbol{Z}_j^\top\boldsymbol{Z}_j\right)\right\|_2\right\}.$$

Then, for any $u > 0$,

$$\Pr\left[\left\|\sum_{j=1}^{m} \boldsymbol{Z}_j\right\|_2 > \sqrt{2\nu^2\{u + \log D\}} + \frac{2R}{3}\{u + \log D\}\right] \leq \exp(-u).$$

In particular, if $\nu^2 \leq mR^2$, then

$$\Pr\left[\left\|\sum_{j=1}^{m} \boldsymbol{Z}_j\right\|_2 > R\sqrt{2m\{u + \log D\}} + \frac{2R}{3}\{u + \log D\}\right] \leq \exp(-u).$$

- *Let $\{\boldsymbol{X}_j, \mathcal{F}_j\}_{j=1}^m$ be an adapted sequence of random matrices of dimension $p_1 \times p_2$ satisfying $\mathbb{E}(\boldsymbol{X}_j \mid \mathcal{F}_{j-1}) = \boldsymbol{0}$. Suppose that $\|\boldsymbol{X}_j\|_2 \leq R_j$ almost surely, where $R_1, \ldots, R_m$ are deterministic. Then, for any $u > 0$,*

$$\Pr\left[\left\|\sum_{j=1}^{m} \boldsymbol{X}_j\right\|_2 > \sqrt{8\{u + \log D\} \sum_{j=1}^{m} R_j^2}\right] \leq \exp(-u).$$

In particular, if $R_j \leq R$ for all j, then

$$\Pr\left[\left\|\sum_{j=1}^{m} \boldsymbol{X}_j\right\|_2 > R\sqrt{8m\{u + \log D\}}\right] \leq \exp(-u).$$

Proof. For a $p_1 \times p_2$ matrix $\boldsymbol{M}$, define its self-adjoint dilation as

$$\mathcal{D}(\boldsymbol{M}) = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{M} \\ \boldsymbol{M}^\top & \boldsymbol{0} \end{pmatrix}.$$

The dilation satisfies

$$\lambda_{\max}\{\mathcal{D}(\boldsymbol{M})\} = \|\boldsymbol{M}\|_2$$

and

$$\mathcal{D}(\boldsymbol{M})^2 = \begin{pmatrix} \boldsymbol{M}\boldsymbol{M}^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}^\top\boldsymbol{M} \end{pmatrix}.$$

For part (i), the rectangular matrix Bernstein inequality (Tropp, 2012, Theorem 1.6) gives, for every $t > 0$,

$$\Pr\left(\left\|\sum_{j=1}^{m} \boldsymbol{Z}_j\right\|_2 > t\right) \le D \exp\left\{-\frac{t^2/2}{\nu^2 + Rt/3}\right\}.$$

Let $x = u + \log D$ and set

$$t = \sqrt{2\nu^2 x} + \frac{2R}{3}x.$$

A direct calculation gives

$$t^2 - 2x\left(\nu^2 + \frac{Rt}{3}\right) = \frac{2R}{3}x\sqrt{2\nu^2 x} \ge 0.$$

It follows that

$$\frac{t^2/2}{\nu^2 + Rt/3} \ge x.$$

Therefore,

$$\Pr\left(\left\|\sum_{j=1}^{m} \boldsymbol{Z}_j\right\|_2 > t\right) \le D \exp(-x) = \exp(-u),$$

which proves the first assertion. The second follows immediately from $\nu^2 \le mR^2$.

For part (ii), consider the self-adjoint matrix differences $\mathcal{D}(\boldsymbol{X}_j)$. They satisfy

$$\mathbb{E}\left[\mathcal{D}(\boldsymbol{X}_j) \mid \mathcal{F}_{j-1}\right] = \mathbf{0}$$

and

$$\mathcal{D}(\boldsymbol{X}_j)^2 \preceq R_j^2 \boldsymbol{I}_D.$$

The matrix Azuma inequality and its rectangular extension (Tropp, 2012, Theorem 7.1 and Remark 7.3) therefore imply that

$$\Pr\left(\left\|\sum_{j=1}^{m}\boldsymbol{X}_j\right\|_2 > t\right) \leq D\exp\left\{-\frac{t^2}{8\sum_{j=1}^{m}R_j^2}\right\}.$$

Taking

$$t = \sqrt{8\{u+\log D\}\sum_{j=1}^{m}R_j^2}$$

yields the desired probability bound. The final assertion follows by using $\sum_{j=1}^{m}R_j^2 \leq mR^2$. $\square$

Lemma 7. *There exist constants $c_5 > 0$ and $c_6 > 0$ and a good event $\mathcal{G}_3$ satisfying*

$$\Pr(\mathcal{G}_3) \geq 1 - \exp\{-c_6\log(nT \vee d)\}$$

such that, on $\mathcal{G}_3$,

$$\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2 \leq c_5 dL\left\{\frac{\log(nT\vee d)}{\sqrt{nT}} + \frac{\{\log(nT\vee d)\}^2}{nT}\right\}.$$

Proof. Throughout this proof, $C_1, C_2, \ldots$ denote positive constants that are local to the present proof and are re-indexed from C_1. They are distinct from the lowercase constants $c_1, c_2, \ldots$ appearing in the statements of the results.

All matrix concentration inequalities used in this proof are equivalent reformulations of those in Tropp (2012); see Lemma 6 for their derivations.

Let $p' = 2 + 2dL$ and

$$\boldsymbol{\Theta}_{i,t} = \boldsymbol{\xi}_{i,t} \left(\boldsymbol{\xi}_{i,t} - \gamma \boldsymbol{U}_{i,t+1} \right)^\top .$$

Then

$$\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \left(\boldsymbol{\Theta}_{i,t} - \mathbb{E} \boldsymbol{\Theta}_{i,t} \right).$$

By Remark 1, $\|\boldsymbol{\xi}_{i,t}\|_2 \leq \sqrt{1 + dc_m^2 L}$ and $\|\boldsymbol{U}_{i,t+1}\|_2 \leq \sqrt{1 + dc_m^2 L}$. Hence,

$$\|\boldsymbol{\Theta}_{i,t}\|_2 \leq (1+\gamma)(1 + dc_m^2 L).$$

Write $R = (1+\gamma)(1 + dc_m^2 L)$. Notice that $R \leq (1+\gamma)(1 + c_m^2) dL$.

