Supplemental Material for
Efficiently Breaking the Curse of Horizon in Off-Policy Evaluation with Double Reinforcement Learning

by Nathan Kallus and Masatoshi Uehara

Appendix A: Notation

We first summarize the notation we use in Table EC.1. Following empirical process theory literature, in the proofs we also use $\mathbb{P}$ to denote expectations interchangeably with $\mathrm{E}[\cdot]$.

Table EC.1 Notation

$\mathcal{M}_1, \mathcal{M}_{1,b}$	NMDP model, see Definition 1
$\mathcal{M}_2, \mathcal{M}_{2,b}$	TMDP model, see Definition 2
$\mathcal{M}_3, \mathcal{M}_{3,b}$	MDP model, see Definition 3
$p^{(0)}_{\pi_b}(s)$, $p^{(0)}_{\pi_e}(s)$	Initial distributions
$\mathcal{J}$	$(s_0, a_0, r_0, s_1, a_1, r_1, \dots)$
$\mathcal{J}_{s_t}$, $\mathcal{J}_{a_t}$	History up to s_t or to a_t, respectively
$\mathcal{H}_{s_t}$, $\mathcal{H}_{a_t}$	History up to s_t or to a_t, respectively, excluding reward variables
$r_t, s_t, a_t,$	Reward, state, and action at t
η_t	$\pi_{e,t}/\pi_{b,t}$
ρ^π	Policy value, $\lim_{T\to\infty} c_T(\gamma)\,\mathrm{E}_\pi[\sum_{t=0}^T \gamma^t r_t]$
ρ^π_T	$c_T(\gamma)\,\mathrm{E}_\pi[\sum_{t=0}^T \gamma^t r_t]$
$p^{(\infty)}_{\pi,\gamma}(s)$	Average visitation distribution of the policy π with discount rate γ
$p^{(\infty)}_{\pi}(s)$	Stationary distribution
$w(s)$	$p^{(\infty)}_{\pi_e,\gamma}(s)/p^{(\infty)}_b(s)$
∇_β	Differentiation with respect to β
$\pi_e(a\|s)$, $\pi_b(a\|s)$	Target and behavior policies respectively
$v(s)$	Value function
$q(s,a)$	q-function
$\nu_t(\mathcal{H}_{a_t})$	Cumulative density ratio $\prod_{k=0}^t \pi_{e,k}/\pi_{b,k}$
$\mu_t(s_t,a_t)$	Marginal density ratio $\mathrm{E}[\nu_t \mid s_t, a_t]$
$\eta(s,a)$	Instantaneous density ratio $\pi_e(a\|s)/\pi_b(a\|s)$
$C, R_{\max}$	Upper bound of density ratio and reward, respectively
$\\|\cdot\\|_p$	L^p-norm $\mathrm{E}[f^p]^{1/p}$
$\mathrm{E}_\pi[\cdot], \mathbb{P}_\pi$	Expectation with respect to a sample from a policy π
$\mathrm{E}[\cdot], \mathbb{P}$	Same as above for $\pi = \pi_b$
$\mathrm{E}_N[\cdot], \mathbb{P}_N, \mathbb{P}_T$	Empirical or time average (based on sample from a behavior policy)
$\mathcal{D}_0, \mathcal{D}_1$	The split samples when using cross-fitting, $\mathcal{D}_0 \cup \mathcal{D}_1 = \{1,\dots,N\}$
N_j	The size of $\mathcal{D}_j$
$\mathrm{E}_{N_j}, \mathbb{P}_{N_j}$	Empirical expectation on $\mathcal{D}_j$
$\mathbb{G}_N$	Empirical process $\sqrt{N}(\mathbb{P}_N - \mathbb{P})$
$A_N = o_p(a_N)$	The term A_N/a_N converges to zero in probability
$A_N = \mathcal{O}_p(a_N)$	The term A_N/a_N is bounded in probability

Appendix B: Technical Background

B.1. Some Definitions

Definition EC.1 ($\mathbb{P}$-Donsker). A class $\mathcal{F}$ of measurable functions f is called $\mathbb{P}$-Donsker when the sequence of processes $\sqrt{n}(\mathbb{P}_n - \mathbb{P})$ converges in distribution to a tight limit process in the space $l^\infty(\mathcal{F})$ as $n \to \infty$.

