Supplementary Material: Off-Policy Evaluation for Episodic Partially Observable Markov Decision Processes under Non-Parametric Models

Contents

1 Introduction **1**

2 Related Work **2**

3 Preliminaries and Notations **3**

4 Identification Results **4**

5 Estimation **5**

6 Theoretical Results **6**

7 Simulation **9**

8 Discussion **10**

A Additional Identification Assumptions **18**

A.1 Basic assumptions on the confounded POMDP structure 18

A.2 Assumptions on the existence of bridge functions 18

A.3 Assumptions on the uniqueness of bridge functions 19

B Additional Results **19**

B.1 Additional Finite-sample error bounds for V-bridge estimation and OPE 20

B.1.1 VC-subgraph class 20

B.1.2 RKHS with exponential eigen-decay 20

B.2 Different choices of proxy variables 20

C Technical Proofs **22**

C.1 Proof of Theorem 4.1 22

C.2 Proof of Theorem 6.1 23

C.2.1 Decomposition of $\mathcal{L}^2$-error of v_1^π 23

C.2.2 Error bounds for projected one-step error 24

C.3 Proof of Theorem 6.2 25

C.3.1 One-step error bound 25

C.3.2 Combined Result 26

C.4 Localized ill-posedness τ_t and one-step transition ill-posedness $C^{(t)}_{t',t'-1}$ 26

C.5 Proofs of Theorems 6.3, B.1 and B.2 27

C.5.1 Decomposition of Off-Policy Value Estimation Error 27
C.5.2 Applying decomposition of OPE error 28

D Auxiliary Lemmas **29**
D.1 Lemmas For Identification 29
D.2 One-step estimation error 29
D.3 Critical radii and local Rademacher complexity 30
D.3.1 Local Rademacher complexity bound by entropy integral 30
D.3.2 Local Rademacher complexity bound for RKHSs 31
D.4 Proof of Lemmas 32
D.4.1 Proof of Lemma C.1 32
D.4.2 Proof of Lemma D.2 33
D.4.3 Proof of Lemma D.5 37
D.4.4 Proof of Lemma D.7 38

E Additional estimation details **39**

F Simulation details **40**
F.1 Simulation setup 40
F.2 Implementation 41

List of Notations

Table 1: List of Notations

$\mathcal{M}$	the episodic and confounded POMDP
$S_t \in \mathcal{S}$	observed state at t and observed state space
$U_t \in \mathcal{U}$	unobserved state at t and unobserved state space
$A_t \in \mathcal{A}$	action at t and discrete action space
T	length of horizon
$r = \{r_t\}_{t=1}^T$	reward functions over $\mathcal{S} \times \mathcal{U} \times \mathcal{A}$
R_t	reward at t
$W_t \in \mathcal{W}$	reward-proxy variable at t and corresponding space
$Z_t \in \mathcal{W}$	action-proxy variable at t and corresponding space
$X_{t,i}$	variable X at t from sample trajectory i
$\pi = \{\pi_t\}_{t=1}^T$	target policy depending on S_t
$\tilde{\pi}_t^b$	behavior policy at t depending on S_t, U_t
$V_t^\pi(s,u)$	state value function
$\mathcal{V}(\pi)$ ($\widehat{\mathcal{V}}(\pi)$)	(estimated) policy value of a target policy π
v_t^π ($\hat{v}_t^\pi$)	(estimated) V-bridge function (or V-bridge for short) at t
q_t^π ($\hat{q}_t^\pi$)	(estimated) Q-bridge function (or Q-bridge for short) at t
$\widetilde{\mathcal{P}}_t$	operator $[\widetilde{\mathcal{P}}_t](Z_t, S_t, A_t) = \mathbb{E}[g(R_t, W_{t+1}, S_{t+1}) \mid Z_t, S_t, A_t]$
$\overline{\mathcal{P}}_t$	operator $[\overline{\mathcal{P}}_t](Z_t, S_t, A_t) = \mathbb{E}[h(W_t, S_t, A_t) \mid Z_t, S_t, A_t]$
$\mathcal{P}_t$ ($\widehat{\mathcal{P}}_t$)	operator $\mathcal{P}_t g = \overline{\mathcal{P}}_t^{-1} \widetilde{\mathcal{P}}_t g$ (estimator of $\mathcal{P}_t$ defined in (7))
$\mathcal{P}_t^\pi$ ($\widehat{\mathcal{P}}_t^\pi$)	operator $\mathcal{P}_t^\pi g = \langle \pi_t, \mathcal{P}_t g \rangle$ (estimator of $\mathcal{P}_t^\pi$: $\widehat{\mathcal{P}}_t^\pi g = \left\langle \pi_t, \widehat{\mathcal{P}}_t g \right\rangle$)
$\mathcal{H}^{(t)}$	user-defined function space on $\mathcal{W} \times \mathcal{S} \times \mathcal{A}$
$\mathcal{F}^{(t)}$	user-defined function space on $\mathcal{Z} \times \mathcal{S} \times \mathcal{A}$
$\mathcal{G}^{(t)}$	user-defined function space on $\mathcal{Z} \times \mathcal{S}$
$\mathcal{R}_n(\mathcal{F}, \delta)$	local Rademacher complexity for function class $\mathcal{F}$ and radius $\delta > 0$
$\widehat{\mathcal{R}}_n(\mathcal{F}, \delta)$	local empirical Rademacher complexity for function class $\mathcal{F}$ and radius $\delta > 0$
$N_n(\epsilon, \mathcal{G})$	the smallest empirical ϵ-covering number of $\mathcal{G}$
$\alpha\mathcal{F}$	$\alpha\mathcal{F} = \{\alpha f : f \in \mathcal{F}\}$ for some $\alpha \in \mathbb{R}$
$\mathcal{F}_B$	$\mathcal{F}_B = \{f \in \mathcal{F} : \|f\|_{\mathcal{F}}^2 \leq B\}$ for any $B > 0$
$\|\text{proj}_t f\|_2$	$\|\text{proj}_t f\|_2 = \sqrt{\mathbb{E}\{f(X) \mid Z_t, S_t, A_t\}^2}$
$\bar{\tau}_1$	ill-posedness $\bar{\tau}_1 = \sup_{g \in \mathcal{G}^{(1)}} \|g(W_1, S_1)\|_2 / \|\mathbb{E}[g(W_1, S_1) \mid Z_1, S_1]\|_2$
τ_t	ill-posedness $\tau_t = \sup_{h \in \mathcal{H}^{(t)}} \|h(W_t, S_t, A_t)\|_2 / \|\text{proj}_t h(W_t, S_t, A_t)\|_2$
$C_{t',t'-1}^{(t)}$	one-step transition ill-posedness defined after Corollary 6.2
$\mathbb{V}(\mathcal{F})$	VC dimension of $\mathcal{F}$
$\zeta(\alpha)$	Riemann Zeta function $\zeta(\alpha) = \sum_{n=1}^\infty (1/n)^\alpha$
$\text{Ker}(K)$	$\text{Ker}(K) = \{g : Kg = 0\}$ null space of linear operator K
$A^\perp$	orthogonal complement of space A
$\|\mathcal{Z}\|$	cardinality of class Z

A Additional Identification Assumptions

In this section, we list Assumptions 3-7 which are needed for Theorem 4.1.

A.1 Basic assumptions on the confounded POMDP structure

For the confounded POMDP with trajectory $(U_t, S_t, W_t, Z_t, A_t, R_t)_{t=1}^T$, we list three basic assumptions below. Let $\perp\!\!\!\perp$ and $\not\perp\!\!\!\perp$ denote statistical independence and dependence respectively.

Assumption 3 (Markovian)**.** For all $1 \le t \le T$, the time-variant transition kernel $\mathbb{P}_t$ satisfies that for any $(s,u) \in \mathcal{S} \times \mathcal{U}$, $a \in \mathcal{A}$ and set $F \in \mathcal{B}(\mathcal{S} \times \mathcal{U})$,

$$\begin{aligned}&\Pr((S_{t+1}, U_{t+1}) \in F \mid S_t = s, U_t = u, A_t = a, \{S_j, U_j, A_j\}_{1 \le j < t}) \\ &= \mathbb{P}_t((S_{t+1}, U_{t+1}) \in F \mid S_t = s, U_t = u, A_t = a),\end{aligned}$$

where $\mathcal{B}(\mathcal{S} \times \mathcal{U})$ is the family of Borel subsets of $\mathcal{S} \times \mathcal{U}$ and $\{S_j, U_j, A_j\}_{1 \le j < t} \neq \emptyset$ if $t = 1$.

Assumption 4 (Reward proxy)**.** $W_t \perp\!\!\!\perp (A_t, U_{t-1}, S_{t-1}) \mid U_t, S_t$ and $W_t \not\perp\!\!\!\perp U_t \mid S_t$ for $1 \le t \le T$.

Assumption 5 (Action proxy)**.** $Z_t \perp\!\!\!\perp W_t \mid (U_t, S_t, A_t)$, $Z_t \perp\!\!\!\perp R_t \mid (U_t, S_t, A_t)$ and $Z_t \perp\!\!\!\perp (S_{t+1}, W_{t+1}) \mid (U_t, S_t, A_t)$, $1 \le t \le T$.

It can be easily verified that the DAG in Figure 1 satisfies Assumptions 3-5. Assumption 3 requires that given the current full state and action (U_t, S_t, A_t), the future are independent of the past.

Assumption 4 requires that the reward proxy W_t is associated with the hidden state U_t after adjusting observed state S_t but W_t is not causally affected by action A_t and past state (U_{t-1}, S_{t-1}) after adjusting the full current state (U_t, S_t). This assumption does not restrict the association between W_t and R_t. Assumption 5 requires that upon conditioning on the current full state and action tuple (U_t, S_t, A_t), the action proxy Z_t does not affect the reward proxy W_t and outcomes R_t, S_{t+1}, W_{t+1} after the action A_t. Again, this assumption does not restrict the association between Z_t and A_t.

However, based on above three assumptions, we cannot directly identify the value of target policy π by adjusting (U_t, S_t) since U_t is unobserved. In addition to Assumptions 4 and 5, we also need Assumption 6 to be stated in Section A.2 below to get around the hidden state U_t.

A.2 Assumptions on the existence of bridge functions

Assumption 6 (Completeness)**.** For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $t = 1, \ldots, T$,

(a) For any square-integrable function g, $\mathbb{E}\{g(U_t) \mid Z_t, S_t = s, A_t = a\} = 0$ a.s. if and only if $g = 0$ a.s;

(b) For any square-integrable function g, $\mathbb{E}\{g(Z_t) \mid W_t, S_t = s, A_t = a\} = 0$ a.s. if and only if $g = 0$ a.s.

Completeness is a commonly made technical assumption in value identification problems, e.g., instrumental variable identification [Newey and Powell, 2003, D'Haultfoeuille, 2011, Chen et al., 2014], and proximal causal inference [Miao et al., 2018a,b, Tchetgen Tchetgen et al., 2020]. Together with the regularity conditions in Assumption 7, we can ensure the existence of Q-bridges q_t^π and V-bridges v_t^π, $1 \le t \le T$.

For a probability measure function μ, let $\mathcal{L}^2\{\mu(x)\}$ denote the space of all squared integrable functions of x with respect to measure $\mu(x)$, which is a Hilbert space endowed with the inner product $\langle g_1, g_2 \rangle = \int g_1(x) g_2(x) \mathrm{d}\mu(x)$. For all s, a, t, define the following operator

$$\begin{aligned}K_{s,a;t} : \mathcal{L}^2\left\{\mu_{W_t|S_t,A_t}(w \mid s,a)\right\} &\to \mathcal{L}^2\left\{\mu_{Z_t|S_t,A_t}(z \mid s,a)\right\} \\ h &\mapsto \mathbb{E}\left\{h(W_t) \mid Z_t = z, S_t = s, A_t = a\right\},\end{aligned}$$

and its adjoint operator

$$\begin{aligned}K^*_{s,a;t} : \mathcal{L}^2\left\{\mu_{Z_t|S_t,A_t}(z \mid s,a)\right\} &\to \mathcal{L}^2\left\{\mu_{W_t|S_t,A_t}(w \mid s,a)\right\} \\ g &\mapsto \mathbb{E}\left\{g(Z_t) \mid W_t = w, S_t = s, A_t = a\right\}.\end{aligned}$$

Assumption 7 (Regularity Conditions)**.** For any $Z_t = z, S_t = s, W_t = w, A_t = a$ and $1 \le t \le T$,

(a) $\iint_{\mathcal{W}\times\mathcal{Z}} f_{W_t|Z_t,S_t,A_t}(w \mid z,s,a) f_{Z_t|W_t,S_t,A_t}(z \mid w,s,a)\mathrm{d}w\mathrm{d}z < \infty$, where $f_{W_t|Z_t,S_t,A_t}$ and $f_{Z_t|W_t,S_t,A_t}$ are conditional density functions.

(b) For any $g \in \mathcal{G}^{(t+1)}$,

$$\int_{\mathcal{Z}} \left[\mathbb{E}\left\{R_t + g(W_{t+1}, S_{t+1}) \mid Z_t = z, S_t = s, A_t = a\right\}\right]^2 f_{Z_t|S_t,A_t}(z \mid s,a)\mathrm{d}z < \infty.$$

(c) There exists a singular decomposition $(\lambda_{s,a;t;\nu}, \phi_{s,a;t;\nu}, \psi_{s,a;t;\nu})_{\nu=1}^{\infty}$ of $K_{s,a;t}$ such that for all $g \in \mathcal{G}^{(t+1)}$,

$$\sum_{\nu=1}^{\infty} \lambda_{s,a;t;\nu}^{-2} \left|\left\langle \mathbb{E}\left\{R_t + g(W_{t+1}, S_{t+1}) \mid Z_t = z, S_t = s, A_t = a\right\}, \psi_{s,a;t;\nu}\right\rangle\right|^2 < \infty.$$

(d) For all $1 \le t \le T$, $v_t^\pi \in \mathcal{G}^{(t)}$ where $\mathcal{G}^{(t)}$ satisfies the regularity conditions (b) and (c) above.

Note that the existence of the singular decomposition of $K_{s,a,;t}$ in Assumption 7 (c) can be ensured by Assumption 7 (a), which is a sufficient condition for the compactness of $K_{s,a;t}$ by Lemma D.1.

For tabular $(\mathcal{U}, \mathcal{W}, \mathcal{Z})$, Corollary A.1 provides a sufficient condition for Assumptions 6 and 7 [Shi et al., 2020].

Corollary A.1. [Shi et al., 2020] Suppose that all $\mathcal{U}$, $\mathcal{W}$, and $\mathcal{Z}$ are tabular. If both Z_t and W_t have at least as many categories as U_t for $1 \le t \le T$, i.e., $|\mathcal{Z}|, |\mathcal{W}| \ge |\mathcal{U}|$ (where $|\mathcal{X}|$ is the cardinality of set $\mathcal{X}$), and transition probability matrices $P_t(\mathbf{W} \mid \mathbf{U}, s) \triangleq [P_t(w_i \mid u_j, s)]_{w_i\in\mathcal{W}, u_j\in\mathcal{U}}$ and $P_t(\mathbf{U} \mid \mathbf{Z}, a, s) \triangleq [P_t(u_i \mid z_j, a, s)]_{u_i\in\mathcal{U}, z_j\in\mathcal{Z}}$ are of full rank with rank $|\mathcal{U}|$ for all a, s, t, then Assumptions 6 and 7 hold.

A.3 Assumptions on the uniqueness of bridge functions

In general, we do not need to impose restrictions on the uniqueness of V-bridges $\{v_t^\pi\}_{t=1}^T$ for policy value identification. To simplify our theoretical analysis on the estimation error of V-bridges, we need the uniqueness of V-bridges $\{v_t^\pi\}_{t=1}^T$ and Q-bridges $\{q_t^\pi\}_{t=1}^T$, which can be ensured by the following Assumption 8.

Assumption 8. For any square-integrable function g and for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, $\mathbb{E}\{g(W_t) \mid Z_t, S_t = s, A_t = a\} = 0$ a.s. if and only $g = 0$ a.s.

Corollary A.2. Under Assumption 8 and all conditions in Theorem 4.1, the V-bridges $\{v_t^\pi\}_{t=1}^T$ that satisfy (3) and Q-bridges $\{q_t^\pi\}_{t=1}^T$ that satisfy (5) are both unique. Moreover, they can be non-parametrically identified by (4).

Proof. Apparently it suffices to prove the uniqueness of Q-bridges $\{q_t^\pi\}_{t=1}^T$. If there is another set of $\{\tilde{q}_t^\pi\}_{t=1}^T$ that is also a solution to (4), then

$$\mathbb{E}\left\{\tilde{q}_t^\pi(W_t, S_t, A_t) - q_t^\pi(W_t, S_t, A_t) \mid Z_t, S_t = s, A_t = a\right\} = 0, \quad \text{a.s.}$$

By Assumption 8, $\tilde{q}_t^\pi(W_t, s, a) = q_t^\pi(W_t, s, a)$ a.s. for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. □

For the tabular case, we have the following corollary for the uniqueness of V-bridges and Q-bridges.

Corollary A.3. [Shi et al., 2020] Under the conditions in Corollary A.1, if $|\mathcal{Z}| = |\mathcal{W}| = |\mathcal{U}|$, then Assumptions 6–8 are satisfied.

B Additional Results

In this section, we derive finite-sample error bounds for V-bridge estimation and OPE when hypothesis spaces $\mathcal{H}^{(t)}$, $\mathcal{G}^{(t)}$ and testing space $\mathcal{F}^{(t)}$ are VC-subgraph classes or RKHSs with exponential eigen-decay. Then we discuss possible choices of proximal variables W_t and Z_t.