Note that $L \asymp (nT)^{1/(2m-1)}$. There exists a constant $C_1 > 0$ such that

$$\log(2p') \leq C_1 \log(nT \vee d).$$

We first consider the case where T is bounded. Let $C_2 > 0$ satisfy $T \leq C_2$, and define

$$\boldsymbol{Z}_i = \frac{1}{T} \sum_{t=0}^{T-1} \left(\boldsymbol{\Theta}_{i,t} - \mathbb{E} \boldsymbol{\Theta}_{i,t} \right).$$

The matrices $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are independent and mean zero, with $\|\boldsymbol{Z}_i\|_2 \leq 2R$. Their matrix variance parameter is bounded by $4nR^2$. Therefore, Lemma 6(i) gives, for every $u > 0$,

$$\Pr\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i \right\|_2 > 2\sqrt{2} R \sqrt{\frac{u + \log(2p')}{n}} + \frac{4R}{3n} \{u + \log(2p')\} \right] \leq \exp(-u).$$

Taking $u = 4\log(nT \vee d)$, the quantity $u + \log(2p')$ is bounded by $(4 + C_1)\log(nT \vee d)$. Since $T \leq C_2$ and $\log(nT \vee d) \geq 1$ for sufficiently large $nT \vee d$, it follows that

$$\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2 \leq C_3 R \frac{\log(nT \vee d)}{\sqrt{nT}}$$

with probability at least $1 - (nT \vee d)^{-4}$, where

$$C_3 = 2\sqrt{2C_2(4 + C_1)} + \frac{4C_2(4 + C_1)}{3}.$$

We next consider the case where T diverges. Define $\boldsymbol{\Theta}_{i,t}^* = \mathbb{E}(\boldsymbol{\Theta}_{i,t} \mid X_{i,t}, A_{i,t})$ and $\boldsymbol{\Theta}_{i,t}^{**} = \mathbb{E}(\boldsymbol{\Theta}_{i,t}^* \mid X_{i,t})$. Then

$$\boldsymbol{\Theta}_{i,t} - \mathbb{E}\boldsymbol{\Theta}_{i,t} = \left(\boldsymbol{\Theta}_{i,t} - \boldsymbol{\Theta}_{i,t}^*\right) + \left(\boldsymbol{\Theta}_{i,t}^* - \boldsymbol{\Theta}_{i,t}^{**}\right) + \left(\boldsymbol{\Theta}_{i,t}^{**} - \mathbb{E}\boldsymbol{\Theta}_{i,t}\right).$$

Accordingly, write

$$\sum_{i=1}^{n}\sum_{t=0}^{T-1}(\boldsymbol{\Theta}_{i,t} - \mathbb{E}\boldsymbol{\Theta}_{i,t}) = \boldsymbol{S}_1 + \boldsymbol{S}_2 + \boldsymbol{S}_3,$$

where

$$\boldsymbol{S}_1 = \sum_{i,t}\left(\boldsymbol{\Theta}_{i,t} - \boldsymbol{\Theta}_{i,t}^*\right), \qquad \boldsymbol{S}_2 = \sum_{i,t}\left(\boldsymbol{\Theta}_{i,t}^* - \boldsymbol{\Theta}_{i,t}^{**}\right),$$

and

$$\boldsymbol{S}_3 = \sum_{i,t}\left(\boldsymbol{\Theta}_{i,t}^{**} - \mathbb{E}\boldsymbol{\Theta}_{i,t}\right).$$

Order the trajectories by $i = 1, \ldots, n$. For each i, let $\mathcal{X}_{<i}$ be the sigma-field generated by the complete trajectories $1, \ldots, i-1$. Define the

within-trajectory filtration by

$$\begin{aligned}\mathcal{H}_{i,2t} &= \sigma\left\{\mathcal{X}_{<i}, X_{i,0}, A_{i,0}, X_{i,1}, \dots, A_{i,t-1}, X_{i,t}\right\},\\ \mathcal{H}_{i,2t+1} &= \mathcal{H}_{i,2t} \vee \sigma(A_{i,t}),\\ \mathcal{H}_{i,2t+2} &= \mathcal{H}_{i,2t+1} \vee \sigma(X_{i,t+1}), \qquad t = 0, \dots, T-1.\end{aligned}$$

These sigma-fields, ordered first by i and then by the second index, form a filtration.

For $\boldsymbol{S}_1$, place the increment $\boldsymbol{\Theta}_{i,t} - \boldsymbol{\Theta}^*_{i,t}$ between $\mathcal{H}_{i,2t+1}$ and $\mathcal{H}_{i,2t+2}$, and place a zero increment between $\mathcal{H}_{i,2t}$ and $\mathcal{H}_{i,2t+1}$. The increment is $\mathcal{H}_{i,2t+2}$-measurable. Moreover, by the Markov property and the independence of the trajectories,

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\Theta}_{i,t} - \boldsymbol{\Theta}^*_{i,t} \mid \mathcal{H}_{i,2t+1}\right] &= \mathbb{E}\left[\boldsymbol{\Theta}_{i,t} \mid \mathcal{H}_{i,2t+1}\right] - \boldsymbol{\Theta}^*_{i,t}\\ &= \mathbb{E}\left[\boldsymbol{\Theta}_{i,t} \mid X_{i,t}, A_{i,t}\right] - \boldsymbol{\Theta}^*_{i,t}\\ &= \mathbf{0}.\end{aligned}$$

Thus, the summands in $\boldsymbol{S}_1$, together with the inserted zero increments, form a matrix martingale difference sequence.

For $\boldsymbol{S}_2$, place the increment $\boldsymbol{\Theta}^*_{i,t} - \boldsymbol{\Theta}^{**}_{i,t}$ between $\mathcal{H}_{i,2t}$ and $\mathcal{H}_{i,2t+1}$, and place a zero increment between $\mathcal{H}_{i,2t+1}$ and $\mathcal{H}_{i,2t+2}$. Since the behavior policy is a fixed stationary Markov policy,

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\Theta}^*_{i,t} - \boldsymbol{\Theta}^{**}_{i,t} \mid \mathcal{H}_{i,2t}\right] &= \mathbb{E}\left[\boldsymbol{\Theta}^*_{i,t} \mid \mathcal{H}_{i,2t}\right] - \boldsymbol{\Theta}^{**}_{i,t} \\ &= \mathbb{E}\left[\boldsymbol{\Theta}^*_{i,t} \mid X_{i,t}\right] - \boldsymbol{\Theta}^{**}_{i,t} \\ &= \mathbf{0}.\end{aligned}$$

Therefore, the summands in $\boldsymbol{S}_2$, together with the inserted zero increments, also form a matrix martingale difference sequence. Because conditional expectations preserve the almost-sure bound $\|\boldsymbol{\Theta}_{i,t}\|_2 \le R$, every increment in $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ has operator norm at most $2R$. Lemma 6(ii) therefore yields, for $j = 1, 2$ and every $u > 0$,

$$\Pr\left[\|\boldsymbol{S}_j\|_2 > 4\sqrt{2}R\sqrt{nT\{u + \log(2p')\}}\right] \le \exp(-u).$$

Taking $u = 4\log(nT \vee d)$ and combining the two bounds gives

$$\frac{\|\boldsymbol{S}_1\|_2 + \|\boldsymbol{S}_2\|_2}{nT} \le 8\sqrt{2(4 + C_1)}\, R\sqrt{\frac{\log(nT \vee d)}{nT}} \tag{S3.46}$$

with probability at least $1 - 2(nT \vee d)^{-4}$.