If the class $\mathcal{F}$ has a square-integrable envelope and a finite uniform entropy integral, then $\mathcal{F}$ is $\mathbb{P}$-Donsker (van der Vaart 1998, Theoreem 19.14).

Next we review the definitions of the standard mixing coefficients (Davidson 1994, Chapter 14).

Definition EC.2 (Mixing coefficients). Consider a stationary stochastic process $X_1, X_2, \dots$, *i.e.*, identically distributed but possibly dependent. Let P denote the joint measure of the process, define $\mathcal{Q}_a^b = \sigma(X_a, \cdots, X_b)$ as the sigma algebra corresponding to subsequences of variable, and let $L^2(Q_a^b)$ be the corresponding space of square-integrable functions. Then we have the following definitions:

$$
\begin{aligned}
\alpha_m &= \sup_t \sup_{G \in \mathcal{Q}_1^t, H \in \mathcal{Q}_{t+m}^\infty} |P(G \cap H) - P(G)P(H)|, \\
\beta_m &= \sup_t \sup_{H \in \mathcal{Q}_{t+m}^\infty} |P(H|\mathcal{Q}_1^t) - P(H)|, \\
\phi_m &= \sup_t \sup_{G \in \mathcal{Q}_1^t, H \in \mathcal{Q}_{t+m}^\infty, P(G) > 0} |P(H|G) - P(H)|, \\
\rho_m &= \sup_t \sup_{g \in L^2(\mathcal{Q}_1^t), h \in L^2(\mathcal{Q}_{t+m}^\infty)} |\mathrm{Corr}(g, h)|.
\end{aligned}
$$

Each of these coefficients measures the amount of dependence when we consider subsequences that are separated by m jumps. The more ergodic the process, the faster these shrink toward zero as m grows. An independent process has all of these always equal zero.

These coefficients satisfy the following relationships (Bradley 2005):

$$
2\alpha_m \le \beta_m \le \phi_m, \quad 4\alpha_m \le \rho_m \le 2\phi_m^{1/2}. \tag{EC.1}
$$

B.2. Semiparametric Theory

In this section, we expand on Section 1.3 to briefly review the precise definitions and results of semiparametric theory that we use following (van der Vaart 1998, Bickel et al. 1998). We focus on scalar estimands for simplicity since our OPE problem involves a scalar estimand. We denote the all of the data $\{O_i\}_{i=1}^n$ as O^n, the estimand as $R: \mathcal{M} \to \mathbb{R}$, and the estimator as $\hat{R}(O^n)$. We let μ be a dominating measure for $\mathcal{M}$ such that $\mu \gg F$ for $F \in \mathcal{M}$.

B.2.1. Definitions

Definition EC.3 (Influence function of estimators). An estimator $\hat{R}(O^n)$ is asymptotically linear (AL) with influence function (IF) $\psi(O)$ if

$$\sqrt{n}(\hat{R}(O^n) - R(F)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(O_i) + \mathrm{o}_p(1/\sqrt{n}).$$

Definition EC.4 (One-dimensional submodel and its score function). A one-dimensional submodel of $\mathcal{M}$ passing through F is a set of distributions $\{F_\epsilon : \epsilon \in (-1,1)\} \subset \mathcal{M}$ such that: (a) $F_0 = F$, (b) the score function

$$s(O;\epsilon) = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\log(\mathrm{d}F_\epsilon/\mathrm{d}\mu)(O)$$

exists, (c) $\mathbb{E}s^2(O;0) < \infty$, and (d) $\mathbb{E}\sup_{\epsilon\in(-1,1)}|\mathrm{d}F_\epsilon/\mathrm{d}\mu(O)| < \infty$.

The score of the submodel at F is defined as $s(O) = s(O;0)$. By definition it belongs to L_2, the Hilbert space of square-integrable functions wrt F.

Definition EC.5 (Tangent space). The tangent space at F wrt $\mathcal{M}$ is the linear closure of the score functions at F over all one-dimensional submodels wrt L_2.

Note that the tangent space is always a cone.