B.1 Additional Finite-sample error bounds for V-bridge estimation and OPE

B.1.1 VC-subgraph class

Theorem B.1. Under Assumptions 1 and 2, and the assumptions in Theorem 6.2 and Corollary 6.1, with probability at least $1-\zeta$, we have

$$\|v_1^\pi - \hat{v}_1^\pi\|_2 \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2} \left\{ \sqrt{\frac{\max_{1\le t\le T}\left\{\mathbb{V}(\mathcal{F}^{(t)}), \mathbb{V}(\mathcal{H}^{(t)}), \mathbb{V}(\mathcal{G}^{(t+1)})\right\}}{n}} + \sqrt{\frac{\log(T/\zeta)}{n}} \right\}, \text{and}$$

$$|\mathcal{V}(\pi) - \hat{\mathcal{V}}(\pi)| \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2} \left\{ \sqrt{\frac{\max_{1\le t\le T}\left\{\mathbb{V}(\mathcal{F}^{(t)}), \mathbb{V}(\mathcal{H}^{(t)}), \mathbb{V}(\mathcal{G}^{(t+1)})\right\}}{n}} + \sqrt{\frac{\log(T/\zeta)}{n}} \right\},$$

where $\texttt{trans-ill} = \max_{1\le t\le T} \exp\{a_t \zeta(\alpha_t)\}$ with $\zeta(\alpha) = \sum_{t=1}^{\infty} t^{-\alpha}$, and $\texttt{ill}_{\max} = \tau_{\pi_1} \max_{1\le t\le T} \tau_t \|\pi_t/\pi_t^b\|_\infty^2$.

The proof of Theorem B.1 is given in Appendix C.5.

B.1.2 RKHS with exponential eigen-decay

Theorem B.2. Under Assumptions 1 and 2, and the assumptions in Theorem 6.2 and Corollary 6.2 (2), with probability at least $1-\zeta$, we have

$$\|v_1^\pi - \hat{v}_1^\pi\|_2 \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2} \left\{ \sqrt{\frac{(\log n)^{1/\min\{\beta_{\mathcal{H}}, \beta_{\mathcal{G}}, \beta_{\mathcal{F}}\}}}{n}} + \sqrt{\frac{\log(T/\zeta)}{n}} \right\}, \text{and}$$

$$|\mathcal{V}(\pi) - \hat{\mathcal{V}}(\pi)| \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2} \left\{ \sqrt{\frac{(\log n)^{1/\min\{\beta_{\mathcal{H}}, \beta_{\mathcal{G}}, \beta_{\mathcal{F}}\}}}{n}} + \sqrt{\frac{\log(T/\zeta)}{n}} \right\}.$$

where $\texttt{trans-ill} = \max_{1\le t\le T} \exp\{a_t \zeta(\alpha_t)\}$ with $\zeta(\alpha) = \sum_{t=1}^{\infty} t^{-\alpha}$, and $\texttt{ill}_{\max} = \tau_{\pi_1} \max_{1\le t\le T} \tau_t \|\pi_t/\pi_t^b\|_\infty^2$.

The proof of Theorem B.2 is given in Appendix C.5.

B.2 Different choices of proxy variables

Here we first provide several options on how to choose proxy variables W_t and Z_t satisfying basic assumptions 3 –5. Then we discuss their effect on the ill-posedness and one step estimation errors. Finally, we comment on some practical issues.

Choice of W_t. In our confounded POMDP setting, typically we need a reward-inducing proxy W_t to be separated from the current observations at time t and satisfy the basic assumptions listed in Appendix A.1. In practice, W_t can be some environmental variables that are correlated with the outcome R_t but A_t cannot affect W_t (see Figure 3). It is worth mentioning that Bennett and Kallus [2021] and Shi et al. [2021] use (part of) the current observed state, i.e., S_t in our paper, as the reward-inducing proxy. In their settings, given the current action A_t, only the hidden state U_t can affect the next hidden state U_{t+1} (Their U_t is the full state variables in our setting). This requires that the proximal variables Z_t and W_t are able to capture the whole information of their hidden state U_t. In this case, Assumption 6 becomes harder to hold. In our setting, however, we allow part of their U_t to be observable. We denote this part by S_t in our paper. This can alleviate the burden on proximal variables Z_t and W_t to capture the whole information of their hidden U_t. Therefore, our completeness assumption 6 is relatively weaker. Moreover, Bennett and Kallus [2021] only consider the evaluation for deterministic target policies, while in our setting, a separate W_t (other than S_t) allows us to evaluate random target policies.

We list some possible causal relationship among W_t, (U_t, S_t) and R_t in Figure 3. We require the causal relationship between U_t and W_t. But the effect of W_t on R_t is optional. In practice, one can

use the observed variables that have no direct effect on the action, for example, measurement of action independent disturbance which may not may not affect the current reward.

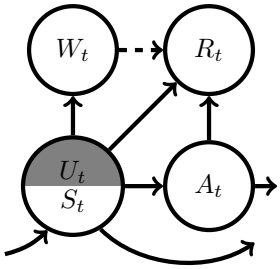
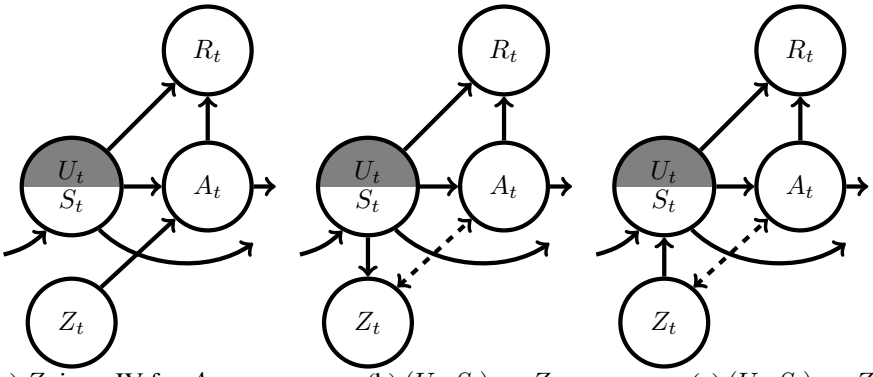


Figure 3: Causal relationship about W_t. Dashed arrows: optional causal effect. W_t may or may not affect R_t.



(a) Z_t is an IV for A_t. (b) $(U_t, S_t) \to Z_t$ (c) $(U_t, S_t) \leftarrow Z_t$

Figure 4: Causal relationship about Z_t. Dashed arrows: optional causal effect $Z_t \to A_t$ or $Z_t \leftarrow A_t$ or no causal effect. (c) is incompatible with Figure 3 (b).

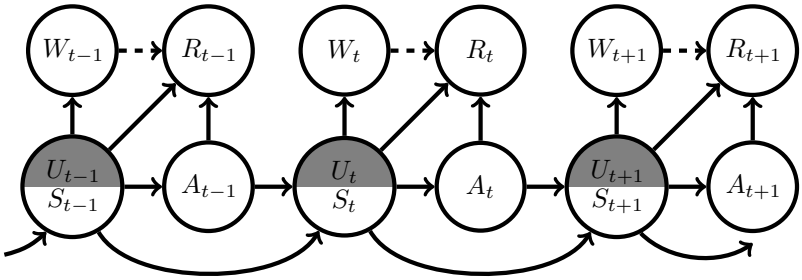


Figure 5: An example of Z_t as the observed history.

Choice of Z_t. Once we determine W_t, there are several proper choices of Z_t that are compatible with W_t (see Figure 4). One choice of Z_t is the observed history up to step $t-1$, e.g., $Z_t = (Z_{t-1}; S_{t-1}, W_{t-1}, A_{t-1}, R_{t-1})$ with some pre-observed history before (U_1, S_1) as Z_1. See Figure 5 for a valid example. In this case Z_{t+1} contains information of Z_t so that we expect that $C^{(t)}_{t',t'+1}$ tends to be smaller. However, this can enlarge the one-step errors $M_{\mathcal{H}}(T-t+1)^2 \left(\bar{\delta}_n^{(t)} + c_0 \sqrt{\frac{\log(c_1 T/\zeta)}{n}} \right)$, where the upper bound of critical radii $\bar{\delta}_n^{(t)}$ becomes larger because the dimension of testing space $\mathcal{F}^{(t)}(\mathcal{Z} \times \mathcal{A} \times \mathcal{S})$ is now $\mathcal{O}(t)$. Fortunately, these one-step errors only contribute to the final error bound for $\mathcal{V}(\pi)$ linearly.

In practice, to reduce the dimension of Z_t, one may use the most recent k-step observed history, or try to learn a low dimensional representation $\phi(Z_t)$ of Z_t and then replace $\mathcal{F}^{(t)}(\mathcal{Z}\times\mathcal{A}\times\mathcal{S})$ by $\tilde{\mathcal{F}}^{(t)}(\phi(\mathcal{Z})\times\mathcal{A}\times\mathcal{S})$ in (7). Similar ideas have been used in kernel IV regression [Singh, 2020].

C Technical Proofs

In this section, we provide the proofs of identification result in Section 3 and the finite sample bounds for V-bridges and OPE in Section 6.

C.1 Proof of Theorem 4.1

Part I. We suppose there exists q_t^π satisfying (4), $1\le t\le T$. Define $v_{T+1}^\pi=0$. Then

$$\begin{aligned}
&\mathbb{E}\left\{R_t+v_{t+1}^\pi(W_{t+1},S_{t+1})\mid Z_t,S_t,A_t\right\}\\
=&\mathbb{E}\left[\mathbb{E}\left\{R_t+v_{t+1}^\pi(W_{t+1},S_{t+1})\mid U_t,Z_t,S_t,A_t\right\}\mid Z_t,S_t,A_t\right]\\
=&\mathbb{E}\left[\mathbb{E}\left\{R_t+v_{t+1}^\pi(W_{t+1},S_{t+1})\mid U_t,S_t,A_t\right\}\mid Z_t,S_t,A_t\right]\ \text{by Assumption 5,}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left\{q_t^\pi(W_t,S_t,A_t)\mid Z_t,S_t,A_t\right\}\\
=&\mathbb{E}\left[\mathbb{E}\left\{q_t^\pi(W_t,S_t,A_t)\mid U_t,Z_t,S_t,A_t\right\}\mid Z_t,S_t,A_t\right]\\
=&\mathbb{E}\left[\mathbb{E}\left\{q_t^\pi(W_t,S_t,A_t)\mid U_t,S_t,A_t\right\}\mid Z_t,S_t,A_t\right]\ \text{by Assumption 5.}
\end{aligned}$$

Therefore, by Assumption 6 (a), we have

$$\mathbb{E}\left\{R_t+v_{t+1}^\pi(W_{t+1},S_{t+1})\mid U_t,S_t,A_t\right\}=\mathbb{E}\left\{q_t^\pi(W_t,S_t,A_t)\mid U_t,S_t,A_t\right\}\quad\text{a.s.}\tag{10}$$

We will use this Bellman-like equation (10) to verify (3) and (5).

Next, we prove that such these $\{q_t^\pi,v_t^\pi\}_{t=1}^T$ obtained by Algorithm 1 can be used as Q-bridges (5) and V-bridges (3).

First, at time T,

$$\begin{aligned}
\mathbb{E}^\pi\left(R_T\mid U_T,S_T\right)&=\sum_{a_T\in\mathcal{A}}\mathbb{E}\left(R_T\mid U_T,S_T,A_T=a_T\right)\pi_T(a_T\mid S_T)\\
&=\sum_{a_T\in\mathcal{A}}\mathbb{E}\left\{q_T^\pi(W_T,S_T,a_T)\mid U_T,S_T,A_T=a_T\right\}\pi_T(a_T\mid S_T)\ \text{by (10)}\\
&=\sum_{a_T\in\mathcal{A}}\mathbb{E}\left\{q_T^\pi(W_T,S_T,a_T)\mid U_T,S_T\right\}\pi_T(a_T\mid S_T)\ \text{by Assumption 4}\\
&=\mathbb{E}\left\{\sum_{a_T\in\mathcal{A}}\pi(a_T\mid S_T)q_T^\pi(W_T,S_T,a_T)\,\middle|\,U_T,S_T\right\}\\
&=\mathbb{E}\left\{v_T^\pi(W_T,S_T)\mid U_T,S_T\right\}\ \text{by definition of } v_T^\pi.
\end{aligned}$$

By induction, suppose that at time $t+1$, $\mathbb{E}^\pi\left[\sum_{t'=t+1}^T R_{t'} \mid S_{t+1}, U_{t+1}\right] = \mathbb{E}\left\{v_{t+1}^\pi(W_{t+1}, S_{t+1}) \mid S_{t+1}, U_{t+1}\right\}$. Then at time t,

$$
\begin{aligned}
&\mathbb{E}^\pi\left(\sum_{t'=t}^T R_{t'} \,\middle|\, U_t, S_t\right)\\
=&\mathbb{E}^\pi\left\{R_t + \mathbb{E}^\pi\left(\sum_{t'=t+1}^T R_{t'} \,\middle|\, U_{t+1}, S_{t+1}, U_t, S_t\right) \,\middle|\, U_t, S_t\right\}\\
=&\mathbb{E}^\pi\left\{R_t + \mathbb{E}^\pi\left(\sum_{t'=t+1}^T R_{t'} \,\middle|\, U_{t+1}, S_{t+1}\right) \,\middle|\, U_t, S_t\right\} \text{ by Assumption 3}\\
=&\mathbb{E}^\pi\left\{R_t + \mathbb{E}\left(v_{t+1}^\pi(W_{t+1}, S_{t+1}) \,\middle|\, U_{t+1}, S_{t+1}\right) \,\middle|\, U_t, S_t\right\}\\
=&\mathbb{E}^\pi\left\{R_t + \mathbb{E}\left(v_{t+1}^\pi(W_{t+1}, S_{t+1}) \,\middle|\, U_{t+1}, S_{t+1}, U_t, S_t\right) \,\middle|\, U_t, S_t\right\} \text{ by Assumption 4}\\
=&\mathbb{E}^\pi\left\{R_t + v_{t+1}^\pi(W_{t+1}, S_{t+1}) \mid U_t, S_t\right\} \text{ by the law of total expectation and Assumption 4}\\
=&\sum_{a_t \in \mathcal{A}} \mathbb{E}\left\{R_t + v_{t+1}^\pi(W_{t+1}, S_{t+1}) \mid U_t, S_t, A_t = a_t\right\} \pi_t(a_t \mid S_t)\\
=&\sum_{a_t \in \mathcal{A}} \mathbb{E}\left\{q_t^\pi(W_t, S_t, a_t) \mid U_t, S_t, A_t = a_t\right\} \pi_t(a_t \mid S_t) \text{ by (10)}\\
=&\sum_{a_t \in \mathcal{A}} \mathbb{E}\left\{q_t^\pi(W_t, S_t, a_t) \mid U_t, S_t\right\} \pi_t(a_t \mid S_t) \text{ by Assumption 4}\\
=&\mathbb{E}\left\{\sum_{a_t \in \mathcal{A}} \pi(a_t \mid S_t) q_t^\pi(W_t, S_t, a_t) \,\middle|\, U_t, S_t\right\}\\
=&\mathbb{E}\left\{v_t^\pi(W_t, S_t) \mid U_t, S_t\right\} \text{ by definition of } v_t^\pi.
\end{aligned}
$$

Therefore (3) hold for all $1 \le t \le T$. The validity of Q-bridge (5) can be similarly verified by restricting on $A_t = a$, for each $a \in \mathcal{A}$.

Part II. Now we prove the existence of the solution to (4).

For $t = T, \ldots, 1$, by Assumption 7 (a), $K_{s,a;t}$ is a compact operator for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ [Carrasco et al., 2007, Example 2.3], so there exists a singular value system stated in Assumption 7 (c) by Lemma D.1. Then by Assumption 6 (b), we have $\mathrm{Ker}(K_{s,a;t}^*) = 0$, since for any $g \in \mathrm{Ker}(K_{s,a;t}^*)$, we have, by the definition of Ker, $K_{s,a;t}^* g = \mathbb{E}\left[g(Z_t) \mid W_t, S_t = s, A_t = a\right] = 0$, which implies that $g = 0$ a.s. Therefore $\mathrm{Ker}(K_{s,a;t}^*) = 0$ and $\mathrm{Ker}(K_{s,a;t}^*)^\perp = \mathcal{L}^2(\mu_{Z_t|S_t,A_t}(z \mid s,a))$. By Assumption 7 (b), $\mathbb{E}\left\{R_t + g(W_{t+1}, S_{t+1}) \mid Z_t = \cdot, S_t = s, A_t = a\right\} \in \mathrm{Ker}(K_{s,a,;t}^*)$ for given $(s,a) \in \mathcal{S}_t \times \mathcal{A}$ and any $g \in \mathcal{G}^{(t+1)}$. Now we have verified the condition (a) in Lemma D.1. The condition (b) is satisfied given Assumption 7 (c). Recursively applying the above argument from $t = T$ to $t = 1$ yields the existence of the solution to (4).

□

C.2 Proof of Theorem 6.1

By definition and Assumptions 6–8, $\mathcal{P}_t^\pi$, $t = 1, \ldots, T$, are linear operators, i.e., $\mathcal{P}_t^\pi(\alpha_1 g_1 + \alpha_2 g_2) = \alpha_1 \mathcal{P}_t^\pi g_1 + \alpha_2 \mathcal{P}_t^\pi g_2$, for any $\alpha_1, \alpha_2 \in \mathbb{R}$ and $g_1, g_2 \in \mathcal{L}^2(\mathcal{R} \times \mathcal{W} \times \mathcal{S})$.

We first decompose $\hat{v}_t^\pi - v_t^\pi$ into a summation of projections of one-step error. Then we bound each one-step error by the projected errors times a product of transition ill-posedness.