It remains to control $\boldsymbol{S}_3$. Let $\widetilde{\boldsymbol{\Theta}}_{i,t} = \boldsymbol{\Theta}^{**}_{i,t} - \mathbb{E}\boldsymbol{\Theta}_{i,t}$. This is a deterministic matrix-valued function of $X_{i,t}$, satisfies $\mathbb{E}\widetilde{\boldsymbol{\Theta}}_{i,t} = \mathbf{0}$, and has operator norm at most $2R$.

By Assumption (A4), the state process is exponentially β-mixing, with $\beta(q) \le C_\beta \rho^q$ for some $C_\beta > 0$ and $\rho \in (0, 1)$. Set

$$q = \left\lceil -\frac{5\log(nT \vee d)}{\log \rho} \right\rceil$$

and $m = \lceil T/q \rceil$. By the growth conditions, $q \le T$ for all sufficiently large $nT \vee d$. Partition $\{0, \dots, T-1\}$ into consecutive blocks $\mathcal{I}_1, \dots, \mathcal{I}_m$, each containing q time points except possibly the last block.

For each i and k, let $Y_{i,k} = (X_{i,t} : t \in \mathcal{I}_k)$ and define

$$\boldsymbol{W}_{i,k} = \sum_{t \in \mathcal{I}_k} \widetilde{\boldsymbol{\Theta}}_{i,t}.$$

Since $\widetilde{\boldsymbol{\Theta}}_{i,t}$ is a deterministic function of $X_{i,t}$, the matrix $\boldsymbol{W}_{i,k}$ is a measurable function of $Y_{i,k}$, and hence $\sigma(\boldsymbol{W}_{i,k}) \subseteq \sigma(Y_{i,k})$.

Fix i and first consider the odd block sums $\boldsymbol{W}_{i,1}, \boldsymbol{W}_{i,3}, \dots$. Regard each $p \times p$ random matrix as a random element of $\mathbb{R}^{p^2}$ through vectorization. Berbee's coupling lemma can then be applied recursively to this sequence. On an enlarged probability space, it produces random matrices $\boldsymbol{W}^*_{i,1}, \boldsymbol{W}^*_{i,3}, \dots$ such that each $\boldsymbol{W}^*_{i,k}$ has the same distribution as $\boldsymbol{W}_{i,k}$, the coupled odd block sums are mutually independent, and

$$\Pr\left(\boldsymbol{W}^*_{i,k} \neq \boldsymbol{W}_{i,k}\right) \le \beta(q)$$

for every odd k. Indeed, the preceding odd blocks and the current odd block are separated by an intervening block of length q, and

$$\beta\left(\sigma\{\boldsymbol{W}_{i,1}, \dots, \boldsymbol{W}_{i,k-2}\}, \sigma(\boldsymbol{W}_{i,k})\right) \le \beta(q),$$

because the sigma-fields generated by the matrix block sums are contained in those generated by the corresponding state blocks.

Applying the same construction to the even block sums produces $\boldsymbol{W}^*_{i,2}, \boldsymbol{W}^*_{i,4}, \ldots$, which are mutually independent and satisfy the same marginal-distribution and mismatch-probability properties. Since the n trajectories are independent, the auxiliary variables in these coupling constructions may be chosen independently across i. Consequently,

$$\left\{\boldsymbol{W}^*_{i,k} : 1 \leq i \leq n,\ k \text{ odd}\right\}$$

is a mutually independent family, and the same is true for the even-numbered coupled block sums.

Let $\mathcal{E}_{\text{cpl}}$ be the event on which $\boldsymbol{W}^*_{i,k} = \boldsymbol{W}_{i,k}$ for every i and k. Since $m \leq 2T/q$,

$$\begin{aligned}\Pr(\mathcal{E}^c_{\text{cpl}}) &\leq nm\beta(q) \\ &\leq \frac{2nT}{q} C_\beta \rho^q \\ &\leq 2C_\beta (nT \vee d)^{-4},\end{aligned}$$

where the last inequality follows from $\rho^q \leq (nT \vee d)^{-5}$.

Each $\boldsymbol{W}^*_{i,k}$ is mean zero and satisfies $\|\boldsymbol{W}^*_{i,k}\|_2 \leq 2qR$. For either parity, there are at most $2nT/q$ such block sums. Therefore, the corresponding matrix variance parameter is bounded by

$$\frac{2nT}{q}(2qR)^2 = 8nTqR^2.$$

Lemma 6(i) gives, for every $u > 0$,

$$\Pr\left[\left\|\sum_{i=1}^{n}\sum_{\substack{1\le k\le m\\ k \text{ odd}}} \boldsymbol{W}^*_{i,k}\right\|_2 > 4R\sqrt{nTq\{u+\log(2p')\}} + \frac{4qR}{3}\{u+\log(2p')\}\right] \le \exp(-u),$$

and the same bound holds for the even block sums.

On $\mathcal{E}_{\mathrm{cpl}}$, $\boldsymbol{S}_3$ is the sum of the coupled odd and even block sums. Taking $u = 4\log(nT \vee d)$ and combining the two parity bounds yields

$$\frac{\|\boldsymbol{S}_3\|_2}{nT} \le 8R\sqrt{\frac{q(4+C_1)\log(nT\vee d)}{nT}} + \frac{8qR(4+C_1)\log(nT\vee d)}{3nT} \quad \text{(S3.47)}$$

except on an event of probability at most

$$2(nT\vee d)^{-4} + 2C_\beta(nT\vee d)^{-4}.$$

Let

$$C_4 = -\frac{5}{\log\rho} + 1.$$

Since $q \le C_4 \log(nT \vee d)$ for all sufficiently large $nT \vee d$, (S3.46) and (S3.47) imply

$$\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2 \le R\left[C_5\frac{\log(nT\vee d)}{\sqrt{nT}} + C_6\frac{\{\log(nT\vee d)\}^2}{nT}\right].$$

The total failure probability in the diverging-T case is at most $(4 + 2C_\beta)(nT\vee d)^{-4}$, which is bounded by $\exp\{-3\log(nT\vee d)\}$ for all sufficiently large $nT\vee d$. The same probability bound holds in the bounded-T case after increasing the threshold if necessary.