Definition EC.6 (Pathwise differentiability). A functional $R(F)$ is pathwise differentiable at F wrt $\mathcal{M}$ if there exists a function $D_F(O)$ such that any one-dimensional submodel $\{F_\epsilon\}$ of $\mathcal{M}$ passing through F with score function $s(O)$ satisfies

$$\frac{dR(F_\epsilon)}{d\epsilon}\,|_{\epsilon=0} = \mathbb{E}[D_F(O)s(O)].$$

The function $D_F(O)$ is called a gradient of $R(F)$ at F wrt $\mathcal{M}$. The efficient IF (EIF, or canonical gradient) of $R(F)$ wrt $\mathcal{M}$ is the unique gradient $\tilde{D}_F(O)$ of $R(F)$ at F wrt $\mathcal{M}$ that belongs to the tangent space at F wrt $\mathcal{M}$.

Next, we define regular estimators, which are those whose limiting distribution is insensitive to local changes to the data-generating process. It excludes, for example, super-efficient pathologies such as the Hodge estimator.

Definition EC.7 (Regular estimators). An estimator sequence $\hat{R}$ is called regular at F for $R(F)$ wrt $\mathcal{M}$, if there exists a probability measure L such that, for any one-dimensional submodel $\{F_\epsilon\}$ of $\mathcal{M}$ passing through F, we have

$$\sqrt{n}\{\hat{R}(O^n) - R(F_{1/\sqrt{n}})\} \text{ with } O^n \sim F^n_{1/\sqrt{n}}, \text{ converges in distribution to } L \text{ as } n \to \infty.$$

B.2.2. Characterizations The following theorems show that influence functions of asymptotically linear estimators for $R(F)$ and gradients of $R(F)$ correspond to one another and how to compute the EIF. These are both based on Theorem 3.1 of van der Vaart (1991).

THEOREM EC.1 **(Influence functions are gradients)**. *Suppose $\hat{R}(O^n)$ is an asymptotically linear estimator of $R(F)$ with influence function $D_F(O)$ and that $R(F)$ is pathwise differentiable at F wrt $\mathcal{M}$. Then, $\hat{R}(O^n)$ is a regular estimator of $R(F)$ wrt $\mathcal{M}$ if and only if and $D_F(O)$ is a gradient of $R(F)$ at F wrt $\mathcal{M}$.*

COROLLARY EC.1 **(Characterization of EIF)**. *The EIF wrt $\mathcal{M}$ is the projection of any gradient wrt $\mathcal{M}$ onto the tangent space wrt $\mathcal{M}$.*

B.2.3. Strategy to calculate the EIF With the above definitions and theorems in mind, our general strategy to construct EIF is as follows.

1. Calculate a gradient $D_F(O)$ of the target functional $R(F)$.
2. Calculate the tangent space wrt $\mathcal{M}$.
3. Either:
   (a) Show that $D_F(O)$ already lies in the tangent space and is thus the EIF, or
   (b) Project $D_F(O)$ onto the tangent space to obtain the EIF.

B.2.4. Optimalites The efficiency bound is defined as the variance of the EIF, $\mathrm{var}_F[\tilde{D}_F(O)]$. It has the following interpretations. First, the efficiency bound is the lower bound in a local asymptotic minimax sense (van der Vaart 1998, Thm. 25.20).

THEOREM EC.2 **(Local Asymptotic Minimax theorem)**. *Let $R(F)$ be pathwise diffentiable at F wrt $\mathcal{M}$ with the EIF $\tilde{D}_F(O)$. Then, for any estimator sequence $\hat{R}(O^n)$, and symmetric quasi-convex loss function $l:\mathbb{R}\to[0,\infty)$,*

$$\sup_{m\in\mathbb{N},\{F_\epsilon^{(1)}\},\dots,\{F_\epsilon^{(m)}\}} \lim_{n\to\infty} \sup_{k=1,\dots,m} \mathrm{E}_{F^{(k)}_{1/\sqrt{n}}}[l(\sqrt{n}\{\hat{R}(O^n)-R(F^{(k)}_{1/\sqrt{n}})\})] \geq \int l(u)\mathrm{d}\mathcal{N}(0,\mathrm{var}_F[\tilde{D}_F(O)])(u),$$

where the first supremum is taken over all finite collections of one-dimensional submodels of $\mathcal{M}$ passing through F.