C.2.1 Decomposition of $\mathcal{L}^2$-error of v_1^π

Following the identification procedure in Algorithm 3, we can decompose v_1^π by

$$v_1^\pi = \mathcal{P}_1^\pi(R_1 + v_2^\pi) = \mathcal{P}_1^\pi(R_1 + \mathcal{P}_2^\pi(R_2 + v_3^\pi)) = \cdots = \mathcal{P}_1^\pi(R_1 + \mathcal{P}_2^\pi(R_2 + \mathcal{P}_3^\pi(\cdots + \mathcal{P}_T^\pi R_T))).$$

Similarly, according to Section 5, we have the empirical version

$$\hat{v}_1^\pi = \widehat{\mathcal{P}}_1^\pi(R_1 + \hat{v}_2^\pi) = \widehat{\mathcal{P}}_1^\pi(R_1 + \widehat{\mathcal{P}}_2^\pi(R_2 + \hat{v}_3^\pi)) = \cdots = \widehat{\mathcal{P}}_1^\pi(R_1 + \widehat{\mathcal{P}}_2^\pi(R_2 + \widehat{\mathcal{P}}_3^\pi(\cdots + \widehat{\mathcal{P}}_T^\pi R_T))).$$

Then for each $t = 1, \ldots, T$, we can decompose $\hat{v}_t^\pi - v_t^\pi$ as

$$\begin{aligned}
\hat{v}_t^\pi - v_t^\pi &= \widehat{\mathcal{P}}_t^\pi(R_t + \hat{v}_{t+1}^\pi) - \mathcal{P}_t^\pi(R_t + v_{t+1}^\pi) \\
&= [\widehat{\mathcal{P}}_t^\pi(R_t + \hat{v}_{t+1}^\pi) - \mathcal{P}_t^\pi(R_t + \hat{v}_{t+1}^\pi)] + [\mathcal{P}_t^\pi(R_t + \hat{v}_{t+1}^\pi) - \mathcal{P}_t^\pi(R_t + v_{t+1}^\pi)] \\
&\triangleq g_t + [\mathcal{P}_t^\pi(R_t + \hat{v}_{t+1}^\pi) - \mathcal{P}_t^\pi(R_t + v_{t+1}^\pi)] \\
&= g_t + \mathcal{P}_t^\pi[\hat{v}_{t+1}^\pi - v_{t+1}^\pi],
\end{aligned} \tag{11}$$

where the last equality is due the the linearity of $\mathcal{P}_t^\pi$, and $v_{T+1}^\pi = \hat{v}_{T+1}^\pi \triangleq 0$. Recursively we have

$$\hat{v}_1^\pi - v_1^\pi = g_1 + \mathcal{P}_1^\pi g_2 + \mathcal{P}_{1:2}^\pi g_3 + \cdots + \mathcal{P}_{1:T-1}^\pi g_T, \tag{12}$$

where $\mathcal{P}_{t':t}^\pi \triangleq \mathcal{P}_{t'}^\pi \ldots \mathcal{P}_t^\pi$. If $t < t'$, $\mathcal{P}_{t':t}^\pi \triangleq \mathcal{I}$, the identity operator.

By the definition of the ill-posedness and combining the above decomposition, we can obtain the discrepancy between $\hat{v}_1^\pi$ and v_1^π:

$$\begin{aligned}
\|v_1^\pi - \hat{v}_1^\pi\|_2 &\le \bar{\tau}_1 \|\mathbb{E}(v_1^\pi - \hat{v}_1^\pi \mid Z_1, S_1)\|_2 \\
&\le \bar{\tau}_1 \sum_{t=1}^{T} \|\mathbb{E}(\mathcal{P}_{1:t-1}^\pi g_t \mid Z_1, S_1)\|_2 \quad \text{by the triangular inequality,}
\end{aligned}$$

where $\bar{\tau}_1 = \sup_{g_1 \in \mathcal{G}^{(1)}} \frac{\|g_1\|_2}{\|\mathbb{E}[g_1(W_1, S_1)|Z_1, S_1]\|_2}$. This indicates that we only need to separately bound the $\mathcal{L}^2$ norm of the projected one-step error defined as

$$\|\mathbb{E}[\mathcal{P}_{1:t-1}^\pi g_t \mid Z_1, S_1]\|_2 = \|\mathbb{E}[\mathcal{P}_{1:t-1}^\pi(\widehat{\mathcal{P}}_t^\pi - \mathcal{P}_t^\pi)(R_t + \hat{v}_{t+1}^\pi) \mid Z_1, S_1]\|_2, \tag{13}$$

for each $t = 1, \ldots, T$.

C.2.2 Error bounds for projected one-step error

To study the one-step $\mathcal{L}^2$ projected error of (13), for each $t = 1, \ldots, T$, motivated by (13), we sequentially define the following functions:

$$\begin{aligned}
g_t &\triangleq (\widehat{\mathcal{P}}_t^\pi - \mathcal{P}_t^\pi)(\hat{v}_{t+1}^\pi + R_t), \\
g_{t,t-1} &\triangleq \mathcal{P}_{t-1}^\pi g_t, \\
g_{t,t-2} &\triangleq \mathcal{P}_{t-2}^\pi g_{t,t-1} = \mathcal{P}_{t-2:t-1}^\pi g_t, \\
&\vdots \\
g_{t,1} &\triangleq \mathcal{P}_1^\pi g_{t,2} = \mathcal{P}_{1:t-1}^\pi g_t.
\end{aligned}$$

For each $1 \le t' < t$,

$$\begin{aligned}
&\|\mathbb{E}[g_{t,t'}(W_{t'}, S_{t'}) \mid Z_{t'}, S_{t'}]\|_2 = \|\mathbb{E}\{[\mathcal{P}_{t'}^\pi g_{t,t'+1}](W_{t'}, S_{t'}) \mid Z_{t'}, S_{t'}\}\|_2 \\
&= \|\mathbb{E}^{\pi_{t'}}[g_{t,t'+1}(W_{t'+1}, S_{t'+1}) \mid Z_{t'}, S_{t'}]\|_2 \\
&\le C_{t'+1,t'}^{(t)} \|\mathbb{E}[g_{t,t'+1}(W_{t'+1}, S_{t'+1}) \mid Z_{t'+1}, S_{t'+1}]\|_2,
\end{aligned}$$

where the local transition ill-posedness $C^{(t)}_{t'+1,t'}$ will be defined later in (17), and the second equality is due to

$$
\begin{aligned}
&\mathbb{E}\{[\mathcal{P}^{\pi}_{t'} g_{t,t'+1}](W_{t'}, S_{t'}) \mid Z_{t'}, S_{t'}\} \\
&= \mathbb{E}\left\{\sum_{a\in\mathcal{A}} \pi_{t'}(a \mid S_{t'})[\mathcal{P}_{t'} g_{t,t'+1}](W_{t'}, S_{t'}, A_{t'} = a) \mid Z_{t'}, S_{t'}\right\} \\
&= \sum_{a\in\mathcal{A}} \pi_{t'}(a \mid S_{t'})\mathbb{E}\left\{[\mathcal{P}_{t'} g_{t,t'+1}](W_{t'}, S_{t'}, A_{t'} = a) \mid Z_{t'}, S_{t'}\right\} \\
&= \sum_{a\in\mathcal{A}} \pi_{t'}(a \mid S_{t'})\mathbb{E}\left\{[\mathcal{P}_{t'} g_{t,t'+1}](W_{t'}, S_{t'}, A_{t'}) \mid Z_{t'}, S_{t'}, A_{t'} = a\right\} \\
&= \sum_{a\in\mathcal{A}} \pi_{t'}(a \mid S_{t'})\mathbb{E}\left\{g_{t,t'+1}(W_{t'+1}, S_{t'+1}) \mid Z_{t'}, S_{t'}, A_{t'} = a\right\} \text{ by Q-bridge} \\
&= \mathbb{E}^{\pi_{t'}}\left\{g_{t,t'+1}(W_{t'+1}, S_{t'+1}) \mid Z_{t'}, S_{t'}\right\}.
\end{aligned}
$$

Then by induction, we can show that

$$
\|\mathbb{E}[\mathcal{P}^{\pi}_{1:t-1} g_t \mid Z_1, S_1]\|_2 \le C^{(t)}_{2,1} \dots C^{(t)}_{t,t-1} \|\mathbb{E}[g_t \mid Z_t, S_t]\|_2
$$

Therefore,

$$
\begin{aligned}
\|v^{\pi}_1 - \hat{v}^{\pi}_1\|_2 &\le \bar{\tau}_1 \sum_{t=1}^{T} \|\mathbb{E}[\mathcal{P}^{\pi}_{1:t-1} g_t \mid Z_1, S_1]\|_2 \\
&\le \bar{\tau}_1 \sum_{t=1}^{T} C^{(t)}_{2,1} \dots C^{(t)}_{t,t-1} \|\mathbb{E}[g_t \mid Z_t, S_t]\|_2
\end{aligned}
$$

Then for each $t = 1, \dots, T$, we need to bound

$$
\begin{aligned}
\|\mathbb{E}[(\widehat{\mathcal{P}}^{\pi}_t - \mathcal{P}^{\pi}_t)(R_t + \hat{v}^{\pi}_{t+1}) \mid Z_t, S_t]\|_2 &\le \|(\widehat{\mathcal{P}}^{\pi}_t - \mathcal{P}^{\pi}_t)(R_t + \hat{v}^{\pi}_{t+1})\|_2 \\
&\le \|(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(R_t + \hat{v}^{\pi}_{t+1})\|_2 \|\pi_t/\pi^b_t\|_\infty \\
&\le \tau_t \|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(R_t + \hat{v}^{\pi}_{t+1})\|_2 \|\pi_t/\pi^b_t\|_\infty, \qquad (14)
\end{aligned}
$$

where τ_t is the local ill-posedness constant at step t, defined in (15).

Finally, we have

$$
\|v^{\pi}_1 - \hat{v}^{\pi}_1\|_2 \le \bar{\tau}_1 \sum_{t=1}^{T} \left\{\prod_{t'=1}^{t} C^{(t)}_{t',t'-1}\right\} \tau_t \|\pi_t/\pi^b_t\|_\infty \|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(R_t + \hat{v}^{\pi}_{t+1})\|_2.
$$

C.3 Proof of Theorem 6.2

For $t = T, \dots, 1$, we iteratively bound $\|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(R_t + \hat{v}^{\pi}_{t+1})\|_2$ by applying Lemma D.2, which depends on the critical radius of the space that contains $\hat{v}^{\pi}_{t+1}$ from the last step $t+1$. Then we give the bound of $\|\hat{v}^{\pi}_t\|^2_{\mathcal{G}^{(t)}}$, which will be used to calculate critical radii in next step.

C.3.1 One-step error bound

Start from $t = T$, $\hat{v}^{\pi}_{T+1} = v^{\pi}_{T+1} \triangleq 0$. By Lemma D.3, we have with probability at least $1 - 3\zeta$,

$$
\begin{aligned}
\|\text{proj}_T(\widehat{\mathcal{P}}_T - \mathcal{P}_T) R_T\|_2 &\lesssim \delta^{(T)}_n [1 + \|q^{\pi}_T\|^2_{\mathcal{H}^{(T)}}] \\
&\le \delta^{(T)}_n [1 + M_{\mathcal{H}}],
\end{aligned}
$$

and

$$
\|\hat{q}^{\pi}_T\|^2_{\mathcal{H}^{(T)}} = \|\widehat{\mathcal{P}}_T R_T\|^2_{\mathcal{H}^{(T)}} \le \|\mathcal{P}_T R_T\|^2_{\mathcal{H}^{(T)}} + C = \|q^{\pi}_T\|^2_{\mathcal{H}^{(T)}} + C \le 2M_{\mathcal{H}},
$$

by Assumption 1 (4) and we let $M_{\mathcal{H}} \ge C$.

Iteratively, at time $1 \le t < T$**,** by Lemma D.2, we have with probability at least $1 - 4\zeta$,

$$\begin{aligned}\|\mathrm{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)[R_t + \hat{v}^\pi_{t+1}]\|_2 &\lesssim (T-t+1)\delta_n^{(t)}[1 + \|\mathcal{P}_t\left(\frac{R_t + \hat{v}^\pi_{t+1}}{T-t+1}\right)\|^2_{\mathcal{H}^{(t)}}] \\ &\le (T-t+1)\delta_n^{(t)}[1 + (T-t+1)M_{\mathcal{H}}], \\ &\lesssim M_{\mathcal{H}}(T-t+1)^2\delta_n^{(t)},\end{aligned}$$

where the second inequality is due to Assumption 1 (2), $\|\mathcal{P}_t\left(\frac{R_t+\hat{v}^\pi_{t+1}}{T-t+1}\right)\|^2_{\mathcal{H}^{(t)}} \le \|\frac{\hat{q}^\pi_{t+1}}{T-t}\|^2_{\mathcal{H}^{(t+1)}} \le (T-t+1)M_{\mathcal{H}}$.

Also,

$$\begin{aligned}\|\frac{\hat{q}^\pi_t}{T-t+1}\|^2_{\mathcal{H}^{(t)}} = \|\widehat{\mathcal{P}}_t\left(\frac{R_t + \hat{v}^\pi_{t+1}}{T-t+1}\right)\|^2_{\mathcal{H}^{(t)}} &\le \|\mathcal{P}_t\left(\frac{R_t + \hat{v}^\pi_{t+1}}{T-t+1}\right)\|^2_{\mathcal{H}^{(t)}} + M_{\mathcal{H}} \\ &\le (T-t+2)M_{\mathcal{H}},\end{aligned}$$

where $\delta_n^{(t)} = \bar{\delta}_n^{(t)} + c_0\sqrt{\log(c_1/\zeta)/n}$, $c_0, c_1 > 0$, $\delta_n^{(t)}$ upper bounds the critical radii of $\mathcal{F}^{(t)}_{3M}(\mathcal{Z}_t \times \mathcal{S}_t \times \mathcal{A}_t)$, $\mathbf{\Omega}^{(t)}$ and $\mathbf{\Xi}^{(t)}$.

Since $\|\frac{\hat{v}^\pi_t}{T-t+1}\|^2_{\mathcal{G}^{(t)}} \le C_{\mathcal{G}}\|\frac{\hat{q}^\pi_t}{T-t+1}\|^2_{\mathcal{H}^{(t)}}$ by Assumption 1 (3), we have that $\|\frac{\hat{v}^\pi_t}{T-t+1}\|^2_{\mathcal{G}^{(t)}} \le C_{\mathcal{G}}\|\frac{\hat{q}^\pi_t}{T-t+1}\|^2_{\mathcal{H}^{(t)}} \le C_{\mathcal{G}}\|\frac{\hat{q}^\pi_{t+1}}{T-t}\|^2_{\mathcal{H}^{(t+1)}} \le C_{\mathcal{G}}(T-t+2)M_{\mathcal{H}}$. Therefore $\frac{\hat{v}^\pi_t}{T-t+1} \in \mathcal{G}^{(t)}_{C_{\mathcal{G}}(T-t+2)M_{\mathcal{H}}}$.

C.3.2 Combined Result

Finally, we replace ζ by $\zeta/(4T)$ and redefine $\delta_n^{(t)} = \bar{\delta}_n^{(t)} + c_0\sqrt{\log(c_1 T/\zeta)/n}$ for $t = 1, \ldots, T$, and consider the intersection of above events, we have with probability at least $1 - \zeta$,

$$\|\mathrm{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^\pi_{t+1} + R_t)\|_2 \lesssim M_{\mathcal{H}}(T-t+1)^2\delta_n^{(t)},$$

uniformly for all $1 \le t \le T$.

C.4 Localized ill-posedness τ_t and one-step transition ill-posedness $C^{(t)}_{t',t'-1}$

Localized ill-posedness. By Theorem 6.2 and (14), we have that with probability at least $1 - \zeta$,

$$\|\mathbb{E}[(\widehat{\mathcal{P}}^\pi_t - \mathcal{P}^\pi_t)(R_t + \hat{v}^\pi_{t+1}) \mid Z_t, S_t]\|_2 \lesssim \tau_t(T-t+1)^2 M_{\mathcal{H}}\delta_n^{(t)}\|\pi_t/\pi_t^b\|_\infty,$$

uniformly for all $1 \le t \le T$, where we define the local ill-posedness [Chen and Reiss, 2011]

$$\begin{aligned}\tau_t \triangleq \sup_{h \in \mathcal{H}^{(t)}} \frac{\|h\|_2}{\|\mathrm{proj}_t h\|_2} \text{ subject to } &\|\mathrm{proj}_t h\|_2 \lesssim (T-t+1)^2 M_{\mathcal{H}}\delta_n^{(t)}, \\ &\|h\|^2_{\mathcal{H}^{(t)}} \lesssim (T-t+1)^3 M_{\mathcal{H}}, \end{aligned} \tag{15}$$

where the bounds for $\|\mathrm{proj}_t h\|_2$ and $\|h\|^2_{\mathcal{H}^{(t)}}$ are adapted from above results in Appendix C.3.1.

We show that under further assumption on the joint distribution of (S_t, A_t, W_t, Z_t), for RKHS $\mathcal{H}^{(t)}$ with kernel $K_{\mathcal{H}^{(t)}}$, the local ill-posedness can be properly controlled. By Mercer's theorem with some regularity conditions, for any $h \in \mathcal{H}^{(t)}$, we have

$$h = \sum_{j=1}^{\infty} a_j e_j,$$

where $\{e_j : \mathcal{W} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ are the eigenfunctions of kernel $K_{\mathcal{H}^{(t)}}$ corresponding to nonincreasing eigenvalues $\{\lambda_j \triangleq \lambda_j^{\downarrow}(K_{\mathcal{H}^{(t)}})\}$. Then we have $\|h\|_2^2 = \sum_j a_j^2$ and $\|h\|^2_{\mathcal{H}} = \sum_j a_j^2/\lambda_j$.

$$\|\mathrm{proj}_t h\|_2^2 = \sum_{i,j} a_i a_j \mathbb{E}\left\{\mathbb{E}[e_i(W_t, S_t, A_t) \mid Z_t, S_t, A_t]\mathbb{E}[e_j(W_t, S_t, A_t) \mid Z_t, S_t, A_t]\right\}.$$

For $m \in \mathbb{N}_+$, let $I = \{1, \ldots, m\}$, $e_I = (e_1, \ldots, e_m)$ and $a_I = (a_1, \ldots, a_m)$ and define

$$\Gamma_m \triangleq \mathbb{E}\left\{\mathbb{E}[e_I(W_t, S_t, A_t) \mid Z_t, S_t, A_t]\mathbb{E}[e_I(W_t, S_t, A_t) \mid Z_t, S_t, A_t]^\top\right\}.$$

With same argument as Dikkala et al. [2020], we impose the assumption that $\lambda_{\min}(\Gamma_m) \geq \nu_m$ for all m almost surely, which means that the projected eigenfunctions are not strongly dependent. And we further assume that for all $i \leq m < j$,

$$|\mathbb{E}\left\{\mathbb{E}[e_i(W_t, S_t, A_t) \mid Z_t, S_t, A_t]\mathbb{E}[e_j(W_t, S_t, A_t) \mid Z_t, S_t, A_t]\right\}| \leq c\nu_m, \tag{16}$$

for some constant $c > 0$. This implies that the projection does not destroy the orthogonality for the first m eigenfunctions and eigenfunctions with indices larger than m too much. Then we can bound the local measure of ill-posedness as follow.