Define

$$\mathcal{G}_3 = \left\{ \left\| \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|_2 \leq c_5 dL \left[\frac{\log(nT \vee d)}{\sqrt{nT}} + \frac{\{\log(nT \vee d)\}^2}{nT} \right] \right\}.$$

Finally, since $R \leq (1+\gamma)(1+c_m^2)dL$, the result follows by taking

$$c_6 = 3$$

and any c_5 satisfying

$$c_5 \geq (1+\gamma)(1+c_m^2)\max\{C_3, C_5, C_6\}.$$

Thus,

$$\Pr(\mathcal{G}_3) \geq 1 - \exp\{-c_6 \log(nT \vee d)\}.$$

This completes the proof. $\square$

S4 Proofs of Results in Section 4

S4.1 Proof of Theorem 5

Proof. Throughout this proof, $C_1, C_2, \ldots$ denote positive constants that are local to the present proof and are re-indexed from C_1. They are distinct from the lowercase constants $c_1, c_2, \ldots$ appearing in the statements of the results.

Recall that $\widehat{\boldsymbol{\kappa}} = (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\oplus}$. Define $S = S_0 \cup \{d + k : k \in S_1\}$ and $\mathcal{S} = G_0 \cup (\cup_{k \in S} G_k)$. By construction, $|S| \leq 2d_0$. For convenience, let

$\widehat{\boldsymbol{\kappa}}_{G_0} = \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$, and, for $k = 1, \ldots, 2d$, let $\widehat{\boldsymbol{\kappa}}_{G_k} = \widehat{\boldsymbol{\beta}}_{G_k} - \boldsymbol{\beta}^*_{G_k}$.

By the definition of the estimator, $\sum_{k=1}^{2d} \|\widehat{\boldsymbol{\beta}}_{G_k}\|_2 \le \sum_{k=1}^{2d} \|\boldsymbol{\beta}^*_{G_k}\|_2$. Moreover, $\boldsymbol{\beta}^*_{G_k} = \mathbf{0}$ for $k \notin S$. Hence,

$$\begin{aligned}\sum_{k=1}^{2d} \|\boldsymbol{\beta}^*_{G_k}\|_2 &\ge \sum_{k\in S} \left\|\boldsymbol{\beta}^*_{G_k} + \widehat{\boldsymbol{\kappa}}_{G_k}\right\|_2 + \sum_{k\notin S} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2 \\ &\ge \sum_{k\in S} \|\boldsymbol{\beta}^*_{G_k}\|_2 - \sum_{k\in S} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2 + \sum_{k\notin S} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2.\end{aligned}$$

It follows that

$$\sum_{k\notin S} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2 \le \sum_{k\in S} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2.$$

Consequently,

$$\begin{aligned}\sum_{k=0}^{2d} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2 &\le \|\widehat{\boldsymbol{\kappa}}_{G_0}\|_2 + 2\sum_{k\in S} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2 \\ &\le 2 \sum_{k\in S\cup\{0\}} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2 \\ &\le 2\sqrt{2d_0+1}\, \|\widehat{\boldsymbol{\kappa}}_{\mathcal{S}}\|_2 \\ &\le 2\sqrt{2d_0+1}\, \|\widehat{\boldsymbol{\kappa}}\|_2.\end{aligned} \tag{S4.48}$$

Since each group contains at most L coordinates, $\|\widehat{\boldsymbol{\kappa}}_{G_k}\|_1 \le \sqrt{L}\|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2$. Therefore,

$$\|\widehat{\boldsymbol{\kappa}}\|_1 = \sum_{k=0}^{2d} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_1 \le \sqrt{L} \sum_{k=0}^{2d} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2.$$

Combining this inequality with (S4.48) gives

$$\|\widehat{\boldsymbol{\kappa}}\|_1 \le 2\sqrt{L(2d_0+1)}\|\widehat{\boldsymbol{\kappa}}\|_2. \tag{S4.49}$$

We next bound the quadratic form $\widehat{\boldsymbol{\kappa}}^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\kappa}}$. By Lemma 4, the oracle coefficient vector is feasible on the event $\mathcal{G}_1$. Since the estimated coefficient vector is feasible by construction, on $\mathcal{G}_1$, the triangle inequality gives

$$\left\| \left[\widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\kappa}} \right]_{G_k} \right\|_2 \leq 2\lambda, \qquad k = 0, \dots, 2d.$$

It follows that, on $\mathcal{G}_1$,

$$\begin{aligned}
\left| \widehat{\boldsymbol{\kappa}}^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\kappa}} \right| &= \left| \sum_{k=0}^{2d} \widehat{\boldsymbol{\kappa}}_{G_k}^\top \left[\widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\kappa}} \right]_{G_k} \right| \\
&\leq \sum_{k=0}^{2d} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2 \left\| \left[\widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\kappa}} \right]_{G_k} \right\|_2 \\
&\leq 2\lambda \sum_{k=0}^{2d} \|\widehat{\boldsymbol{\kappa}}_{G_k}\|_2 \\
&\leq 4\lambda \sqrt{2d_0 + 1}\, \|\widehat{\boldsymbol{\kappa}}\|_2.
\end{aligned}$$

On the other hand, Lemma 3 implies

$$\left| \widehat{\boldsymbol{\kappa}}^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{\kappa}} \right| \geq \frac{\bar{c}}{2} \|\widehat{\boldsymbol{\kappa}}\|_2^2.$$

Therefore,

$$\begin{aligned}
\left| \widehat{\boldsymbol{\kappa}}^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\kappa}} \right| &\geq \left| \widehat{\boldsymbol{\kappa}}^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{\kappa}} \right| - \left| \widehat{\boldsymbol{\kappa}}^\top (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \widehat{\boldsymbol{\kappa}} \right| \\
&\geq \frac{\bar{c}}{2} \|\widehat{\boldsymbol{\kappa}}\|_2^2 - \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \|\widehat{\boldsymbol{\kappa}}\|_1^2.
\end{aligned}$$

By (S4.49) and Lemma 5, on the event $\mathcal{G}_2$,

$$\left| \widehat{\boldsymbol{\kappa}}^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\kappa}} \right| \geq \left[\frac{\bar{c}}{2} - 4(2d_0 + 1) c_3 L^2 \log(nT \vee d) \sqrt{\frac{\log(nT \vee d)}{nT}} \right] \|\widehat{\boldsymbol{\kappa}}\|_2^2.$$

Under $L \asymp (nT)^{1/(2m-1)}$ with $m \geq 3$, for all sufficiently large nT,

$$\frac{\bar{c}}{2} - 4(2d_0 + 1) c_3 L^2 \log(nT \vee d) \sqrt{\frac{\log(nT \vee d)}{nT}} \geq \frac{\bar{c}}{4}. \tag{S4.50}$$