The above suprema appear somewhat complicated. It also implies the following weaker but easier to interpret result.

COROLLARY EC.2. *Under the same assumptions of Theorem EC.2,*

$$\inf_{\delta>0}\liminf_{n\to\infty} \sup_{\|Q-F\|_{TV}\leq\delta} \mathrm{E}_Q[l(\sqrt{n}\{\hat{R}(O^n)-R(Q)\})] \geq \int l(u)\mathrm{d}\mathcal{N}(0,\mathrm{var}_F[\tilde{D}_F(O)])(u),$$

where $\|\cdot\|_{TV}$ is the total variation distance.

Second, the efficiency bound can be seen as a pointwise lower bound bound at the specific instance F, if we restrict to regular estimators (van der Vaart 1998, Thm. 25.21).

THEOREM EC.3 **(Convolution theorem)**. *Let $l:\mathbb{R}\to[0,\infty)$ be a symmetric quasi-convex loss function. Let $R(F)$ be pathwise differentiable at F wrt $\mathcal{M}$ with EIF $\tilde{D}_F(O)$. Let $\hat{R}(O^n)$ be a regular estimator sequence at F wrt $\mathcal{M}$ with limiting distribution L. Then,*

$$\int l(u)\mathrm{d}L(u)\geq\int l(u)\mathrm{d}\mathcal{N}(0,\mathrm{var}_F[\tilde{D}_F(O)])(u).$$

Equality holds obviously holds when $L=\mathcal{N}(0,\mathrm{var}_F[\tilde{D}_F(O)])$. We therefore say an estimator is efficient if it is regular and its limiting distribution is $\mathcal{N}(0,\mathrm{var}_F[\tilde{D}_F(O)])$. Note that such would be implied if it were regular and AL with influence function $\tilde{D}_F(O)$. The following implies such is in fact both necessary and sufficient (van der Vaart 1998, Lemma 25.23).

THEOREM EC.4. *Let $R(F)$ be pathwise differentiable at F wrt $\mathcal{M}$ with the EIF $\tilde{D}_F(O)$. Then an estimator sequence is efficient (regular wrt $\mathcal{M}$ and with limiting distribution $\mathcal{N}(0,\mathrm{var}_F[\tilde{D}_F(O)])$) if and only if is AL with influence function $\tilde{D}_F(O)$.*

Appendix C: Proofs

Proof of Theorem 1 We omit the proof since as it can (almost) be concluded as corollary of Theorem 1 of Kallus and Uehara (2020) where we simply set $T=\infty$ and redefine rewards by multiplying by γ^t. There are only two subtle differences. They prove the efficiency bound when $p(r_t|\mathcal{J}_{a_t}),p(a_t|\mathcal{J}_{s_t}),p(s_t|\mathcal{J}_{r_t})$ are replaced with $p(r_t|\mathcal{H}_{a_t}),p(a_t|\mathcal{H}_{s_t}),p(s_t|\mathcal{H}_{a_t})$, and the initial density is unknown. The former change simply means our conditioning sets are in the efficiency bound are slightly different. The latter change simply eliminates the $k=0$ term in the efficiency bound.

Proof of Theorem 2 Again we omit the proof as it can (almost) be concluded as a corollary of Theorem 3 of (Kallus and Uehara 2020). There is a single subtle difference. They prove the efficiency bound when the initial density is unknown. This change simply eliminates the $k=0$ term in the efficiency bound.

Proof of Corollary 1 We prove the statement for NMDP. For TMDP, the statement is confirmed similarly. We have

$$\begin{aligned}&\lim_{T\to\infty}(1-\gamma)^2\sum_{k=1}^{T}\mathrm{E}[\gamma^{2(k-1)}\nu_{k-1}^2\mathrm{var}\{r_{k-1}+v(s_k)|s_{k-1},a_{k-1}\}]\\ &\lessapprox\lim_{T\to\infty}(1-\gamma)^2\sum_{k=1}^{T}\gamma^{2(k-1)}\|\nu_k\|_\infty^2<\infty.\end{aligned}$$

The final statement is clear because TMDP is included in NMDP.