Lemma C.1 (Dikkala et al. [2020], Lemma 11)**.** Suppose that $\lambda_{\min}(\Gamma_m) \geq \nu_m$ and (16) holds for all $i \leq m < j$ and some constant $c > 0$. Then

$$[\tau^*(\delta, B)]^2 \triangleq \min_{m \in \mathbb{N}_+} \left\{ \delta^2/\nu_m + B\left(2c\sqrt{\sum_{i=1}^{\infty}\lambda_i}\sqrt{\sum_{j>m}\lambda_j} + \lambda_{m+1}\right)\right\}.$$

The optimal m_* is such that $\delta^2/\nu_m \asymp B\left(2c\sqrt{\sum_{i=1}^{\infty}\lambda_i}\sqrt{\sum_{j>m}\lambda_j} + \lambda_{m+1}\right)$.

• For a mild ill-posed case, if $\lambda_m \leq m^{-2\alpha_{\mathcal{H}}}$ for $\alpha_{\mathcal{H}} > 1/2$ and $\nu_m > m^{-2b}$ for $b > 0$, then $m_* \sim \left[\delta^2/B\right]^{-\frac{1}{2(\alpha_{\mathcal{H}}-1/2+b)}}$ and thus

$$\begin{aligned}
&\|(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^\pi_{t+1} + R_t)\|_2 \\
\lesssim\ & \tau_t^*(\|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^\pi_{t+1} + R_t)\|_2, \|(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^\pi_{t+1} + R_t)\|^2_{\mathcal{H}^{(t)}}) \\
\lesssim\ & \tau_t^*\left[(T-t+1)^2 M_{\mathcal{H}}\delta_n^{(t)}, (T-t+1)^3 M_{\mathcal{H}}\right] \\
\lesssim\ & (T-t+1)^{\frac{2(\alpha_{\mathcal{H}}-1/2)+3b}{(\alpha_{\mathcal{H}}-1/2)+2b}}[\delta_n^{(t)}]^{\frac{\alpha_{\mathcal{H}}-1/2}{\alpha_{\mathcal{H}}-1/2+b}}.
\end{aligned}$$

• For a severe ill-posed case, if $\lambda_m \leq m^{-2\alpha_{\mathcal{H}}}$ for $\alpha_{\mathcal{H}} > 1/2$ and $\nu_m \sim e^{-m^b}$ for $b > 0$, then $m_* \sim \left[\log\left(\frac{B}{\delta^2}\right)\right]^{\frac{1}{b}}$, by the same argument above,

$$\|(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^\pi_{t+1} + R_t)\|_2 \lesssim \left[\log\left(\frac{1}{(T-t+1)[\delta_n^{(t)}]^2}\right)\right]^{\frac{\alpha_{\mathcal{H}}-1/2}{2b}}(T-t+1)^{3/2}.$$

One-step transition ill-posedness. For each t, from $t' = t-1$ to $t' = 1$, we can recursively define a sequence of local transition ill-posedness as the following:

$$\begin{aligned}
C^{(t)}_{t'+1,t'} \triangleq & \sup_{g \in \mathcal{G}(W_{t'+1} \times S_{t'+1})} \frac{\|\mathbb{E}^{\pi_{t'}}[g(W_{t'+1}, S_{t'+1}) \mid Z_{t'}, S_{t'}]\|_2}{\|\mathbb{E}[g(W_{t'+1}, S_{t'+1}) \mid Z_{t'+1}, S_{t'+1}]\|_2} \\
\text{subject to } & \|\mathbb{E}[g(W_{t'+1}, S_{t'+1}) \mid Z_{t'+1}, S_{t'+1}]\|_2 \\
& \lesssim \tau_t(T-t+1)^2 M_{\mathcal{H}}\delta_n^{(t)}\|\pi_{t'}/\pi^b_{t'}\|_\infty \prod_{s=t'+1}^{t-1} C^{(t)}_{s+1,s}.
\end{aligned} \tag{17}$$

Then we have with probability at least $1-\zeta$,

$$\begin{aligned}
&\|\mathbb{E}[\mathcal{P}^\pi_{1:t-1}(\widehat{\mathcal{P}}^\pi_t - \mathcal{P}^\pi_t)(R_t + \hat{v}^\pi_{t+1}) \mid Z_t, S_t]\|_2 \\
&\leq \left\{\prod_{t'=1}^{t} C^{(t)}_{t',t'-1}\right\}\tau_t(T-t+1)^2 M_{\mathcal{H}}\delta_n^{(t)}\tau_t\|\pi_t/\pi^b_t\|_\infty,
\end{aligned}$$

uniformly for all $1 \leq t \leq T$.

C.5 Proofs of Theorems 6.3, B.1 and B.2

C.5.1 Decomposition of Off-Policy Value Estimation Error

Our objective is to give an upper bound of

$$\begin{aligned}
|\mathbb{E}v^\pi_1(W_1, S_1) - \mathbb{E}_n\hat{v}^\pi_1(W_1, S_1)| \leq\ & |\mathbb{E}v^\pi_1 - \mathbb{E}_n v^\pi_1| + |\mathbb{E}(v^\pi_1 - \hat{v}^\pi_1)| \\
& + |\mathbb{E}_n(v^\pi_1 - \hat{v}^\pi_1) - \mathbb{E}(v^\pi_1 - \hat{v}^\pi_1)| \\
=\ & (I) + (II) + (III),
\end{aligned}$$

For (I), by applying Hoeffding's inequality, we have with probability at least $1-\zeta/T$,

$$(I) = |\mathbb{E}v_1^\pi - \mathbb{E}_n v_1^\pi| \lesssim \|v_1^\pi\|_\infty \sqrt{\frac{\log(c_1 T/\zeta)}{n}} \lesssim T\sqrt{\frac{\log(c_1 T/\zeta)}{n}}.$$

For (II), obviously $(II) = |\mathbb{E}(v_1^\pi - \hat{v}_1^\pi)| \le \mathbb{E}|v_1^\pi - \hat{v}_1^\pi| \le \|v_1^\pi - \hat{v}_1^\pi\|_2$.

For (III), by applying Theorem 14.20 of Wainwright [2019], we have with probability at least $1-\zeta$,

$$(III) = |\mathbb{E}_n(v_1^\pi - \hat{v}_1^\pi) - \mathbb{E}(v_1^\pi - \hat{v}_1^\pi)| \lesssim \delta_n^{(0)}(\|v_1^\pi - \hat{v}_1^\pi\|_2 + T\delta_n^{(0)}),$$

where $\delta_n^{(0)} = \bar{\delta}_n^{(0)} + c_0\sqrt{\frac{\log(c_1 T/\zeta)}{n}}$, and $\bar{\delta}_n^{(0)}$ is the critical radius of $\mathcal{G}_{C_{\mathcal{G}}(T+1)M_{\mathcal{H}}}$.

The $\mathcal{L}^2$-error $\|v_1^\pi - \hat{v}^\pi\|_2$ in the upper bounds of (II) and (III) can be bound by combining Theorems 6.1 and 6.2.

C.5.2 Applying decomposition of OPE error

By Assumption 2, we can define $\texttt{trans-ill} = \max_{1\le t\le T}\exp\{a_t\zeta(\alpha_t)\}$ since $\prod_{t'=1}^t C_{t',t'-1}^{(t)} \le \exp\{a_t\zeta(\alpha_t)\}$, $1 \le t \le T$ are bounded by Corollary 6.3. Define $\texttt{ill}_{\max} = \bar{\tau}_1 \max_{1\le t\le T}\tau_t\|\pi_t/\pi_t^b\|_\infty$.

By applying Theorems 6.1 and 6.2, and crtical radii results in Example 1 – 3 in Appendix D.3, we have the following results:

For Theorem 6.3. With probability at least $1-\zeta$,

$$\|v_1^\pi - \hat{v}_1^\pi\|_2 \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2}\sqrt{\log(c_1 T/\zeta)}n^{-\frac{1}{2+\max\{1/\alpha_{\mathcal{H}},1/\alpha_{\mathcal{G}},1/\alpha_{\mathcal{F}}\}}}\log(n),$$

by Corollary 6.2 (1). Then by above decomposition, with probability at least $1-\zeta$,

$$|\mathcal{V}(\pi) - \hat{\mathcal{V}}(\pi)| \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2}\sqrt{\log(c_1 T/\zeta)}n^{-\frac{1}{2+\max\{1/\alpha_{\mathcal{H}},1/\alpha_{\mathcal{G}},1/\alpha_{\mathcal{F}}\}}}\log(n).$$

For Theorem B.1. With probability at least $1-\zeta$, with probability at least $1-\zeta$,

$$\|v_1^\pi - \hat{v}_1^\pi\|_2 \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2}\left\{\sqrt{\frac{\max_{1\le t\le T}\{\mathbb{V}(\mathcal{F}^{(t)}),\mathbb{V}(\mathcal{H}^{(t)}),\mathbb{V}(\mathcal{G}^{(t+1)})\}}{n}} + \sqrt{\frac{\log(T/\zeta)}{n}}\right\},$$

by Corollary 6.1. Then by above decomposition, with probability at least $1-\zeta$,

$$|\mathcal{V}(\pi) - \hat{\mathcal{V}}(\pi)| \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2}\left\{\sqrt{\frac{\max_{1\le t\le T}\{\mathbb{V}(\mathcal{F}^{(t)}),\mathbb{V}(\mathcal{H}^{(t)}),\mathbb{V}(\mathcal{G}^{(t+1)})\}}{n}} + \sqrt{\frac{\log(T/\zeta)}{n}}\right\}.$$

For Theorem B.2. With probability at least $1-\zeta$,

$$\|v_1^\pi - \hat{v}_1^\pi\|_2 \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2}\left\{\sqrt{\frac{(\log n)^{1/\min\{\beta_{\mathcal{H}},\beta_{\mathcal{G}},\beta_{\mathcal{F}}\}}}{n}} + \sqrt{\frac{\log(T/\zeta)}{n}}\right\},$$

by Corollary 6.2 (1). Then by above decomposition,

$$|\mathcal{V}(\pi) - \hat{\mathcal{V}}(\pi)| \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2}\left\{\sqrt{\frac{(\log n)^{1/\min\{\beta_{\mathcal{H}},\beta_{\mathcal{G}},\beta_{\mathcal{F}}\}}}{n}} + \sqrt{\frac{\log(T/\zeta)}{n}}\right\}.$$

For Corollary under mild and severe ill-posed cases. Under assumptions in main Theorem 6.3, by directly applying Lemma C.1, we have that

$$\|v_1^\pi - \hat{v}_1^\pi\|_2 \lesssim \bar{\tau}_1 \max_{1\le t\le T} \|\pi_t/\pi_t^b\|_\infty \times \texttt{trans-ill} \times \eta(n,T,\zeta,\alpha_{\mathcal{H}},\alpha_{\mathcal{F}},\alpha_{\mathcal{G}},b),$$
$$|\mathcal{V}(\pi) - \hat{\mathcal{V}}(\pi)| \lesssim \bar{\tau}_1 \max_{1\le t\le T} \|\pi_t/\pi_t^b\|_\infty \times \texttt{trans-ill} \times \eta(n,T,\zeta,\alpha_{\mathcal{H}},\alpha_{\mathcal{F}},\alpha_{\mathcal{G}},b),$$

where, for mild ill-posed case that $\nu_m \sim m^{-2b}$ for $b>0$:

$$\eta(n,T,\zeta,\alpha_{\mathcal{H}},\alpha_{\mathcal{F}},\alpha_{\mathcal{G}},b) = T^{\frac{7(\alpha_{\mathcal{H}}-1/2)+10b}{2(\alpha_{\mathcal{H}}-1/2)+4b}} \left(\sqrt{\log(c_1T/\zeta)}n^{-\frac{1}{2+\max\{1/\alpha_{\mathcal{H}},1/\alpha_{\mathcal{G}},1/\alpha_{\mathcal{F}}\}}}\log(n)\right)^{\frac{(\alpha_{\mathcal{H}}-1/2)}{(\alpha_{\mathcal{H}}-1/2)+2b}}.$$

for severe ill-posed case that $\nu_m \sim e^{-m^b}$ for $b>0$:

$$\eta(n,T,\zeta,\alpha_{\mathcal{H}},\alpha_{\mathcal{F}},\alpha_{\mathcal{G}},b) = \sum_{t=1}^{T}(T-t+1)^{3/2}\left\{\log\frac{n^{\frac{2}{2+\max\{1/\alpha_{\mathcal{H}},1/\alpha_{\mathcal{G}},1/\alpha_{\mathcal{F}}\}}}}{(\log n)^2(T-t+1)^2\log(T/\zeta)^2}\right\}^{-\frac{\alpha_{\mathcal{H}}-1/2}{2b}}.$$

D Auxiliary Lemmas

In this section, we provide some auxiliary lemmas which are needed to prove Theorem 4.1 – 6.3 and their proofs.

D.1 Lemmas For Identification

Lemma D.1 (Picard's Theorem, Theorem 15.16 of Kress [1989])**.** Given Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, a compact operator $K:\mathcal{H}_1\to\mathcal{H}_2$ and its adjoint operator $K^*:\mathcal{H}_2\to\mathcal{H}_1$, there exists a singular system $(\lambda_\nu,\phi_\nu,\psi_\nu)_{\nu=1}^\infty$ of K, with singular values $\{\lambda_\nu\}$ and orthogonal sequences $\{\phi_\nu\}\subset\mathcal{H}_1$ and $\{\psi_\nu\}\subset\mathcal{H}_2$ such that $K\phi_\nu=\lambda_\nu\psi_\nu$ and $K^*\psi_\nu=\lambda_\nu\phi_\nu$.

Given $g\in\mathcal{H}_2$, the Fredholm integral equation of the first kind $Kh=g$ is solvable if and only if

(a) $g\in\mathrm{Ker}(K^*)^\perp$ and

(b) $\sum_{\nu=1}^\infty \lambda_\nu^{-2}|\langle g,\psi_\nu\rangle|^2<\infty$,

where $\mathrm{Ker}(K^*)=\{h:K^*h=0\}$ is the null space of K^*, and $^\perp$ denotes the orthogonal complement to a set.

D.2 One-step estimation error

Consider the problem of estimating a function h that satisfying the conditional moment restriction

$$\mathbb{E}\{g(W)-h(X)\mid Z\}=0, \tag{18}$$

where $Z\in\mathcal{Z}$, $X\in\mathcal{X}$, $W\in\mathcal{W}$, $h\in\mathcal{H}\subset\{h\in\mathbb{R}^{\mathcal{X}}:\|h\|_\infty\le 1\}$, $g\in\mathcal{G}\subset\{g\in\mathbb{R}^{\mathcal{W}}:\|g\|_\infty\le 1\}$. Suppose that $h_g^*\in\mathcal{H}$ is the true h that satisfies the conditional moment restriction (18).

Suppose that we observe an i.i.d. sample $\{(W_i,X_i,Z_i)\}_{i=1}^n$ of sample size n drawn from an unknown distribution. Consider the minimax estimator

$$\hat{h}_g = \arg\min_{h\in\mathcal{H}}\sup_{f\in\mathcal{F}} \Psi_n(h,f,g) - \lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right) + \lambda\mu\|h\|_{\mathcal{H}}^2, \tag{19}$$

where $\Psi_n(h,f,g)=n^{-1}\sum_{i=1}^n\{g(W_i)-h(X_i)\}f(Z_i)$ with the population version $\Psi(h,f,g)=\mathbb{E}\{g(W)-h(X)\}f(Z)$ and $\lambda,\delta,\mu,U>0$ are tuning parameters.

Lemma D.2 ($\mathcal{L}^2$-error rate for minimax estimator)**.** Let $\mathcal{F}\subset\{f\in\mathbb{R}^{\mathcal{Z}}:\|f\|_\infty\le 1\}$ be a symmetric and star-convex set of test functions. Define $\delta=\delta_n+c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$ for some univeral constants $c_0,c_1>0$ and δ_n the upper bound of critical radii of $\mathcal{F}_{3U}$,

$$\mathbf{\Omega}=\left\{(x,w,z)\mapsto r(h_g^*(x)-g(w))f(z): g\in\mathcal{G}, f\in\mathcal{F}_{3U}, r\in[0,1]\right\},$$

and

$$\mathbf{\Xi} = \left\{(x,z) \mapsto r[h-h_g^*](x) f_\Delta^{L^2 B}(z); h \in \mathcal{H}, (h-h_g^*) \in \mathcal{H}_B, g \in \mathcal{G}, r \in [0,1]\right\},$$

where $f_\Delta^{L^2 B} = \arg\min_{f \in \mathcal{F}_{L^2 B}} \|f - \text{proj}_Z(h-h_g^*)\|_2$. Moreover, suppose that $\forall h \in \mathcal{H}$, $g \in \mathcal{G}$, $\|f_\Delta - \text{proj}_Z(h-h_g^*)\|_2 \le \eta_n \lesssim \delta_n$, where $f_\Delta \in \arg\inf_{f \in \mathcal{F}_{L^2\|h-h_g^*\|_{\mathcal{H}}^2}} \|f - \text{proj}_Z(h-h_g^*)\|_2$. If the tuning parameters satisfy $324 C_\lambda \delta^2/U \le \lambda \le 324 C'_\lambda \delta^2/U$ and $\mu \ge \frac{4}{3}L^2 + \frac{18(C_f+1)}{B}\frac{\delta^2}{\lambda}$, then with probability $1-4\zeta$,

$$\sup_{g \in \mathcal{G}} \|\text{proj}_Z(\hat{h}_g - h_g^*)\|_2 \lesssim (1 + \sup_{g \in \mathcal{G}} \|h_g^*\|_{\mathcal{H}}^2)\delta,$$

and for all $g \in \mathcal{G}$ uniformly,

$$\|\hat{h}_g\|_{\mathcal{H}}^2 \le C + \|h_g^*\|_{\mathcal{H}}^2.$$

The proof of Lemma D.2 is given in Appendix D.4.2.