Combining the upper and lower bounds for the quadratic form, on the event $\mathcal{G}_1 \cap \mathcal{G}_2$, yields

$$\frac{\bar{c}}{4}\|\widehat{\boldsymbol{\kappa}}\|_2^2 \leq 4\lambda\sqrt{2d_0+1}\,\|\widehat{\boldsymbol{\kappa}}\|_2.$$

Thus, $\|\widehat{\boldsymbol{\kappa}}\|_2 \lesssim \lambda\sqrt{d_0}$. Since $\lambda \asymp \sqrt{L}\log(nT\vee d)/\sqrt{nT}$, we obtain

$$\|\widehat{\boldsymbol{\kappa}}\|_2 \lesssim \frac{\log(nT\vee d)}{\sqrt{nT}}\sqrt{d_0 L}.$$

Equation (S4.49) further gives

$$\|\widehat{\boldsymbol{\kappa}}\|_1 \lesssim \frac{\log(nT\vee d)}{\sqrt{nT}}d_0 L.$$

These bounds hold on the event $\mathcal{G}_1 \cap \mathcal{G}_2$, where

$$P(\mathcal{G}_1 \cap \mathcal{G}_2) \geq 1 - \exp\{-c_2\log(nT\vee d)\} - \exp\{-c_4\log(nT\vee d)\}.$$

We finally consider the special case where $d = O(1)$. By Lemma 7, on the event $\mathcal{G}_3$,

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 \leq c_5 dL\left\{\frac{\log(nT\vee d)}{\sqrt{nT}} + \frac{\{\log(nT\vee d)\}^2}{nT}\right\}.$$

Since

$$\left|\widehat{\boldsymbol{\kappa}}^\top(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\boldsymbol{\kappa}}\right| \leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2\|\widehat{\boldsymbol{\kappa}}\|_2^2,$$

condition (S4.50) can be replaced by

$$\frac{\bar{c}}{2} - c_5 dL\left\{\frac{\log(nT\vee d)}{\sqrt{nT}} + \frac{\{\log(nT\vee d)\}^2}{nT}\right\} \geq \frac{\bar{c}}{4}.$$

Under $d = O(1)$, it is sufficient that

$$L\left\{\frac{\log(nT \vee d)}{\sqrt{nT}} + \frac{\{\log(nT \vee d)\}^2}{nT}\right\} = o(1).$$

Recall that the feasibility requires

$$\frac{(nT)^{1/(2m)}}{L} = o(1).$$

In particular, if $L \asymp (nT)^{1/(2m-1)}$, this condition holds whenever $2m-1 > 2$.

On the event $\mathcal{G}_1 \cap \mathcal{G}_3$, the same argument then gives the preceding ℓ_2- and ℓ_1-error bounds, where

$$P(\mathcal{G}_1 \cap \mathcal{G}_3) \geq 1 - \exp\{-c_2 \log(nT \vee d)\} - \exp\{-c_6 \log(nT \vee d)\}.$$

$\square$

S4.2 Proof of Theorem 6

Proof. Throughout this proof, $C_1, C_2, \ldots$ denote positive constants that are local to the present proof and are re-indexed from C_1. They are distinct from the lowercase constants $c_1, c_2, \ldots$ appearing in the statements of the results.

It follows from the triangle inequality that

$$|\hat{V}^{\pi}(\tilde{x}) - V^{\pi}(\tilde{x})| \leq |\hat{V}^{\pi}(\tilde{x}) - V^{\pi*}(\tilde{x})| + |V^{\pi*}(\tilde{x}) - V^{\pi}(\tilde{x})|.$$

From inequality (3.5), we have

$$|V^{\pi}(\tilde{x}) - V^{\pi*}(\tilde{x})| \leq d_0 C_1 L^{-m},$$

and

$$\mathbb{E}_{\tilde{x}}|V^{\pi}(\tilde{x}) - V^{\pi*}(\tilde{x})| \leq d_0 C_1 L^{-m}.$$

Thus, the remaining proof focuses on the estimation error

$$|\hat{V}^{\pi}(\tilde{x}) - V^{\pi*}(\tilde{x})|.$$

We have

$$\begin{aligned} |\hat{V}^{\pi}(\tilde{x}) - V^{\pi*}(\tilde{x})| &= |U^T(\tilde{x})\hat{\boldsymbol{\kappa}}| \\ &= \left|\sum_{k=0}^{2d} U_{G_k}^T(\tilde{x})\hat{\boldsymbol{\kappa}}_{G_k}\right| \\ &\leq \sum_{k=0}^{2d} \left|U_{G_k}^T(\tilde{x})\hat{\boldsymbol{\kappa}}_{G_k}\right| \\ &\leq \sum_{k=0}^{2d} \|U_{G_k}^T(\tilde{x})\|_2 \|\hat{\boldsymbol{\kappa}}_{G_k}\|_2. \end{aligned} \tag{S4.51}$$

Note that

$$\|U_{G_0}^T(\tilde{x})\|_2 = \left\{\pi(0 \mid \tilde{x})^2 + \pi(1 \mid \tilde{x})^2\right\}^{1/2} \leq 1.$$

Moreover, for every $1 \leq s \leq d$,

$$\|U_{G_s}^T(\tilde{x})\|_2 = \pi(0 \mid \tilde{x})\|\Phi_s(\tilde{x}_s)\|_2 \leq \|\Phi_s(\tilde{x}_s)\|_2,$$

and

$$\|U_{G_{s+d}}^T(\tilde{x})\|_2 = \pi(1 \mid \tilde{x})\|\Phi_s(\tilde{x}_s)\|_2 \leq \|\Phi_s(\tilde{x}_s)\|_2.$$

Combined with Remark 1, we have

$$\begin{aligned} \sum_{k=0}^{2d} \|U_{G_k}^T(\tilde{x})\|_2 \|\hat{\boldsymbol{\kappa}}_{G_k}\|_2 &\leq c_m\sqrt{L}\sum_{k=0}^{2d} \|\hat{\boldsymbol{\kappa}}_{G_k}\|_2 \\ &\leq \sqrt{4c_m^2 L(2d_0+1)}\|\hat{\boldsymbol{\kappa}}\|_2. \end{aligned} \tag{S4.52}$$

The second inequality follows from equation (S4.48). Combined with Theorem 5, on the event $\mathcal{G}_1 \cap \mathcal{G}_2$, we have

$$\sup_{\tilde{x} \in \mathcal{X}^d} |\hat{V}^{\pi}(\tilde{x}) - V^{\pi *}(\tilde{x})| \lesssim d_0 L \frac{\log(nT \vee d)}{\sqrt{nT}}. \tag{S4.53}$$