Proof of Theorem 3 Let $o=(s,a,r,s')$ and let $L_2=\{f(o):\mathbb{E}f^2(o)<\infty\}$. Recall the model is

$$\mathcal{M}_3=\{p(o)=p(s)p(a|s)p(r|s,a)p(s'|s,a)\,:\,p(s)p(a\mid s)>0\}.$$

Calculation of the tangent space We first prove that the corresponding tangent space is

$$\mathcal{T} = L_2 \cap (\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 + \mathcal{T}_4), \quad \text{where}$$
$$\mathcal{T}_1 = \{f(s) : \mathbb{E}[f] = 0\}, \mathcal{T}_2 = \{f(s,a) : \mathbb{E}[f(s,a)|s] = 0\},$$
$$\mathcal{T}_3 = \{f(s,a,s') : \mathbb{E}[f(s,a,s')|s,a] = 0\},$$
$$\mathcal{T}_4 = \{f(s,a,r) : \mathbb{E}[f(s,a,r)|s,a] = 0\}.$$

Notice that $\mathcal{T}_j$ are orthogonal so the above sum is a direct sum.

For any submodel

$$\{p_\theta(s)p_\theta(a \mid s)p_\theta(s'|s,a)p_\theta(r|s,a)\},$$

passing through p at $\theta = 0$, the score function can be decomposed as

$$g(o) = g_s(s) + g_{a|s}(s,a) + g_{r|s,a}(s,a,r) + g_{s'|s,a}(s,a,s'), \text{ where}$$
$$g_s(s) \coloneqq \frac{d}{d\theta}\log p_\theta(s),\ g_{a|s}(s,a) \coloneqq \frac{d}{d\theta}\log p_\theta(a|s),$$
$$g_{r|s,a}(a,s,r) \coloneqq \frac{d}{d\theta}\log p_\theta(r|s,a),\ g_{s'|s,a}(s,a,a') \coloneqq \frac{d}{d\theta}\log p_\theta(s'|s,a),$$

where these satisfy

$$\mathbb{E}[g_s(s)] = 0, \mathbb{E}[g_{a|s}(s,a) \mid s] = 0, \mathbb{E}[g_{r|s,a}(s,a,r) \mid s,a] = 0, \mathbb{E}[g_{s'|s,a}(s,a,s') \mid s,a] = 0.$$

Therefore, $\mathcal{T}$ contains the tangent space.

Next, consider any $g(o) \in \mathcal{T}$. Since g is in the direct sum $\mathcal{T}_j$ we can write $g(o) = g_s(s) + g_{a|s}(s,a) + g_{r|s,a}(s,a,r) + g_{s'|s,a}(s,a,s')$. Let $k(x) = 2(1+e^{-2x})^{-1}$. Notice that k is bounded by 2 and $k(0) = k'(0) = 1$. Consider the submodel

$$p_\theta(s)p_\theta(a|s)p_\theta(r|s,a)p_\theta(s'|s,a), \quad \theta \in (-\epsilon, \epsilon)$$

where

$$p_\theta(s) = p(s)k(\theta g_s(s)),\ p_\theta(a|s) = p(a|s)k(\theta g_{a|s}(s,a)),$$
$$p_\theta(s'|s,a) = p(s'|s,a)k(\theta g_{s'|s,a}(s,a,s')),\ p_\theta(r|s,a) = p(r|s,a)k(\theta g_{r|s,a}(s,a,r)).$$

Taking the score of this submodel we see it exactly coincides with g, completing the argument.

Calculation of gradient We show the statement in two steps: (1) calculating some gradient, (2) showing this gradient is in the tangent space of $\mathcal{M}_3$. Then, we can conclude that this gradient is actually the EIF.