Lemma D.3 (Dikkala et al. [2020], Theorem 1)**.** Consider the problem of estimating a function h that satisfies

$$\mathbb{E}\{Y - h(X) \mid Z\} = 0,$$

where $Z \in \mathcal{Z}$, $X \in \mathcal{X}$, $W \in \mathcal{W}$, $h \in \mathcal{H} \subset \{h \in \mathbb{R}^{\mathcal{X}} : \|h\|_\infty \le 1\}$, $|Y| \le 1$. Suppose that there exists $h^* \in \mathcal{H}$ that satisfies the conditional moment equation. Suppose that we observed an i.i.d. sample $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ of sample size n drawn from an unknown distribution. Consider the minimax estimator

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \sup_{f \in \mathcal{F}} \Phi_n(h,f) - \lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right) + \lambda\mu\|h\|_{\mathcal{H}}^2, \tag{20}$$

where $\Phi_n(h,f) = n^{-1}\sum_{i=1}^n \{Y_i - h(X_i)\} f(Z_i)$ with the population version $\Phi(h,f) = \mathbb{E}\{Y - h(X)\}f(Z)$ and $\lambda, \delta, \mu, U > 0$ are tuning parameters.

Let $\mathcal{F} \subset \{f \in \mathbb{R}^{\mathcal{Z}} : \|f\|_\infty \le 1\}$ be a symmetric and star-convex set of test functions. Define $\delta = \delta_n + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$ for some univeral constants $c_0, c_1 > 0$ and δ_n the upper bound of critical radii of $\mathcal{F}_{3U}$ and

$$\bar{\mathbf{\Xi}} = \left\{(x,z) \mapsto r[h-h^*](x) f_\Delta^{L^2 B}(z); h \in \mathcal{H}, (h-h^*) \in \mathcal{H}_B, r \in [0,1]\right\},$$

where $f_\Delta^{L^2 B} = \arg\min_{f \in \mathcal{F}_{L^2 B}} \|f - \text{proj}_Z(h-h^*)\|_2$. Moreover, suppose that $\forall h \in \mathcal{H}$, $\|f_\Delta - \text{proj}_Z(h-h^*)\|_2 \le \eta_n \lesssim \delta_n$, where $f_\Delta \in \arg\inf_{f \in \mathcal{F}_{L^2\|h-h^*\|_{\mathcal{H}}^2}} \|f - \text{proj}_Z(h-h^*)\|_2$. Suppose tuning parameters satistying $324 C_\lambda \delta^2/U \le \lambda \le 324 C'_\lambda \delta^2/U$ and $\mu \ge \frac{4}{3}L^2 + \frac{18(C_f+1)}{B}\frac{\delta^2}{\lambda}$. Then with probability $1-3\zeta$,

$$\|\text{proj}_Z(\hat{h} - h^*)\|_2 \lesssim (1 + \|h^*\|_{\mathcal{H}}^2)\delta,$$

and

$$\|\hat{h}\|_{\mathcal{H}}^2 \le C + \|h^*\|_{\mathcal{H}}^2.$$

D.3 Critical radii and local Rademacher complexity

In this section we list several ways to bound the critical radii of $\mathcal{F}_{3U}$, $\mathbf{\Omega}$ and $\mathbf{\Xi}$ for Lemmas in Appendix D.2. We restrict $\mathcal{G} = \mathcal{G}_D = \{g \in \mathcal{G} : \|g\|_{\mathcal{G}}^2 \le D\}$ for some $D > 0$ in this section.

D.3.1 Local Rademacher complexity bound by entropy integral

In this subsection, we introduce an entropy integral based approach to bound the local Rademacher complexity and critical radii. Similar to local Rademacher complexity, for a star-shaped and b-uniformly bounded function class $\mathcal{F}$, the *local empirical Rademacher complexity*, a data-dependent quantity, is defined by

$$\widehat{R}_n(\delta; \mathcal{F}) \triangleq \mathbb{E}\left[\sup_{f \in \mathcal{F}, \|f\|_n \le \delta} |\frac{1}{n}\epsilon_i f(X_i)| \,\middle|\, \{X_i\}_{i=1}^n\right]$$

where $\{\epsilon_i\}_{i=1}^n$ are i.i.d. Rademacher variables. The *empirical critical radius* $\hat{\delta}_n$ is the smallest positive solution to

$$\widehat{R}_n(\delta) \leq \frac{\delta^2}{b}. \tag{21}$$

Wainwright [2019, Proposition 14.25] gives the relationship that with probability at least $1-\zeta$,

$$\delta_n \leq \mathcal{O}(\hat{\delta}_n + \sqrt{\frac{\log(1/\zeta)}{n}}).$$

Therefore, we can study the critical radius δ_n by empirical critical radius $\hat{\delta}_n$.

Given a space $\mathcal{G}$, an *empirical ϵ-covering* of $\mathcal{G}$ is defined as any function class $\mathcal{G}^\epsilon$ such that for all $g \in \mathcal{G}$, $\inf_{g_\epsilon \in \mathcal{G}^\epsilon} \|g_\epsilon - g\|_n \leq \epsilon$. Denote the smallest empirical ϵ-covering of $\mathcal{G}$ by $N_n(\epsilon, \mathcal{G})$. Let $\mathbb{B}_n(\delta; \mathcal{G}) \triangleq \{g \in \mathcal{G} : \|g\|_n \leq \delta\}$. Then we have the following Lemma to bound the empirical critical radius by Dudley's entropy integral.

Lemma D.4. [Wainwright, 2019, Corollary 14.3] The empirical critical inequality (21) is satisfied for any $\delta > 0$ such that

$$\frac{64}{\sqrt{n}} \int_{\frac{\delta^2}{2b}}^{\delta} \sqrt{\log N_n(t, \mathbb{B}_n(\delta, \mathcal{G}))} \mathrm{d}t \leq \frac{\delta^2}{b}.$$

Lemma D.5. Suppose that $\|h_g^*\|_{\mathcal{H}}^2 \leq A\|g\|_{\mathcal{G}}^2$ for all $g \in \mathcal{G}$, so that $\|h_g^*\|_{\mathcal{H}}^2 \leq AD$. Let $\widehat{\delta}_n > 0$ satisfy the inequality

$$\frac{64}{\sqrt{n}} \int_{\frac{\delta^2}{2}}^{4\delta} \sqrt{\log N_n(t, \mathrm{star}\{\mathcal{F}_{3U \vee L^2 B}\}) + \log N_n(t, \mathrm{star}\{\mathcal{H}_{AD \vee B}\}) + \log N_n(t, \mathrm{star}\{\mathcal{G}_D\})} \mathrm{d}t \leq \delta^2.$$

Then with probability $1-\zeta$, we have $\delta_n \leq \mathcal{O}(\widehat{\delta}_n + \sqrt{\frac{\log(1/\zeta)}{n}})$, where δ_n is the maximum critical radii of $\mathcal{F}_{3U}$, $\mathbf{\Omega}$ and $\mathbf{\Xi}$, with

$$\mathbf{\Omega} = \left\{(x, w, z) \mapsto r(h_g^*(x) - g(w)) f(z) : g \in \mathcal{G}_D, f \in \mathcal{F}_{3U}, r \in [0, 1]\right\}.$$

The proof of Lemma D.5 is given in Appendix D.4.3.

Example 1 (Critical radii for VC subspaces)**.** If star shaped $\mathcal{F}, \mathcal{H}$ and $\mathcal{G}$ are VC subspaces with VC dimensions $\mathbb{V}(\mathcal{F}), \mathbb{V}(\mathcal{H})$ and $\mathbb{V}(\mathcal{G})$, respectively, then $\log N_n(t, \mathcal{F}) + \log N_n(t, \mathcal{H}) + \log N_n(t, \mathcal{G}) \lesssim [\mathbb{V}(\mathcal{F}) + \mathbb{V}(\mathcal{H}) + \mathbb{V}(\mathcal{G})] \log(1/t) \lesssim \max\{\mathbb{V}(\mathcal{F}), \mathbb{V}(\mathcal{H}), \mathbb{V}(\mathcal{G})\} \log(1/t)$. By Lemma D.4 and Lemma D.5, we have with probability at least $1-\zeta$, $\delta_n \lesssim \sqrt{\frac{\max\{\mathbb{V}(\mathcal{F}), \mathbb{V}(\mathcal{H}), \mathbb{V}(\mathcal{G})\}}{n}} + \sqrt{\frac{\log(1/\zeta)}{n}}$, where the δ_n is defined in Lemma D.5.

D.3.2 Local Rademacher complexity bound for RKHSs

Lemma D.6 (Critical radii for RKHSs, Corollary 14.5 of Wainwright [2019])**.** Let $\mathcal{F}_B = \left\{f \in \mathcal{F} \mid \|f\|_{\mathcal{F}}^2 \leq B\right\}$ be the B-ball of a RKHS $\mathcal{F}$. Suppose that $K_{\mathcal{F}}$ is the reproducing kernel of $\mathcal{F}$ with eigenvalues $\{\lambda_j^{\downarrow}(K_{\mathcal{F}})\}_{j=1}^{\infty}$ sorted in a decreasing order. Then the localized population Rademacher complexity is upper bounded by

$$\mathcal{R}_n(\mathcal{F}_B, \delta) \leq \sqrt{\frac{2B}{n}} \sqrt{\sum_{j=1}^{\infty} \min\left\{\lambda_j^{\downarrow}(K_{\mathcal{F}}), \delta^2\right\}}.$$

Lemma D.7 (Critical radii for $\mathbf{\Omega}$ and $\mathbf{\Xi}$ when $\mathcal{H}, \mathcal{F}, \mathcal{G}$ are RKHSs)**.** Suppose that $\mathcal{F}$,$\mathcal{H}$, and $\mathcal{G}$ are RKHSs endowed with reproducing kernels $K_{\mathcal{F}}$, $K_{\mathcal{H}}$, and $K_{\mathcal{G}}$ with decreasingly sorted eigenvalues $\left\{\lambda_j^{\downarrow}(K_{\mathcal{F}})\right\}_{j=1}^{\infty}$, $\left\{\lambda_j^{\downarrow}(K_{\mathcal{H}})\right\}_{j=1}^{\infty}$, and $\left\{\lambda_j^{\downarrow}(K_{\mathcal{G}})\right\}_{j=1}^{\infty}$, respectively. Then

$$\mathcal{R}_n(\mathbf{\Xi}, \delta) \leq LB\sqrt{\frac{2}{n}} \sqrt{\sum_{i,j=1}^{\infty} \min\left\{\lambda_i^{\downarrow}(K_{\mathcal{H}})\lambda_j^{\downarrow}(K_{\mathcal{F}}), \delta^2\right\}}, \quad \text{and}$$

$$\mathcal{R}_n(\mathbf{\Omega}, \delta) \leq \sqrt{D}(1+\sqrt{A})\sqrt{\frac{12U}{n}}\sqrt{\sum_{i,j=1}^{\infty} \min\left\{[(\lambda_i^{\downarrow}(K_{\mathcal{H}}) + \lambda_i^{\downarrow}(K_{\mathcal{G}})]\lambda_j^{\downarrow}(K_{\mathcal{F}}), \delta^2\right\}}.$$

The proof of Lemma D.7 is given in Appendix D.4.4.

We give the following two examples as directly applications of Lemma D.6 and D.7.

Example 2 (Critical radii for RKHSs endowed with kernels with polynomial decay)**.** With the same conditions in Lemma D.7, when $\lambda_j^{\downarrow}(K_{\mathcal{F}}) \leq cj^{-2\alpha_{\mathcal{F}}}$, $\lambda_j^{\downarrow}(K_{\mathcal{G}}) \leq cj^{-2\alpha_{\mathcal{G}}}$, $\lambda_j^{\downarrow}(K_{\mathcal{H}}) \leq cj^{-2\alpha_{\mathcal{H}}}$, where constant $\alpha_{\mathcal{H}}, \alpha_{\mathcal{G}}, \alpha_{\mathcal{F}} > 1/2, c > 0$, then by Krieg [2018] we have the upper bound of critical radii of $\mathcal{F}_{3U}$, $\mathbf{\Omega}$ and $\mathbf{\Xi}$ satisfies

$$\delta_n \lesssim \max\{\sqrt{B}, LB, \sqrt{6DU}(1+\sqrt{A})\} n^{-\frac{1}{2+\max\{1/\alpha_{\mathcal{F}}, 1/\alpha_{\mathcal{G}}, 1/\alpha_{\mathcal{H}}\}}} \log(n).$$

Example 3 (Critical radii for RKHSs endowed with kernels with exponential decay)**.** With the same conditions in Lemma D.7, when $\lambda_j^{\downarrow}(K_{\mathcal{H}}) \leq a_1 e^{-a_2 j^{\beta_{\mathcal{H}}}}$, $\lambda_j^{\downarrow}(K_{\mathcal{G}}) \leq a_1 e^{-a_2 j^{\beta_{\mathcal{G}}}}$ and $\lambda_j^{\downarrow}(K_{\mathcal{F}}) \leq a_1 e^{-a_2 j^{\beta_{\mathcal{F}}}}$, for constants $a_1, a_2, \beta_{\mathcal{H}}, \beta_{\mathcal{G}}, \beta_{\mathcal{F}} > 0$, then we have the upper bound of critical radii of $\mathcal{F}_{3U}$, $\mathbf{\Omega}$ and $\mathbf{\Xi}$ satisfies

$$\delta_n \lesssim \max\{\sqrt{B}, LB, \sqrt{6DU}(1+\sqrt{A})\}\sqrt{\frac{(\log n)^{1/\min\{\beta_{\mathcal{F}}, \beta_{\mathcal{G}}, \beta_{\mathcal{H}}\}}}{n}}.$$

D.4 Proof of Lemmas

D.4.1 Proof of Lemma C.1

Proof. For any $m \in \mathbb{N}_+$,

$$\begin{aligned}
\|\text{proj}_t h\|_2^2 &= a_I^\top \Gamma_m a_I + 2\sum_{i\leq m<j} a_i a_j \mathbb{E}\left\{\mathbb{E}[e_i(W_t, S_t, A_t) \mid Z_t, S_t, A_t]\mathbb{E}[e_j(W_t, S_t, A_t) \mid Z_t, S_t, A_t]\right\} \\
&\quad + \mathbb{E}\left(\sum_{j>m} a_j \mathbb{E}[e_j(W_t, S_t, A_t) \mid Z_t, S_t, A_t]\right) \\
&\geq a_I^\top \Gamma_m a_I - 2\sum_{i\leq m<j} |a_i a_j| \mathbb{E}\left\{\mathbb{E}[e_i(W_t, S_t, A_t) \mid Z_t, S_t, A_t]\mathbb{E}[e_j(W_t, S_t, A_t) \mid Z_t, S_t, A_t]\right\} \\
&\geq a_I^\top \Gamma_m a_I - 2\sum_{i\leq m<j} |a_i a_j| c\nu_m \\
&\geq \nu_m \|a_I\|_2^2 - 2c\nu_m \sum_{i\leq m} |a_i| \sum_{j>m} |a_j| \\
&\geq \nu_m \|a_I\|_2^2 - 2c\nu_m \sqrt{\sum_{i\leq m} \lambda_i} \sqrt{\sum_{i\leq m} \frac{|a_i|^2}{\lambda_i}} \sqrt{\sum_{j>m} \lambda_j} \sqrt{\sum_{j>m} \frac{|a_j|^2}{\lambda_j}} \\
&\geq \nu_m \|a_I\|_2^2 - 2c\nu_m B \sqrt{\sum_{i=1}^{\infty} \lambda_i} \sqrt{\sum_{j>m} \lambda_j}, \quad \text{since } \sum_{j=1}^{\infty} \frac{|a_j|^2}{\lambda_j} \leq B.
\end{aligned}$$

Therefore, $\|h\|_2^2 \leq \|a_I\|^2 + B\lambda_{m+1} \leq \|\text{proj}_t h\|_2^2/\nu_m + 2cB\sqrt{\sum_{i=1}^{\infty}\lambda_i}\sqrt{\sum_{j>m}\lambda_j} + B\lambda_{m+1}$. Because $\|\text{proj}_t h\|_2 \leq \delta$, by taking minimum over $m \in \mathbb{N}_+$, we have that

$$[\tau^*(\delta, B)]^2 \leq \min_{m\in\mathbb{N}_+}\left\{\delta^2/\nu_m + B\left(2c\sqrt{\sum_{i=1}^{\infty}\lambda_i}\sqrt{\sum_{j>m}\lambda_j} + \lambda_{m+1}\right)\right\}.$$

□

D.4.2 Proof of Lemma D.2

Proof. Let $\mathcal{H}_B = \left\{h \in \mathcal{H} : \|h\|_{\mathcal{H}}^2 \le B\right\}$ and $\mathcal{F}_U = \left\{f \in \mathcal{F} : \|f\|_{\mathcal{F}}^2 \le U\right\}$. Moreover, let

$$\Psi^{\lambda}(f,g,h) = \Psi(h,f,g) - \lambda\left(\frac{2}{3}\|f\|_{\mathcal{F}}^2 + \frac{U}{2\delta^2}\|f\|_2^2\right), \quad \text{and}$$
$$\Psi_n^{\lambda}(f,g,h) = \Psi_n(h,f,g) - \lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right).$$