Under the choice $L = (nT)^{1/(2m-1)}$, the functional approximation error satisfies

$$|V^{\pi}(\tilde{x}) - V^{\pi *}(\tilde{x})| \leq d_0 C_1 L^{-m} = d_0 C_1 (\sqrt{nT})^{-2m/(2m-1)} \lesssim d_0 (\sqrt{nT})^{-1}.$$

Therefore, on the same event,

$$\begin{aligned} \sup_{\tilde{x} \in \mathcal{X}^d} |\hat{V}^{\pi}(\tilde{x}) - V^{\pi}(\tilde{x})| &\lesssim d_0 L \frac{\log(nT \vee d)}{\sqrt{nT}} + d_0 (\sqrt{nT})^{-1} \\ &\lesssim d_0 L \frac{\log(nT \vee d)}{\sqrt{nT}}. \end{aligned} \tag{S4.54}$$

Next, we establish the bound for the generalized test error. Assume that the density $\mu_{\tilde{x}}$ is uniformly bounded by $\tilde{c}_\mu$. By definition,

$$\begin{aligned} &\left| \mathbb{E}_{\tilde{x}} \hat{V}^{\pi}(\tilde{x}) - \mathbb{E}_{\tilde{x}} V^{\pi *}(\tilde{x}) \right| \\ &\quad = \left| \int_{\tilde{x} \in \mathcal{X}^d} U^T(\tilde{x}) \hat{\boldsymbol{\kappa}} \mu_{\tilde{x}} \, d\tilde{x} \right| \\ &\quad \leq \tilde{c}_\mu \sum_{k=0}^{2d} \int_{\tilde{x} \in \mathcal{X}^d} \left| U_{G_k}^T(\tilde{x}) \hat{\boldsymbol{\kappa}}_{G_k} \right| \, d\tilde{x}. \end{aligned} \tag{S4.55}$$

It follows from the Cauchy–Schwarz inequality that

$$\begin{aligned}
&\int_{\tilde{x}\in\mathcal{X}^d} \left|U_{G_k}^T(\tilde{x})\hat{\boldsymbol{\kappa}}_{G_k}\right| \, d\tilde{x} \\
&\quad = \int_{\tilde{x}\in\mathcal{X}^d} \left\{\hat{\boldsymbol{\kappa}}_{G_k}^T U_{G_k}(\tilde{x}) U_{G_k}^T(\tilde{x})\hat{\boldsymbol{\kappa}}_{G_k}\right\}^{1/2} \, d\tilde{x} \\
&\quad \lesssim \left\{\int_{\tilde{x}\in\mathcal{X}^d} \hat{\boldsymbol{\kappa}}_{G_k}^T U_{G_k}(\tilde{x}) U_{G_k}^T(\tilde{x})\hat{\boldsymbol{\kappa}}_{G_k} \, d\tilde{x}\right\}^{1/2} \\
&\quad = \left[\hat{\boldsymbol{\kappa}}_{G_k}^T \left\{\int_{\tilde{x}\in\mathcal{X}^d} U_{G_k}(\tilde{x}) U_{G_k}^T(\tilde{x}) \, d\tilde{x}\right\} \hat{\boldsymbol{\kappa}}_{G_k}\right]^{1/2},
\end{aligned} \tag{S4.56}$$

where the first inequality also uses the boundedness of $\mathcal{X}^d$.

When π is a stationary deterministic policy, we have $0 \le \pi(a \mid \tilde{x}) \le 1$. For every $1 \le s \le d$, the action-specific group definitions give

$$U_{G_s}(\tilde{x}) = \Phi_s(\tilde{x}_s)\pi(0 \mid \tilde{x}), \qquad U_{G_{s+d}}(\tilde{x}) = \Phi_s(\tilde{x}_s)\pi(1 \mid \tilde{x}).$$

Therefore,

$$U_{G_s}(\tilde{x})U_{G_s}^T(\tilde{x}) = \Phi_s(\tilde{x}_s)\Phi_s^T(\tilde{x}_s)\pi(0 \mid \tilde{x})^2,$$

and

$$U_{G_{s+d}}(\tilde{x})U_{G_{s+d}}^T(\tilde{x}) = \Phi_s(\tilde{x}_s)\Phi_s^T(\tilde{x}_s)\pi(1 \mid \tilde{x})^2.$$

Since $0 \le \pi(a \mid \tilde{x})^2 \le 1$, we have, in the positive-semidefinite order,

$$\int_{\mathcal{X}^d} U_{G_s}(\tilde{x})U_{G_s}^T(\tilde{x}) \, d\tilde{x} \preceq \int_{\mathcal{X}^d} \Phi_s(\tilde{x}_s)\Phi_s^T(\tilde{x}_s) \, d\tilde{x},$$

and

$$\int_{\mathcal{X}^d} U_{G_{s+d}}(\tilde{x})U_{G_{s+d}}^T(\tilde{x}) \, d\tilde{x} \preceq \int_{\mathcal{X}^d} \Phi_s(\tilde{x}_s)\Phi_s^T(\tilde{x}_s) \, d\tilde{x}.$$

Because $\mathcal{X} = [0,1]$, integration over the remaining state coordinates gives

$$\int_{\mathcal{X}^d} \Phi_s(\tilde{x}_s)\Phi_s^T(\tilde{x}_s)\, d\tilde{x} = \int_{\mathcal{X}} \Phi_s(z)\Phi_s^T(z)\, dz.$$

Hence, Remark 1 implies

$$\lambda_{\max}\left\{\int_{\mathcal{X}^d} U_{G_k}(\tilde{x})U_{G_k}^T(\tilde{x})\, d\tilde{x}\right\} \lesssim 1, \qquad 1 \leq k \leq 2d.$$

For the intercept group,

$$U_{G_0}(\tilde{x}) = \{\pi(0 \mid \tilde{x}), \pi(1 \mid \tilde{x})\}^T,$$

and therefore

$$\begin{aligned}\lambda_{\max}\left\{\int_{\mathcal{X}^d} U_{G_0}(\tilde{x})U_{G_0}^T(\tilde{x})\, d\tilde{x}\right\} &\leq \int_{\mathcal{X}^d} \|U_{G_0}(\tilde{x})\|_2^2\, d\tilde{x} \\ &\leq 1.\end{aligned}$$

Consequently,

$$\lambda_{\max}\left\{\int_{\mathcal{X}^d} U_{G_k}(\tilde{x})U_{G_k}^T(\tilde{x})\, d\tilde{x}\right\} \lesssim 1, \qquad 0 \leq k \leq 2d. \tag{S4.57}$$

Combining (S4.56) and (S4.57), we obtain

$$\left|\mathbb{E}_{\tilde{x}}\hat{V}^{\pi}(\tilde{x}) - \mathbb{E}_{\tilde{x}}V^{\pi*}(\tilde{x})\right| \lesssim \sum_{k=0}^{2d} \|\hat{\boldsymbol{\kappa}}_{G_k}\|_2. \tag{S4.58}$$