We first study the relationship between the empirical penalty $\lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right)$ and population penalty $\lambda\left(\frac{2}{3}\|f\|_{\mathcal{F}}^2 + \frac{U}{2\delta^2}\|f\|_2^2\right)$. Let $\delta = \delta_n + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$, where δ_n upper bounds the critical radius of $\mathcal{F}_{3U}$ and c_0, c_1 are universal constants, by Theorem 14.1 of Wainwright [2019], with probablity $1-\zeta$, uniformly for any $f \in \mathcal{F}$, we have

$$\left|\|f\|_n^2 - \|f\|_2^2\right| \le \frac{1}{2}\|f\|_2^2 + \delta^2 \max\left\{1, \frac{\|f\|_{\mathcal{F}}^2}{3U}\right\}, \text{and thus} \tag{22}$$

$$\begin{aligned}
\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2 &\ge \|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\left[\frac{1}{2}\|f\|_2^2 - \delta^2 \max\left\{1, \frac{\|f\|_{\mathcal{F}}^2}{3U}\right\}\right] \\
&\ge \|f\|_{\mathcal{F}}^2 + \frac{U}{2\delta^2}\|f\|_2^2 - \max\left\{U, \frac{1}{3}\|f\|_{\mathcal{F}}^2\right\} \\
&\ge \frac{2}{3}\|f\|_{\mathcal{F}}^2 + \frac{U}{2\delta^2}\|f\|_2^2 - U.
\end{aligned} \tag{23}$$

In the following proof, we obtain the error rate of the uniform projected RMSE $\sup_{g\in\mathcal{G}}\|\text{proj}_Z(\hat{h}_g - h_g^*)\|_2$ by combinding upper and lower bounds of the sup-loss

$$\sup_{f\in\mathcal{F}} \Psi_n(\hat{h}_g, f, g) - \Psi_n(h_g^*, f, g) - 2\lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right). \tag{24}$$

Upper bound of sup-loss (24). By a simple decomposition of $\Psi_n^{\lambda}(h,f,g)$, we have

$$\begin{aligned}
\Psi_n^{\lambda}(h,f,g) &= \Psi_n(h,f,g) - \Psi_n(h_g^*,f,g) + \Psi_n(h_g^*,f,g) - \lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right) \\
&\ge \Psi_n(h,f,g) - \Psi_n(h_g^*,f,g) - 2\lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right) \\
&\quad + \inf_{f\in\mathcal{F}}\left\{\Psi_n(h_g^*,f,g) + \lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right)\right\} \\
&= \Psi_n(h,f,g) - \Psi_n(h_g^*,f,g) - 2\lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right) \\
&\quad - \sup_{f\in\mathcal{F}} \Psi_n^{\lambda}(h_g^*,f,g), \quad \text{since } \mathcal{F} \text{ is symmetric about } 0.
\end{aligned}$$

Taking $\sup_{f\in\mathcal{F}}$ on both sides and picking $h \leftarrow \hat{h}_g$ yields the basic inequality:

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \Psi_n(\hat{h}_g, f, g) - \Psi_n(h_g^*, f, g) - 2\lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right) \\
\le &\sup_{f\in\mathcal{F}} \Psi_n^{\lambda}(h_g^*, f, g) + \sup_{f\in\mathcal{F}} \Psi_n^{\lambda}(\hat{h}_g, f, g) \\
\le &2\sup_{f\in\mathcal{F}} \Psi_n^{\lambda}(h_g^*, f, g) + \lambda\mu(\|h_g^*\|_{\mathcal{H}}^2 - \|\hat{h}_g\|_{\mathcal{H}}^2),
\end{aligned} \tag{25}$$

where the last inequality is given by the definition of $\hat{h}_g$ in (19). Now it suffices to obtain the upper bound of $\sup_{f\in\mathcal{F}} \Psi_n^{\lambda}(h_g^*, f, g)$ uniformly over $g \in \mathcal{G}$.

For upper bound of $\sup_{f\in\mathcal{F}}\Psi_n^\lambda(h_g^*,f,g)$**.** By the assumption that $\|g\|_\infty\le 1$, $\|h\|_\infty\le 1$ and $\|f\|_\infty\le 1$, we have $\|\frac{1}{2}\{g(W)-h(X)\}f(Z)\|_\infty\le 1$. Then we apply Lemma 11 of Foster and Syrgkanis [2019], with $\mathcal{L}_{\frac{1}{2}(g-h_g^*)f}=\frac{1}{2}(g-h_g^*)f$. Let δ_n be the upper bound of critical radii of $\mathbf{\Omega}$. By choosing $\delta=\delta_n+c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$, we have with probability $1-\zeta$, uniformly for any $f\in\mathcal{F}_{3U}$ and $g\in\mathcal{G}$:

$$\begin{aligned}&\frac{1}{2}\left|\left\{\Psi_n(h_g^*,f,g)-\Psi_n(h_g^*,0,g)\right\}-\left\{\Psi(h_g^*,f,g)-\Psi(h_g^*,0,g)\right\}\right|\\&\le 18\delta\left(\|\frac{1}{2}(g-h_g^*)f\|_2+\delta\right)\\&\le 18\delta\left(\|f\|_2+\delta\right),\end{aligned}$$

where, by definition, $\Psi_n(h_g^*,0,g)=\Psi(h_g^*,0,g)=0$. If $\|f\|_{\mathcal{F}}^2\ge 3U$, applying the above inequality with $f\leftarrow f\sqrt{3U}/\|f\|_{\mathcal{F}}$, we have with probability $1-\zeta$, for all $f\in\mathcal{F}$ and $g\in\mathcal{G}$:

$$\begin{aligned}\left|\Psi_n(h_g^*,f,g)-\Psi(h_g^*,f,g)\right|&\le 36\delta\left\{\|f\|_2+\max\left\{1,\frac{\|f\|_{\mathcal{F}}}{\sqrt{3U}}\right\}\delta\right\}\\&\le 36\delta\left\{\|f\|_2+\left(1+\frac{\|f\|_{\mathcal{F}}}{\sqrt{3U}}\right)\delta\right\}.\end{aligned}\tag{26}$$

By using (26) and (23) sequentially, we have with probability $1-2\zeta$, for all $f\in\mathcal{F}$ and $g\in\mathcal{G}$:

$$\begin{aligned}&\Psi_n^\lambda(h_g^*,f,g)=\Psi_n(h_g^*,f,g)-\lambda\left(\|f\|_{\mathcal{F}}^2+\frac{U}{\delta^2}\|f\|_n^2\right)\\\le&\Psi(h_g^*,f,g)+36\delta\left\{\|f\|_2+\left(1+\frac{\|f\|_{\mathcal{F}}}{\sqrt{3U}}\right)\delta\right\}-\lambda\left(\|f\|_{\mathcal{F}}^2+\frac{U}{\delta^2}\|f\|_n^2\right)\\\le&\Psi(h_g^*,f,g)+36\delta\left\{\|f\|_2+\left(1+\frac{\|f\|_{\mathcal{F}}}{\sqrt{3U}}\right)\delta\right\}-\lambda\left(\frac{2}{3}\|f\|_{\mathcal{F}}^2+\frac{U}{2\delta^2}\|f\|_2^2\right)+\lambda U\\=&\Psi^{\lambda/2}(h_g^*,f,g)+36\delta^2+\lambda U+\left(36\delta\|f\|_2-\frac{\lambda U}{4\delta^2}\|f\|_2^2\right)+\left(\frac{36\delta}{\sqrt{3U}}\delta\|f\|_{\mathcal{F}}-\frac{\lambda}{3}\|f\|_{\mathcal{F}}^2\right).\end{aligned}$$

With the assumption that $\lambda\ge 324C_\lambda\delta^2/U$, by completing squares, we have

$$36\delta\|f\|_2-\frac{\lambda U}{4\delta^2}\|f\|_2^2\le\frac{(36\delta)^2}{4\frac{\lambda U}{4\delta^2}}\le\frac{4\delta^2}{C_\lambda},\quad\text{and}$$

$$\frac{36\delta^2}{\sqrt{3U}}\|f\|_{\mathcal{F}}-\frac{\lambda}{3}\|f\|_{\mathcal{F}}^2\le\frac{324\delta^4}{\lambda U}\le\frac{\delta^2}{C_\lambda}.$$

Therefore, with probability $1-2\zeta$, for all $f\in\mathcal{F}$ and $g\in\mathcal{G}$:

$$\Psi_n^\lambda(h_g^*,f,g)\le\Psi^{\lambda/2}(h_g^*,f,g)+\lambda U+\left(36+\frac{5}{C_\lambda}\right)\delta^2.\tag{27}$$

Now we go back to (25). By applying two upper bounds above, we have with probability $1-2\zeta$, uniformly for all $g\in\mathcal{G}$:

$$\begin{aligned}&\sup_{f\in\mathcal{F}}\Psi_n(\hat{h}_g,f,g)-\Psi_n(h_g^*,f,g)-2\lambda\left(\|f\|_{\mathcal{F}}^2+\frac{U}{\delta^2}\|f\|_n^2\right)\\\le&2\sup_{f\in\mathcal{F}}\Psi_n^\lambda(h_g^*,f,g)+\lambda\mu(\|h_g^*\|_{\mathcal{H}}^2-\|\hat{h}_g\|_{\mathcal{H}}^2)\\\le&2\sup_{f\in\mathcal{F}}\Psi^{\lambda/2}(h_g^*,f,g)+2\lambda U+(72+10/C_\lambda)\delta^2+\lambda\mu(\|h_g^*\|_{\mathcal{H}}^2-\|\hat{h}_g\|_{\mathcal{H}}^2)\\=&2\lambda U+(72+10/C_\lambda)\delta^2+\lambda\mu(\|h_g^*\|_{\mathcal{H}}^2-\|\hat{h}_g\|_{\mathcal{H}}^2),\end{aligned}\tag{28}$$

where $\sup_{f\in\mathcal{F}}\Psi^{\lambda/2}(h_g^*,f,g^*)=0$ since $\mathbb{E}\left\{g(W)-h_g^*(X)\right\}f(Z)=0$.

We can also obtain the upper bound of $\|\hat{h}_g\|_{\mathcal{H}}$ by (28). By choosing $f = 0$, the LHS of (28) is 0, so the supremum of LHS is nonnegative. Then with probability $1 - 2\zeta$,

$$\begin{aligned}\|\hat{h}_g\|_{\mathcal{H}}^2 &\le \frac{1}{\lambda\mu}\Big\{2\lambda U + (72 + 10/C_\lambda)\delta^2\Big\} + \|h_g^*\|_{\mathcal{H}}^2 \\ &\le \frac{36C_\lambda + 3 + \frac{5}{9C_\lambda}}{\frac{24C_\lambda L^2}{U} + \frac{C_f+1}{B}} + \|h_g^*\|_{\mathcal{H}}^2. \end{aligned} \tag{29}$$

Lower bound of sup-loss (24). For any h and g, by our assumption that $\|f_\Delta - \text{proj}_Z(h - h_g^*)\|_2 \le \eta_n$, where $f_\Delta = \arg\min_{f \in \mathcal{F}_{L^2\|h-h_g^*\|_{\mathcal{H}}^2}} \|f - \text{proj}_Z(h - h_g^*)\|_2$. Let $\widehat{\Delta}_g = \hat{h}_g - h_g^*$, and $f_{\widehat{\Delta}_g} = \arg\min_{f \in \mathcal{F}_{L^2\|\hat{h}_g - h_g^*\|_{\mathcal{H}}}} \|f - \text{proj}_Z(h - h_g^*)\|_2$.

If $\|f_{\widehat{\Delta}_g}\|_2 < C_f\delta$, then by the triangle inequality, we have

$$\|\text{proj}_Z(\hat{h}_g - h_g^*)\|_2 \le \|f_{\widehat{\Delta}_g}\|_2 + \|f_{\widehat{\Delta}_g} - \text{proj}_Z(\hat{h}_g - h_g^*)\|_2 \le C_f\delta + \eta_n.$$

If $\|f_{\widehat{\Delta}_g}\|_2 \ge C_f\delta$, let $r = \frac{C_f\delta}{2\|f_{\widehat{\Delta}_g}\|_2} \in [0, 1/2]$. By star-convexity, $r f_{\widehat{\Delta}_g} \in \mathcal{F}_{L^2\|\hat{h}_g - h_g^*\|_{\mathcal{H}}^2}$. Therefore, for any $g \in \mathcal{G}$,

$$\begin{aligned}&\sup_{f\in\mathcal{F}} \Psi_n(\hat{h}_g, f, g) - \Psi_n(h_g^*, f, g) - 2\lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right) \\ &\qquad \ge \underbrace{r\Big\{\Psi_n(\hat{h}_g, f_{\widehat{\Delta}_g}, g) - \Psi_n(h_g^*, f_{\widehat{\Delta}_g}, g)\Big\}}_{(I)} - \underbrace{2\lambda r^2\left(\|f_{\widehat{\Delta}_g}\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f_{\widehat{\Delta}_g}\|_n^2\right)}_{(II)}.\end{aligned}$$

For (II): We have

$$\begin{aligned}(II) &= r^2\left(\|f_{\widehat{\Delta}_g}\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f_{\widehat{\Delta}_g}\|_n^2\right) \le \frac{1}{4}\|f_{\widehat{\Delta}_g}\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}r^2\|f_{\widehat{\Delta}_g}\|_n^2 \\ &\le \frac{1}{4}\|f_{\widehat{\Delta}_g}\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}r^2\left(\frac{3}{2}\|f_{\widehat{\Delta}_g}\|_2^2 + \delta^2 + \delta^2\frac{\|f_{\widehat{\Delta}_g}\|_{\mathcal{F}}^2}{3U}\right) \text{ with probability } 1-\zeta \text{ by (22)} \\ &\le \frac{1}{3}\|f_{\widehat{\Delta}_g}\|_{\mathcal{F}}^2 + \frac{1}{4}U + \frac{3}{8}C_f^2U \text{ by definition of } r \\ &\le \frac{1}{3}L^2\|\hat{h}_g - h_g^*\|_{\mathcal{H}}^2 + (\frac{1}{4} + \frac{3}{8}C_f^2)U \text{ since } f_{\widehat{\Delta}_g} \in \mathcal{F}_{L^2\|\hat{h}_g - h_g^*\|_{\mathcal{H}}^2}.\end{aligned}$$

For (I): Note that $\Psi_n(h, f, g) - \Psi_n(h_g^*, f, g) = \frac{1}{n}\sum_{i=1}^n [h - h_g^*](X_i)f(Z_i)$. We apply Lemma 11 of Foster and Syrgkanis [2019], with $\mathcal{L}_{(h-h_g^*)f} = (h - h_g^*)f$. Recall that

$$\Xi = \Big\{(x, z) \mapsto r[h - h_g^*](x) f_\Delta^{L^2B}(z) : h \in \mathcal{H}, (h - h_g^*) \in \mathcal{H}_B, g \in \mathcal{G}, r \in [0, 1]\Big\},$$

where $f_\Delta^{L^2B} = \arg\min_{f\in\mathcal{F}_{L^2B}} \|f - \text{proj}_Z(h - h_g^*)\|_2$. Since δ_n upper bounds critical radius of Ξ, we have with probability $1 - \zeta$, uniformly for all $g \in \mathcal{G}$, and $h \in \mathcal{H}$ such that $\Delta = h - h_g^* \in \mathcal{H}_B$,

$$\begin{aligned}&\Big|\big\{\Psi_n(h, f_\Delta, g) - \Psi_n(h_g^*, f_\Delta, g)\big\} - \big\{\Psi(h, f_\Delta, g) - \Psi(h_g^*, f_\Delta, g)\big\}\Big| \\ &\le 18\delta\left(\|(h - h_g^*)f_\Delta\|_2 + \delta\right) \\ &\le 18\delta(\|f_\Delta\|_2 + \delta),\end{aligned}$$

where in the second inequality, we use the fact that $h - h_g^* \in \mathcal{H}_B$, so that $\|h - h_g^*\|_\infty \le 1$. When $\|\Delta\|_{\mathcal{H}}^2 = \|h - h_g^*\|_{\mathcal{H}}^2 > B$, by replacing $h - h_g^*$ by $(h - h_g^*)\sqrt{B}/\|h - h_g^*\|_{\mathcal{H}}$ and multiplying both sides by $\|h - h_g^*\|_{\mathcal{H}}^2/B$, we have with probability $1 - \zeta$, uniformly for all $h \in \mathcal{H}$, $g \in \mathcal{G}$,

$$\begin{aligned}&\Big|\big\{\Psi_n(h, f_\Delta, g) - \Psi_n(h_g^*, f_\Delta, g)\big\} - \big\{\Psi(h, f_\Delta, g) - \Psi(h_g^*, f_\Delta, g)\big\}\Big| \\ &\le 18\delta(\|f_\Delta\|_2 + \delta)\max\left\{1, \frac{\|h - h_g^*\|_{\mathcal{H}}^2}{B}\right\}.\end{aligned}$$