Combining this result with Theorem 5, on the event $\mathcal{G}_1 \cap \mathcal{G}_2$, we obtain

$$\left|\mathbb{E}_{\tilde{x}}\hat{V}^{\pi}(\tilde{x}) - \mathbb{E}_{\tilde{x}}V^{\pi*}(\tilde{x})\right| \lesssim d_0\sqrt{L}\frac{\log(nT \vee d)}{\sqrt{nT}}. \tag{S4.59}$$

Moreover, by inequality (3.5),

$$\begin{aligned}\left|\mathbb{E}_{\tilde{x}}\left[V^{\pi*}(\tilde{x}) - V^{\pi}(\tilde{x})\right]\right| &\leq \mathbb{E}_{\tilde{x}}\left|V^{\pi*}(\tilde{x}) - V^{\pi}(\tilde{x})\right| \\ &\leq d_0 C_1 L^{-m}.\end{aligned} \tag{S4.60}$$

Therefore, by the triangle inequality, on $\mathcal{G}_1 \cap \mathcal{G}_2$,

$$\begin{aligned}\left|\mathbb{E}_{\tilde{x}}\hat{V}^{\pi}(\tilde{x}) - \mathbb{E}_{\tilde{x}}V^{\pi}(\tilde{x})\right| &\leq \left|\mathbb{E}_{\tilde{x}}\hat{V}^{\pi}(\tilde{x}) - \mathbb{E}_{\tilde{x}}V^{\pi*}(\tilde{x})\right| \\ &\quad + \left|\mathbb{E}_{\tilde{x}}V^{\pi*}(\tilde{x}) - \mathbb{E}_{\tilde{x}}V^{\pi}(\tilde{x})\right| \qquad \text{(S4.61)} \\ &\lesssim d_0\sqrt{L}\frac{\log(nT \vee d)}{\sqrt{nT}} + d_0 L^{-m}.\end{aligned}$$

Under the choice $L = (nT)^{1/(2m-1)}$, the approximation error $d_0 L^{-m}$ is dominated by the preceding estimation error. Hence, on the same event,

$$\left|\mathbb{E}_{\tilde{x}}\hat{V}^{\pi}(\tilde{x}) - \mathbb{E}_{\tilde{x}}V^{\pi}(\tilde{x})\right| \lesssim d_0\sqrt{L}\frac{\log(nT \vee d)}{\sqrt{nT}}. \qquad \text{(S4.62)}$$

Finally,

$$P(\mathcal{G}_1 \cap \mathcal{G}_2) \geq 1 - \exp\{-c_2 \log(nT \vee d)\} - \exp\{-c_4 \log(nT \vee d)\}.$$

This completes the proof. □

S4.3 Proof of Theorem 7

Proof. Throughout this proof, $C_1, C_2, \ldots$ denote positive constants that are local to the present proof and are re-indexed from C_1. They are distinct from the lowercase constants $c_1, c_2, \ldots$ appearing in the statements of the results.

Denote

$$q_{s,a}^{\pi*}(x_s) = \Phi_s^T(x_s)\boldsymbol{\beta}_{s,a}^*.$$

In this proof, we omit the superscript π from q, $\widehat{q}$, and q^* for simplicity. The expectation is taken with respect to the reference distribution $x \sim \mathbb{G}$.

Consider an active feature x_s, where $s \in \mathcal{K}$. Under the condition in Theorem 7, there exists an action $a \in \{0, 1\}$ such that

$$\mathbb{E}\left\{q_{s,a}^2(x_s)\right\} \geq v.$$

By (3.5),

$$\left|q_{s,a}(x_s) - q_{s,a}^*(x_s)\right| \leq C_1 L^{-m}$$

uniformly over all action-state pairs. We have

$$\begin{aligned}
\mathbb{E}\left\{q_{s,a}^2(x_s)\right\} &= \mathbb{E}\left[\left\{q_{s,a}^*(x_s) + q_{s,a}(x_s) - q_{s,a}^*(x_s)\right\}^2\right] \\
&\leq 2\mathbb{E}\left\{\left(q_{s,a}^*(x_s)\right)^2\right\} + 2\mathbb{E}\left[\left\{q_{s,a}(x_s) - q_{s,a}^*(x_s)\right\}^2\right] \qquad \text{(S4.63)} \\
&\leq 2\mathbb{E}\left\{\left(q_{s,a}^*(x_s)\right)^2\right\} + 2C_1^2 L^{-2m}.
\end{aligned}$$

Thus,

$$\mathbb{E}\left\{\left(q_{s,a}^*(x_s)\right)^2\right\} \geq \frac{v}{2} - C_1^2 L^{-2m}.$$

Since

$$L \asymp (nT)^{1/(2m-1)},$$

for all sufficiently large nT,

$$\mathbb{E}\left\{\left(q_{s,a}^*(x_s)\right)^2\right\} \geq \frac{v}{4}.$$

On the event $\mathcal{G}_1 \cap \mathcal{G}_2$, Theorem 5 implies that

$$\|\widehat{\boldsymbol{\kappa}}\|_2 \leq c_\kappa \sqrt{L} \frac{\log(nT \vee d)}{\sqrt{nT}}.$$

Using the same inequality as in (S4.63), we have

$$\begin{aligned}
\mathbb{E}\left\{\left(q^*_{s,a}(x_s)\right)^2\right\} &= \mathbb{E}\left[\left\{\widehat{q}_{s,a}(x_s) + q^*_{s,a}(x_s) - \widehat{q}_{s,a}(x_s)\right\}^2\right] \\
&\leq (1+0.5)\mathbb{E}\left\{\widehat{q}^2_{s,a}(x_s)\right\} + (1+2)\mathbb{E}\left[\left\{\widehat{q}_{s,a}(x_s) - q^*_{s,a}(x_s)\right\}^2\right] \\
&\leq 1.5\mathbb{E}\left\{\widehat{q}^2_{s,a}(x_s)\right\} + 3\mathbb{E}\left\{\|\Phi_s(x_s)\|_2^2\|\widehat{\boldsymbol{\kappa}}_{G_s}\|_2^2\right\} \\
&\leq 1.5\mathbb{E}\left\{\widehat{q}^2_{s,a}(x_s)\right\} + 3\left\{c_m c_\kappa L\frac{\log(nT \vee d)}{\sqrt{nT}}\right\}^2.
\end{aligned} \tag{S4.64}$$