When $\|f_{\widehat{\Delta}_g}\|_2 \geq C_f\delta$, with probability $1-\zeta$, uniformly for all $g \in \mathcal{G}$,

$$
\begin{aligned}
(I) \geq & r\left\{\Psi(\hat{h}_g, f_{\widehat{\Delta}_g}, g) - \Psi(h_g^*, f_{\widehat{\Delta}_g}, g)\right\} - 18\delta r\left[\|f_{\widehat{\Delta}_g}\|_2 + \delta\right] \max\left\{1, \frac{\|\hat{h}_g - h_g^*\|_{\mathcal{H}}^2}{B}\right\} \\
\geq & \underbrace{r\left\{\Psi(\hat{h}_g, f_{\widehat{\Delta}_g}, g) - \Psi(h_g^*, f_{\widehat{\Delta}_g}, g)\right\}}_{(I.1)} - 9\delta\left[C_f\delta + \delta\right] \max\left\{1, \frac{\|\hat{h}_g - h_g^*\|_{\mathcal{H}}^2}{B}\right\},
\end{aligned}
$$

where the second inequality is due to the definition of $r = \frac{C_f\delta}{2\|f_{\widehat{\Delta}_g}\|_2} \leq \frac{1}{2}$, and

$$
\begin{aligned}
(I.1) &= \frac{C_f\delta}{2\|f_{\widehat{\Delta}_g}\|_2}\left\{\Psi(\hat{h}_g, f_{\widehat{\Delta}_g}, g) - \Psi(h_g^*, f_{\widehat{\Delta}_g}, g)\right\} \\
&= \frac{C_f\delta}{2\|f_{\widehat{\Delta}_g}\|_2}\mathbb{E}\left\{\hat{h}_g(X) - h_g^*(X)\right\} f_{\widehat{\Delta}_g}(Z) \\
&= \frac{C_f\delta}{2\|f_{\widehat{\Delta}_g}\|_2}\mathbb{E}\left(f_{\widehat{\Delta}_g}(Z)\mathbb{E}\left[\hat{h}_g(X) - h_g^*(X) \mid Z\right]\right) \\
&= \frac{C_f\delta}{2\|f_{\widehat{\Delta}_g}\|_2}\mathbb{E}\left(f_{\widehat{\Delta}_g}(Z)\left\{\text{proj}_Z(\hat{h}_g - h_g^*)(Z)\right\}\right) \\
&= \frac{C_f\delta}{2\|f_{\widehat{\Delta}_g}\|_2}\mathbb{E}\left[f_{\widehat{\Delta}_g}(Z)^2 - \left\{f_{\widehat{\Delta}_g}(Z) - \text{proj}_Z(\hat{h}_g - h_g^*)(Z)\right\} f_{\widehat{\Delta}_g}(Z)\right] \\
&\geq \frac{C_f\delta}{2}\left(\|f_{\widehat{\Delta}_g}\|_2 - \|f_{\widehat{\Delta}_g} - \text{proj}_Z(\hat{h}_g - h_g^*)\|_2\right) \text{ by Cauchy-Schwartz inequality} \\
&\geq \frac{C_f\delta}{2}\left(\|f_{\widehat{\Delta}_g}\|_2 - \eta_n\right) \text{ since } \|f_{\widehat{\Delta}_g} - \text{proj}_Z(h_g^* - \hat{h}_g)\|_2 \leq \eta_n \\
&\geq \frac{C_f\delta}{2}\left(\|\text{proj}_Z(\hat{h}_g - h_g^*)\|_2 - 2\eta_n\right) \text{ by triangle inequality.}
\end{aligned}
$$

Finally, we have either $\|f_{\widehat{\Delta}_g}\|_2 < C_f\delta$ or with probability $1-2\zeta$, uniformly for all $g \in \mathcal{G}$:

$$
\begin{aligned}
&\sup_{f\in\mathcal{F}} \Psi_n(\hat{h}_g, f, g) - \Psi_n(h_g^*, f, g) - 2\lambda\left(\|f\|_{\mathcal{F}}^2 + \frac{U}{\delta^2}\|f\|_n^2\right) \geq (I) - 2\lambda(II) \\
\geq & \frac{C_f\delta}{2}\left(\|\text{proj}_Z(\hat{h}_g - h_g^*)\|_2 - 2\eta_n\right) - 9(C_f+1)\delta^2 \max\left\{1, \frac{\|\hat{h}_g - h_g^*\|_{\mathcal{H}}^2}{B}\right\} \\
& - \frac{2\lambda}{3}L^2\|\hat{h}_g - h_g^*\|_{\mathcal{H}}^2 - 2\lambda(\frac{1}{4} + \frac{3}{8}C_f^2)U. \qquad (30)
\end{aligned}
$$

Combine upper and lower bounds of (24). Combining the upper bound (28) and lower bound (30), we have either $\|f_{\hat{h}_g}\|_2 < C_f\delta$ or with probability $1-4\zeta$, uniformly for all $g \in \mathcal{G}$:

$$
\begin{aligned}
\frac{C_f\delta}{2}\|\text{proj}_Z(\hat{h}_g - h_g^*)\|_2 \leq & 2\lambda U + \left(72 + \frac{10}{C_\lambda}\right)\delta^2 + \lambda\mu(\|h_g^*\|_{\mathcal{H}}^2 - \|\hat{h}_g\|_{\mathcal{H}}^2) \\
& + C_f\delta\eta_n + 9(C_f+1)\delta^2 \max\left\{1, \frac{\|\hat{h}_g - h_g^*\|_{\mathcal{H}}^2}{B}\right\} \\
& + \frac{2\lambda}{3}L^2\|\hat{h}_g - h_g^*\|_{\mathcal{H}}^2 + \left(\frac{1}{2} + \frac{3}{4}C_f^2\right)\lambda U \\
= & \lambda\mu(\|h_g^*\|_{\mathcal{H}}^2 - \|\hat{h}_g\|_{\mathcal{H}}^2) + \left(\frac{2\lambda}{3}L^2 + \frac{9(C_f+1)\delta^2}{B}\right)\|\hat{h}_g - h_g^*\|_{\mathcal{H}}^2 \\
& + \left(\frac{5}{2} + \frac{3}{4}C_f^2\right)\lambda U + C_f\delta\eta_n + \left(72 + \frac{10}{C_\lambda} + 9(C_f+1)\right)\delta^2.
\end{aligned}
$$

Then, with the assumption that $\mu \geq \frac{4}{3}L^2 + \frac{18(C_f+1)}{B}\frac{\delta^2}{\lambda}$, we have

$$\begin{aligned}
&\lambda\mu(\|h_g^*\|_{\mathcal{H}}^2 - \|\hat{h}_g\|_{\mathcal{H}}^2) + \left(\frac{2\lambda}{3}L^2 + \frac{9(C_f+1)\delta^2}{B}\right)\|\hat{h}_g - h_g^*\|_{\mathcal{H}}^2 \\
\leq &\lambda\mu(\|h_g^*\|_{\mathcal{H}}^2 - \|\hat{h}_g\|_{\mathcal{H}}^2) + 2\left(\frac{2\lambda}{3}L^2 + \frac{9(C_f+1)\delta^2}{B}\right)\left(\|\hat{h}_g\|_{\mathcal{H}}^2 + \|h_g^*\|_{\mathcal{H}}^2\right) \\
\leq &2\lambda\mu\|h_g^*\|_{\mathcal{H}}^2 \leq 2\lambda\mu\sup_{g\in\mathcal{G}}\|h_g^*\|_{\mathcal{H}}^2.
\end{aligned}$$

Finally, with probability $1-4\zeta$, uniformly for all $g \in \mathcal{G}$:

$$\begin{aligned}
&\sup_{g\in\mathcal{G}}\|\text{proj}_Z(\hat{h}_g - h_g^*)\|_2 \\
\leq &\left(\frac{4\mu\sup_{g\in\mathcal{G}}\|h_g^*\|_{\mathcal{H}}^2 + 5U}{C_f} + \frac{3U}{2}C_f\right)\frac{\lambda}{\delta} + 2\eta_n + \left(\frac{162 + 20/C_\lambda}{C_f} + 18\right)\delta \\
\lesssim &\left[324C_\lambda'\left(\frac{4\mu\sup_{g\in\mathcal{G}}\|h_g^*\|_{\mathcal{H}}^2/U + 5}{C_f} + \frac{3}{2}C_f\right) + \frac{162 + 20/C_\lambda}{C_f} + 18\right]\delta + 2\eta_n \\
\lesssim &(1 + \sup_{g\in\mathcal{G}}\|h_g^*\|_{\mathcal{H}}^2)\delta,
\end{aligned}$$

where the second inequality is due to the assumption that $324C_\lambda\delta^2/U \leq \lambda \leq 324C_\lambda'\delta^2/U$, and the last inequality is due to the assumption that $\eta_n \lesssim \delta_n$. □

D.4.3 Proof of Lemma D.5

Proof. **Step 1. Critical radius of $\mathcal{F}_{3U}$.** Directly applying Lemma D.4, we only require that $\widehat{\delta}_n$ satisfies the inequality

$$\frac{64}{\sqrt{n}}\int_{\frac{\delta^2}{2}}^{\delta}\sqrt{\log N_n(t, \text{star}\{\mathcal{F}_{3U}\})}\text{d}t \leq \delta^2.$$

Then with probability $1-\zeta$, we have $\delta_n \leq \mathcal{O}(\widehat{\delta}_n + \sqrt{\frac{\log(1/\zeta)}{n}})$, where δ_n is the maximum critical radii of $\boldsymbol{\Omega}$.

Step 2. Critical radius of $\boldsymbol{\Xi}$.

Since $\boldsymbol{\Xi} \subset \{(x,z) \mapsto rh(x)f(z) : h \in \mathcal{H}_B, f \in \mathcal{F}_{L^2B}, r \in [0,1]\} \triangleq \tilde{\boldsymbol{\Xi}}$, we only need to consider a conservative critical radius for $\tilde{\boldsymbol{\Xi}}$.

Suppose that $\mathcal{H}_B^\epsilon$ is an empirical ϵ-covering of star$\{\mathcal{H}_B\}$ and $\mathcal{F}_{L^2B}^\epsilon$ is an empirical ϵ-covering of star$\{\mathcal{F}_{L^2B}\}$. Then for any $rhf \in \tilde{\boldsymbol{\Xi}}$, $r \in [0,1]$,

$$\begin{aligned}
\inf_{h_\epsilon\in\mathcal{H}_B^\epsilon, f_\epsilon\in\mathcal{F}_{L^2B}^\epsilon}\|h_\epsilon f_\epsilon - rhf\|_n &\leq \inf_{h_\epsilon\in\mathcal{H}_B^\epsilon}\|(h_\epsilon - h)f_\epsilon\|_n + \inf_{f_\epsilon\in\mathcal{F}_{L^2B}^\epsilon}\|h(rf - f_\epsilon)\|_n \\
&\leq \inf_{h_\epsilon\in\mathcal{H}_B^\epsilon}\|h_\epsilon - h\|_n + \inf_{f_\epsilon\in\mathcal{F}_{L^2B}^\epsilon}\|rf - f_\epsilon\|_n \\
&\leq 2\epsilon.
\end{aligned}$$

Therefore, $\mathcal{H}_B^{\epsilon/2} \times \mathcal{F}_{L^2B}^{\epsilon/2}$ is an empirical ϵ-covering of $\boldsymbol{\Xi}$. Since

$$\begin{aligned}
\log N_n(t, \mathbb{B}_n(\delta, G_\Delta)) &\leq \log N_n(t, \mathbb{B}_n(\delta, \tilde{\mathcal{G}}_\Psi)) \leq \log N_n(t, \tilde{\mathcal{G}}_\Psi) \\
&\leq \log N_n(t/2, \text{star}\{\mathcal{H}_B\}) + \log N_n(t/2, \text{star}\{\mathcal{F}_{L^2B}\}),
\end{aligned}$$

by Lemma D.4, we only require that $\hat{\delta}_n$ satisfies the inequality

$$\frac{64}{\sqrt{n}}\int_{\frac{\delta^2}{2}}^{\delta}\sqrt{\log N_n(t/2, \text{star}\{\mathcal{H}_B\}) + \log N_n(t/2, \text{star}\{\mathcal{F}_{L^2B}\})}\text{d}t \leq \delta^2.$$

Then with probability $1-\zeta$, we have $\delta_n \le \mathcal{O}(\widehat{\delta}_n + \sqrt{\frac{\log(1/\zeta)}{n}})$, where δ_n is the maximum critical radii of $\mathbf{\Omega}$.

Step 3. Critical radius of $\mathbf{\Omega}$.

$$\begin{aligned}\mathbf{\Omega} &\triangleq \left\{(x,w,z) \mapsto r(h_g^*(x) - g(w))f(z) : g \in \mathcal{G}_D, f \in \mathcal{F}_{3U}, r \in [0,1]\right\} \\ &\subset \{(x,w,z) \mapsto r(h(x) - g(w))f(z) : g \in \mathcal{G}_D, h \in \mathcal{H}_{AD}, f \in \mathcal{F}_{3U}, r \in [0,1]\} \\ &\triangleq \tilde{\mathcal{G}}_\Psi,\end{aligned}$$

where the second line is due to $\|h_g^*\|_{\mathcal{H}}^2 \le A\|g\|_{\mathcal{G}}^2$ for all $g \in \mathcal{G}$. Suppose that $\mathcal{H}_{AD}^\epsilon$ is an empirical ϵ-covering of star$\{\mathcal{H}_{AD}\}$ and $\mathcal{G}_D^\epsilon$ is that of star$\{\mathcal{G}_D\}$, $\mathcal{F}_{3U}^\epsilon$ is that of star$\{\mathcal{F}_{3U}\}$. Then for any $r(h-g)f \in \tilde{\mathcal{G}}_\Psi$, $r \in [0,1]$,

$$\begin{aligned}&\inf_{h_\epsilon \in \mathcal{H}_{AD}^\epsilon, f_\epsilon \in \mathcal{F}_{3U}^\epsilon, g_\epsilon \in \mathcal{G}_D^\epsilon} \|r(h-g)f - (h_\epsilon - g_\epsilon)f_\epsilon\| \\ &\le \inf_{f_\epsilon \in \mathcal{F}_{3U}^\epsilon} \|(h-g)(f_\epsilon - rf)\|_n + \inf_{h_\epsilon \in \mathcal{H}_{AD}^\epsilon} \|(h_\epsilon - h)f_\epsilon\|_n + \inf_{g_\epsilon \in \mathcal{G}_D^\epsilon} \|(g_\epsilon - g)f_\epsilon\|_n \\ &\le \inf_{f_\epsilon \in \mathcal{F}_{3U}^\epsilon} 2\|f_\epsilon - rf\|_n + \inf_{h_\epsilon \in \mathcal{H}_{AD}^\epsilon} \|h_\epsilon - h\|_n + \inf_{g_\epsilon \in \mathcal{G}_D^\epsilon} \|g_\epsilon - g\|_n \\ &\le 4\epsilon,\end{aligned}$$

where the second inequality is from triangular inequality and the thrid inequality is due to the fact that $\|h-g\|_\infty \le 2$ and $\|f_\epsilon\|_\infty \le 1$.

Therefore, $\mathcal{H}_{AD}^{\epsilon/4} \times \mathcal{G}_D^{\epsilon/4} \times \mathcal{F}_{3U}^{\epsilon/4}$ is an empirical ϵ-covering of $\mathbf{\Omega}$.

By Lemma D.4, we only require that $\widehat{\delta}_n$ satisfies the Dudley's integral inequality. Actually, since

$$\begin{aligned}\log N_n(t, \mathbb{B}_n(\delta, \mathbf{\Omega})) &\le \log N_n(t, \mathbb{B}_n(\delta, \tilde{\mathcal{G}}_\Psi)) \le \log N_n(t, \tilde{\mathcal{G}}_\Psi) \\ &\le \log N_n(t/4, \text{star}\{\mathcal{H}_{AD}\}) + \log N_n(t/4, \text{star}\{\mathcal{G}_D\}) \\ &\quad + \log N_n(t/4, \text{star}\{\mathcal{F}_{3U}\}),\end{aligned}$$

when $\widehat{\delta}_n$ satisfies the inequality

$$\frac{64}{\sqrt{n}} \int_{\frac{\delta^2}{2}}^{\delta} \sqrt{\log N_n(t/4, \text{star}\{\mathcal{H}_{AD}\}) + \log N_n(t/4, \text{star}\{\mathcal{G}_D\}) + \log N_n(t/4, \text{star}\{\mathcal{F}_{3U}\})}\mathrm{d}t \le \delta^2,$$

then with probability $1-\zeta$, we have $\delta_n \le \mathcal{O}(\widehat{\delta}_n + \sqrt{\frac{\log(1/\zeta)}{n}})$, where δ_n is the maximum critical radii of $\mathbf{\Omega}$. Finally, after combining Steps 1-3, we have that if $\widehat{\delta}_n$ satisfies the inequality

$$\frac{64}{\sqrt{n}} \int_{\frac{\delta^2}{2}}^{4\delta} \sqrt{\log N_n(t, \text{star}\{\mathcal{F}_{3U \vee L^2 B}\}) + \log N_n(t, \text{star}\{\mathcal{H}_{AD \vee B}\}) + \log N_n(t, \text{star}\{\mathcal{G}_D\})}\mathrm{d}t \le \delta^2,$$

then with probability $1-\zeta$, we have $\delta_n \le \mathcal{O}(\widehat{\delta}_n + \sqrt{\frac{\log(1/\zeta)}{n}})$, where δ_n is the maximum critical radii of $\mathcal{F}_{3U}$, $\mathbf{\Xi}$ and $\mathbf{\Omega}$. □

D.4.4 Proof of Lemma D.7

Proof. **Critical radius of $\mathbf{\Xi}$.** We consider a conservative critical radius for $\tilde{\mathcal{G}}_\Delta$, which is a tensor product of two RKHSs $\mathcal{H}_B$ and $\mathcal{F}_{L^2 B}$. Suppose that $\mathcal{H}$ and $\mathcal{F}$ are endowed with reproducing kernels $K_{\mathcal{H}}$ and $K_{\mathcal{F}}$, with ordered eigenvalues $\left\{\lambda_j^\downarrow(K_{\mathcal{H}})\right\}_{j=1}^\infty$ and $\left\{\lambda_j^\downarrow(K_{\mathcal{F}})\right\}_{j=1}^\infty$, respectively. Then the RKHS $\tilde{\mathcal{G}}_\Delta$ has reproducing kernel $K_{\mathbf{\Xi}} = K_{\mathcal{H}} \otimes K_{\mathcal{F}}$, with eigenvalues $\left\{\lambda_j^\downarrow(K_{\mathcal{H}})\right\}_{j=1}^\infty \times \left\{\lambda_j^\downarrow(K_{\mathcal{F}})\right\}_{j=1}^\infty$. Therefore, by Lemma D.6,

$$\mathcal{R}_n(\tilde{\mathcal{G}}_\Delta, \delta) \le \sqrt{\frac{2L^2B^2}{n}} \sqrt{\sum_{i,j=1}^\infty \min\left\{\lambda_i^\downarrow(K_{\mathcal{H}})\lambda_j^\downarrow(K_{\mathcal{F}}), \delta^2\right\}}.$$