The last inequality follows from

$$\|\widehat{\boldsymbol{\kappa}}_{G_s}\|_2 \leq \|\widehat{\boldsymbol{\kappa}}\|_2$$

and the definition

$$c_\kappa = \frac{8c_1\sqrt{2d_0+1}}{\bar{c}}.$$

Suppose that there exists an index s such that $s \in \mathcal{K}$ but $s \notin \widehat{\mathcal{K}}$. Then, for every $a \in \{0,1\}$,

$$\begin{aligned}
\mathbb{E}\left\{\widehat{q}^2_{s,a}(x_s)\right\} &\leq \mathbb{E}\left\{\|\Phi_s(x_s)\|_2^2\|\widehat{\boldsymbol{\beta}}_{G_s}\|_2^2\right\} \\
&\leq (c_m\sqrt{L})^2 h^2 \\
&\leq \frac{v}{8}.
\end{aligned} \tag{S4.65}$$

Combining (S4.63), (S4.64), and (S4.65), we obtain

$$\frac{v}{4} \leq \frac{3v}{16} + 3\left\{c_m c_\kappa L\frac{\log(nT \vee d)}{\sqrt{nT}}\right\}^2.$$

Consequently,

$$\frac{v}{16} \leq 3\left\{c_m c_\kappa L\frac{\log(nT \vee d)}{\sqrt{nT}}\right\}^2,$$

and hence

$$v \lesssim \left\{ L \frac{\log(nT \vee d)}{\sqrt{nT}} \right\}^2.$$

However, under

$$L \asymp (nT)^{1/(2m-1)},$$

we have

$$L \frac{\log(nT \vee d)}{\sqrt{nT}} \longrightarrow 0$$

as $nT \to \infty$. This contradicts the assumption that v is bounded away from zero. Therefore, on the event $\mathcal{G}_1 \cap \mathcal{G}_2$,

$$\mathcal{K} \subseteq \widehat{\mathcal{K}}.$$

Since $\mathcal{G}_1 \cap \mathcal{G}_2$ occurs with the probability stated in Theorem 5, the desired result follows. $\square$

S5 Numerical Results

Table 1: Simulation results for $d = 10$. Each entry reports bias (SD).

$n \backslash T$		$T = 20$	$T = 30$	$T = 40$
$n = 20$	SAVE	0.12 (0.46)	0.01 (0.25)	0.05 (0.15)
	LFQ	-0.17 (0.01)	-0.16 (0.02)	-0.14 (0.06)
	GD	0.02 (0.17)	-0.03 (0.16)	0.01 (0.12)
$n = 30$	SAVE	-0.02 (0.19)	-0.02 (0.14)	0.03 (0.12)
	LFQ	-0.17 (0.00)	-0.16 (0.01)	-0.15 (0.04)
	GD	0.02 (0.10)	-0.03 (0.11)	0.02 (0.10)
$n = 40$	SAVE	0.03 (0.11)	-0.01 (0.09)	0.03 (0.09)
	LFQ	-0.17 (0.00)	-0.17 (0.01)	-0.15 (0.03)
	GD	0.02 (0.08)	-0.01 (0.08)	0.02 (0.09)
$n = 50$	SAVE	0.04 (0.13)	0.01 (0.09)	0.03 (0.07)
	LFQ	-0.17 (0.00)	-0.17 (0.01)	-0.15 (0.02)
	GD	0.01 (0.07)	0.01 (0.08)	0.02 (0.07)

Table 2: Simulation results for $d = 100$. Each entry reports bias (SD).

$n \backslash T$		$T = 20$	$T = 30$	$T = 40$
$n = 20$	SAVE	0.09 (0.18)	0.07 (0.24)	-0.03 (0.28)
	LFQ	0.73 (0.19)	0.42 (0.23)	0.23 (0.19)
	GD	0.03 (0.21)	-0.01 (0.18)	-0.06 (0.20)
$n = 30$	SAVE	0.07 (0.15)	-0.02 (0.22)	-1.83 (5.64)
	LFQ	0.66 (0.15)	0.43 (0.16)	0.26 (0.14)
	GD	0.02 (0.14)	-0.02 (0.16)	-0.05 (0.16)
$n = 40$	SAVE	0.09 (0.11)	-1.34 (2.95)	0.58 (1.78)
	LFQ	0.66 (0.13)	0.44 (0.09)	0.25 (0.14)
	GD	0.03 (0.12)	0.00 (0.15)	-0.05 (0.12)
$n = 50$	SAVE	0.05 (0.16)	7.68 (17.78)	0.27 (1.49)
	LFQ	0.65 (0.11)	0.42 (0.09)	0.24 (0.15)
	GD	0.01 (0.10)	-0.03 (0.12)	-0.04 (0.12)

Table 3: Simulation Results of Cartpole Environment. Each entry reports bias (SD)

$n\backslash T$		$T = 50$	$T = 60$	$T = 70$
$n = 50$	SAVE	-0.33 (12.85)	-9.46 (58.79)	0.59 (5.43)
	LFQ	-12.85 (0.07)	-7.38 (0.20)	-3.25 (0.33)
	GD	8.18 (3.28)	2.66 (1.04)	1.75 (0.61)
$n = 60$	SAVE	-0.88 (14.76)	0.95 (25.33)	-0.31 (4.06)
	LFQ	-12.84 (0.07)	-7.52 (0.21)	-3.43 (0.17)
	GD	6.25 (2.32)	2.97 (1.00)	1.66 (0.91)
$n = 70$	SAVE	-6.14 (16.76)	-8.80 (24.89)	0.42 (6.61)
	LFQ	-12.87 (0.08)	-7.49 (0.16)	-3.47 (0.19)
	GD	6.57 (3.13)	2.52 (0.90)	1.54 (0.64)
$n = 80$	SAVE	-82.87 (357.27)	-3.05 (18.03)	2.41 (9.88)
	LFQ	-12.88 (0.05)	-7.56 (0.13)	-3.46 (0.15)
	GD	4.61 (1.34)	2.12 (0.94)	1.30 (0.41)

References

Chen, X. and T. M. Christensen (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics 188*(2), 447–465. Heterogeneity in Panel Data and in Nonparametric Analysis in honor of Professor Cheng Hsiao.

de Boor, C. (1976). Splines as linear combinations of B-splines. In G. G. Lorentz, C. K. Chui, and L. L. Schumaker (Eds.), *Approximation Theory II*, pp. 1–47. New York: Academic Press.

Ma, T., J. Zhu, H. Cai, Z. Qi, Y. Chen, C. Shi, and E. B. Laber (2026). Sequential knockoffs for variable selection in reinforcement learning. *Journal of the American Statistical Association 0*(ja), 1–29.

Shi, C., S. Zhang, W. Lu, and R. Song (2022). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(3), 765–793.

Tropp, J. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability 16*(none), 262 – 270.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics 12*(4), 389–435.