Critical radius of $\boldsymbol{\Omega}$. We consider a conservative critical radius for

$$\tilde{\mathcal{G}}_\Psi = \{(x,w,z) \mapsto r(h(x)-g(w))f(z) : g \in \mathcal{G}_D, h \in \mathcal{H}_{AD}, f \in \mathcal{F}_{3U}, r \in [0,1]\}.$$

Let $\tilde{h}(x,w) = h(x)$ and $\tilde{g}(x,w) = g(w)$, $x \in \mathcal{X}, w \in \mathcal{W}$. In addition, $\tilde{h} \in \tilde{\mathcal{H}}_{AD}$ on $\mathcal{X} \times \mathcal{W}$ with kernel $K_{\tilde{\mathcal{H}}} = K_{\mathcal{H}} \otimes 1$ and $\tilde{g} \in \tilde{\mathcal{G}}_D$ on $\mathcal{X} \times \mathcal{W}$ with kernel $K_{\tilde{\mathcal{G}}} = 1 \otimes K_{\mathcal{G}}$. Notice that $h - g \in \tilde{\mathcal{H}}_{AD} + \tilde{\mathcal{G}}_D$, which is a RKHS endowed with RKHS norm $\|f\|_{\tilde{\mathcal{H}}+\tilde{\mathcal{G}}} = \min_{f=\tilde{h}+\tilde{g}, \tilde{h}\in\tilde{\mathcal{H}}, \tilde{g}\in\tilde{\mathcal{G}}} \|\tilde{h}\|_{\tilde{\mathcal{H}}} + \|\tilde{g}\|_{\tilde{\mathcal{G}}}$, and reproducing kernel $K_{\tilde{\mathcal{H}}} + K_{\tilde{\mathcal{G}}}$. As a result, $\|h-g\|_{\tilde{\mathcal{H}}+\tilde{\mathcal{G}}} \leq \sqrt{AD} + \sqrt{D}$ for all $h - g \in \tilde{\mathcal{H}}_{AD} + \tilde{\mathcal{G}}_D$. According to Weyl's inequality for compact self-adjoint operators in Hilbert spaces (see the s-number sequence theory in Hinrichs [2006] and Pietsch [1987, 2.11.9]), $\lambda^{\downarrow}_{i+j-1}(K_{\tilde{\mathcal{H}}} + K_{\tilde{\mathcal{G}}}) \leq \lambda^{\downarrow}_i(K_{\tilde{\mathcal{H}}}) + \lambda^{\downarrow}_j(K_{\tilde{\mathcal{G}}}) = \lambda^{\downarrow}_i(K_{\mathcal{H}}) + \lambda^{\downarrow}_j(K_{\mathcal{G}})$ whenever $i, j \geq 1$, so we have $\lambda^{\downarrow}_j(K_{\tilde{\mathcal{H}}} + K_{\tilde{\mathcal{G}}}) \leq \lambda^{\downarrow}_{[(j+1)/2]}(K_{\mathcal{H}}) + \lambda^{\downarrow}_{[(j+1)/2]}(K_{\mathcal{G}})$ whenever $j \geq 1$.

Since $(\tilde{\mathcal{H}} + \tilde{\mathcal{G}}) \otimes \mathcal{F}$ is a RKHS with reproducing kernel $(K_{\tilde{\mathcal{H}}} + K_{\tilde{\mathcal{G}}}) \otimes K_{\mathcal{F}}$, by the same argument for $\boldsymbol{\Xi}$, we have

$$\begin{aligned}\mathcal{R}_n(\tilde{\mathcal{G}}_\Psi, \delta) &\leq \sqrt{D}(1+\sqrt{A})\sqrt{\frac{6U}{n}}\sqrt{\sum_{i,j=1}^{\infty} \min\left\{[\lambda^{\downarrow}_{[(i+1)/2]}(K_{\mathcal{H}}) + \lambda^{\downarrow}_{[(i+1)/2]}(K_{\mathcal{G}})]\lambda^{\downarrow}_j(K_{\mathcal{F}}), \delta^2\right\}} \\ &\leq \sqrt{D}(1+\sqrt{A})\sqrt{\frac{12U}{n}}\sqrt{\sum_{i,j=1}^{\infty} \min\left\{[\lambda^{\downarrow}_i(K_{\mathcal{H}}) + \lambda^{\downarrow}_i(K_{\mathcal{G}})]\lambda^{\downarrow}_j(K_{\mathcal{F}}), \delta^2\right\}}.\end{aligned}$$

□

E Additional estimation details

In this section we demonstrate the performance of the proposed FQE-type algorithm introduced in Section 5 for the case where $\mathcal{H}^{(t)}$ and $\mathcal{F}^{(t)}$ are Reproducing kernel Hilbert spaces (RKHSs) endowed with reproducing kernels $K_{\mathcal{H}^{(t)}}$ and $K_{\mathcal{F}^{(t)}}$ respectively and canonical RKHS norms $\|\bullet\|_{\mathcal{H}^{(t)}} = \|\bullet\|_{K_{\mathcal{H}^{(t)}}}$, $\|\bullet\|_{\mathcal{F}^{(t)}} = \|\bullet\|_{K_{\mathcal{F}^{(t)}}}$ respectively, for $1 \leq t \leq T$.

For each $1 \leq t \leq T$, based on observed batch data $\{S_{t,i}, W_{t,i}, Z_{t,i}, A_{t,i}, R_{t,i}\}_{i=1}^n$, we can obtain the Gram matrices $\mathbf{K}_{\mathcal{H}^{(t)}} = [K_{\mathcal{H}^{(t)}}([W_{t,i}, S_{t,i}, A_{t,i}], [W_{t,j}, S_{t,j}, A_{t,j}])]_{i,j=1}^n$ and $\mathbf{K}_{\mathcal{F}^{(t)}} = [K_{\mathcal{F}^{(t)}}([Z_{t,i}, S_{t,i}, A_{t,i}], [Z_{t,j}, S_{t,j}, A_{t,j}])]_{i,j=1}^n$. Then we compute $\hat{q}_t^\pi = \widehat{\mathcal{P}}_t(\hat{v}_{t+1}^\pi + R_t)$ via (7) with $g = \hat{v}_{t+1}^\pi + R_t$. Specifically, $\hat{q}_t^\pi$ has the following form:

$$\hat{q}_t^\pi(w,s,a) = [\widehat{\mathcal{P}}_t(\hat{v}_{t+1}^\pi + R_t)](w,s,a) = \sum_{i=1}^n \alpha_i K_{\mathcal{H}^{(t)}}([W_{t,i}, S_{t,i}, A_{t,i}], [w,s,a]), \tag{31}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^\top = \left(\mathbf{K}_{\mathcal{H}^{(t)}}\mathbf{M}^{(t)}\mathbf{K}_{\mathcal{H}^{(t)}} + 4\lambda^2\mu\mathbf{K}_{\mathcal{H}^{(t)}}\right)^\dagger \mathbf{K}_{\mathcal{H}^{(t)}}\mathbf{M}^{(t)}\mathbf{Y}_t$ with $\mathbf{M}^{(t)} = \mathbf{K}_{\mathcal{F}^{(t)}}^{1/2}(\frac{M}{n\delta^2}\mathbf{K}_{\mathcal{F}^{(t)}} + \mathbf{I}_n)^{-1}\mathbf{K}_{\mathcal{F}^{(t)}}^{1/2}$, and $\mathbf{Y}_t = \mathbf{R}_t + \hat{\mathbf{v}}_{t+1}^\pi$ with $\mathbf{R}_t = [R_{t,1}, \ldots, R_{t,k}]^\top$ and $\hat{\mathbf{v}}_{t+1}^\pi = [\hat{v}_{t+1}^\pi(W_{t+1,1}, S_{t+1,1}), \ldots, \hat{v}_{t+1}^\pi(W_{t+1,n}, S_{t+1,n})]^\top$. Here $\mathbf{A}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{A}$.

Selection of hyper-parameters. There are several hyper-parameters in (31) for each $1 \leq t \leq T$. In each step, we treat $\mathbf{Y}_t = R_t + \hat{\mathbf{v}}_{t+1}^\pi$ as the response vector and use cross-validation to tune M/δ^2 and $\lambda^2\mu$ in (31). We adopt the tricks of Dikkala et al. [2020] and use the recommended defaults in their Python package `mliv`, where two scaling functions are defined by $\varsigma(n) = 5/n^{0.4}$ and $\zeta(\texttt{scale}, n) = \texttt{scale} \times \varsigma^4(n)/2$.

For cross-validation, let $I^{(1)}, \ldots, I^{(K)}$ denote the index sets of the randomly partitioned K folds of the indices $\{1, \ldots, n\}$ and $I^{(-k)} = \{1, \ldots, n\} \backslash I^{(k)}$, $k = 1, \ldots, K$. We summarize the one-step NPIV estimation with cross-validation in Algorithm 2.

Algorithm 2: Min-max NPIV estimation with RKHSs

Input: $\{S_{t,i}, W_{t,i}, Z_{t,i}, A_{t,i}, Y_{t,i} = R_{t,i} + \hat{v}^{\pi}_{t+1}(W_{t+1,i}, S_{t+1,i})\}_{i=1}^{n}$, target policy π_t, kernels $K_{\mathcal{H}^{(t)}}, K_{\mathcal{F}^{(t)}}$, SCALE as some positive scaling factors, the number of cross-validation partition K.

Repeat for $\texttt{scale} \in \texttt{SCALE}$:

  Repeat for $k = 1, \ldots, K$:

    $[M/\delta^2]^{(-k)} = 1/\varsigma^2(|I^{(-k)}|)$, $[\lambda^2\mu]^{(-k)} = \zeta(\texttt{scale}, |I^{(-k)}|)$.

    Obtain $\hat{q}_t^{\pi\ (-k)}$ by (31) with data whose indices are in $I^{(-k)}$.

    $[M/\delta^2]^{(k)} = 1/\varsigma^2(|I^{(k)}|)$.

    Calculate $\epsilon_i = Y_{t,i} - \hat{q}_t^{\pi\ (-k)}(W_{t,i}, S_{t,i}, A_{t,i})$ for $i \in I^{(k)}$.

    $\texttt{Loss}^{(k)}(\texttt{scale}) = \epsilon^\top \mathbf{M}_{I^{(k)}} \epsilon$, where $\epsilon = [\epsilon_i]_{i \in I^{(k)}}^\top$ and $\mathbf{M}_{I^{(k)}}$ is obtained by data in $I^{(k)}$.

  $\texttt{Loss}(\texttt{scale}) = K^{-1} \sum_{k=1}^{K} \texttt{Loss}^{(k)}(\texttt{scale})$.

$\texttt{scale}^* = \arg\min_{\texttt{scale} \in \texttt{SCALE}} \texttt{Loss}(\texttt{scale})$.

Obtain $\hat{q}_t^{\pi}$ by (31) with all data and $M/\delta^2 = 1/\varsigma^2(n)$, $\lambda^2\mu = \zeta(\texttt{scale}^*, n)$.

Output: $\{\hat{v}_t^{\pi}(W_{t,i}, S_{t,i}) = \sum_{a \in \mathcal{A}} \hat{q}_t^{\pi}(W_{t,i}, S_{t,i}, a)\pi(a \mid S_{t,i})\}_{i=1}^{n}$.

Below we summarize our proposed FQE-type algorithm using a sequential NPIV estimation with tuning procedure described in Algorithm 3.

Algorithm 3: A FQE-type algorithm by sequential min-max NPIV estimation

Input: Batch Data $\mathcal{D}_n = \{\{S_{t,i}, W_{t,i}, Z_{t,i}, A_{t,i}, R_{t,i}\}_{t=1}^{T}\}_{i=1}^{n}$, a target policy $\pi = \{\pi_t\}_{t=1}^{T}$, kernels $\{K_{\mathcal{H}^{(t)}}, K_{\mathcal{F}^{(t)}}\}_{t=1}^{T}$, set SCALE as some positive scaling factors, number of cross-validation partition K.

Let $\hat{v}^{\pi}_{T+1} = 0$.

Repeat for $t = T, \ldots, 1$:

  Obtain $\{\hat{v}_t^{\pi}(W_{t,i}, S_{t,i})\}_{i=1}^{n}$ by Algorithm 2.

Output: $\widehat{\mathcal{V}}(\pi) = n^{-1} \sum_{k=1}^{n} \hat{v}_1^{\pi}(W_{1,k}, S_{1,k})$.

F Simulation details

In this section, we perform a simulation study to evaluate the performance of our proposed OPE estimation and to verify the finite-sample error bound of our OPE estimator in the main result Theorem 6.3.

F.1 Simulation setup

Let $\mathcal{S} = \mathbb{R}^2, \mathcal{U} = \mathbb{R}, \mathcal{W} = \mathbb{R}, \mathcal{Z} = \mathbb{R}$, and $\mathcal{A} = \{1, -1\}$.

MDP setting. At time t, given (S_t, U_t, A_t), we generate

$$S_{t+1} = S_t + A_t U_t \mathbf{1}_2 + e_{S_{t+1}},$$

where $\mathbf{1}_2 = [1, 1]^\top$ and the random error $e_{S_{t+1}} \sim \mathcal{N}([0, 0]^\top, \mathbf{I}_2)$ with $\mathbf{I}_2$ denoting the 2-by-2 identity matrix.

The behavior policy is

$$\tilde{\pi}_t^b(A_t \mid U_t, S_t) = \text{expit}\left\{-A_t\left(t_0 + t_u U_t + t_s^\top S_t\right)\right\},$$

where $t_0 = 0$, $t_u = 1$, and $t_s^\top = [-0.5, -0.5]$.

By this behavior policy

$$\pi_t^b(A_t \mid S_t) = \mathbb{E}[\tilde{\pi}_t^b(A_t \mid U_t, S_t) \mid A_t, S_t] = \text{expit}\{-A_t\left(t_0 + t_u\kappa_0 + (t_s + t_u\kappa_s)^\top S_t\right)\},$$

provided that the following conditional distribution is used.

We generate the hidden state U_t, and two proximal variables Z_t and W_t by the following conditional multivariate normal distribution given (S_t, A_t):

$$(Z_t, W_t, U_t) \mid (S_t, A_t) \sim N\left(\begin{bmatrix} \alpha_0 + \alpha_a A_t + \alpha_s S_t \\ \mu_0 + \mu_a A_t + \mu_s S_t \\ \kappa_0 + \kappa_a A_t + \kappa_s S_t \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_z^2 & \sigma_{zw} & \sigma_{zu} \\ \sigma_{zw} & \sigma_w^2 & \sigma_{wu} \\ \sigma_{zu} & \sigma_{wu} & \sigma_u^2 \end{bmatrix} \right),$$

where

- $\alpha_0 = 0$, $\alpha_a = 0.5$, $\alpha_s^\top = [0.5, 0.5]$,
- $\mu_0 = 0$, $\mu_a = -0.25$, $\mu_s^\top = [0.5, 0.5]$,
- $\kappa_0 = 0$, $\kappa_a = -0.5$, $\kappa_s^\top = [0.5, 0.5]$
- the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 0.25 & 0.5 \\ 0.25 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}$$

The initial S_1 is uniformly sampled[2] from $\mathbb{R}^2$.

Reward setting. The reward is given by

$$R_t = \text{expit}\left\{ \frac{1}{2} A_t (U_t + [1, -2] S_t) \right\} + e_t,$$

where $e_t \sim \text{Uniform}[-0.1, 0.1]$. One can verify that our simulation setting satisfies the conditions in Section A.1 so that our method can be applied.

Target policy. We evaluate a ϵ-greedy policy $\pi(a \mid S_t)$ maximizing the immediate reward:

$$A_t \mid S_t \sim \begin{cases} \text{sign}\{\mathbb{E}[U_t + [1, -2] S_t \mid S_t]\} & \text{with probability } 1 - \epsilon, \\ \text{Uniform}\{-1, 1\} & \text{with probability } \epsilon. \end{cases}$$

We set $\epsilon = 0.2$.

F.2 Implementation

We present the results of policy evaluation for the simulation setup above. Specifically, to evaluate the finite-sample error bound of the proposed estimator in terms of the sample size n, we consider $T = 1, 3, 5$ and let $n = 256, 512, 1024, 2048, 4096$; to evaluate the estimation error of our OPE estimator in terms of the length of horizon T, we fix $n = 512$ and let $T = 1, 2, 4, 8, 16, 24, 32, 48, 64$. For each setting of (n, T), we repeat 100 times. All simulation are computed on a desktop with one AMD Ryzen 3800X CPU, 32GB of DDR4 RAM and one Nvidia RTX 3080 GPU.

We choose $\mathcal{F}^{(t)}$ and $\mathcal{H}^{(t)}$ as RKHSs endowed with Gaussian kernels, with bandwidths selected according to the median heuristic trick by Fukumizu et al. [2009] for each $1 \leq t \leq T$. The pool of scaling factors `SCALE` contains 30 positive numbers spaced evenly on a log scale between 0.001 to 0.05. The number of cross-validation partition $K = 5$. The true target policy value of π is estimated by the mean cumulative rewards of $50,000$ Monte Carlo trajectories with policy π.

[2] Sample by `gym` package build in function `spaces.sample()` from `spaces.Box(low=-np.inf, high=np.inf, shape=(2,), dtype=np.float32)`.