To calculate a gradient of the target functional $J(\pi_e)$ wrt the model ($\mathcal{M}_3$) we seek ϕ satisfying

$$\nabla_\theta \rho^{\pi_e}(\theta)|_{\theta=0} = \mathrm{E}_{p_b}[\phi(s,a,r,s')g(s,a,r,s')]|_{\theta=0},$$
$$\rho^{\pi_e}(\theta) = \lim_{T\to\infty} \mathrm{E}[c_T(\gamma)\sum_{t=0}^{T}\gamma^t r_t | s_0 \sim p_e(s_0), a_0 \sim \pi_e(a_1|s_1), s_1 \sim p_\theta(s_1|s_0,a_0), r_1 \sim p_\theta(r_1|s_1,a_1), \cdots],$$

where $p_\theta(s)p_\theta(a \mid s)p_\theta(s'|s,a)p_\theta(r|s,a)$ is any submodel, whose score function belongs to the tangent set $\mathcal{T}$. Note

$$\begin{aligned}\rho^{\pi_e}(\theta) &= \lim_{T\to\infty} c_T(\gamma)\sum_{t=0}^{T}\int \gamma^t r_t p_\theta(r_t|s_t,a_t)\left\{\prod_{k=0}^{t}\pi_e(a_k|s_k)p_\theta(s_{k+1}|s_k,a_k)\right\}p^{(0)}_{\pi_e}(s_0)\mathrm{d}\lambda(\mathcal{J}_{s_{t+1}})\\ &= \int\int c_t(\gamma)\gamma^t r_t p_\theta(r_t|s_t,a_t)\left\{\prod_{k=0}^{t}\pi_e(a_k|s_k)p_\theta(s_{k+1}|s_k,a_k)\right\}p^{(0)}_{\pi_e}(s_0)\mathrm{d}\lambda(\mathcal{J}_{s_{t+1}})\mathrm{d}\mu'(t)\end{aligned}$$

where μ' is a counting measure on $\mathbb{Z}$. This is only to emphasize that $\lim_{T\to\infty}\sum_{t=0}^{T}$ is an integral wrt the counting measure.

Then, we have

$$\nabla_\theta \rho^{\pi_e}(\theta)|_{\theta=0} = C(\theta)|_{\theta=0}, \tag{EC.2}$$
$$\begin{aligned}C(\theta) &= \lim_{T\to\infty} c_T(\gamma)\sum_{t=0}^{T}\int \gamma^t r_t \nabla p_\theta(r_t|s_t,a_t)\left\{\prod_{k=0}^{t}\pi_e(a_k|s_k)p_\theta(s_{k+1}|s_k,a_k)\right\}p^{(0)}_{\pi_e}(s_0)\mathrm{d}\lambda(\mathcal{J}_{s_{t+1}})\\ &+ \lim_{T\to\infty} c_T(\gamma)\sum_{t=0}^{T}\int \gamma^t r_t p_\theta(r_t|s_t,a_t)\left\{\prod_{k=0}^{t}\pi_e(a_k|s_k)\nabla p_\theta(s_{k+1}|s_k,a_k)\right\}p^{(0)}_{\pi_e}(s_0)\mathrm{d}\lambda(\mathcal{J}_{s_{t+1}}).\end{aligned}$$

This exchange of integration and differentiation is justified by showing

$$\lim_{T\to\infty} c_T(\gamma)\sum_{t=0}^{T}\int \gamma^t r_t |\nabla p_\theta(r_t|s_t,a_t)|\left\{\prod_{k=0}^{t}\pi_e(a_k|s_k)p_\theta(s_{k+1}|s_k,a_k)\right\}p^{(0)}_{\pi_e}(s_0)\mathrm{d}\lambda(\mathcal{J}_{s_{t+1}}) \tag{EC.3}$$
$$+\lim_{T\to\infty} c_T(\gamma)\sum_{t=0}^{T}\int \gamma^t r_t p_\theta(r_t|s_t,a_t)\left\{\prod_{k=0}^{t}\pi_e(a_k|s_k)|\nabla p_\theta(s_{k+1}|s_k,a_k)|\right\}p^{(0)}_{\pi_e}(s_0)\mathrm{d}\lambda(\mathcal{J}_{s_{t+1}}). \tag{EC.4}$$

is uniformly upper-bounded by some value around some neighborhood of $\theta = 0$. The first term (EC.3) is equal to

$$\begin{aligned}&\lim_{T\to\infty}\int c_T(\gamma)\sum_{t=1}^{T}\mathrm{E}_{\pi_e,\theta}[\gamma^t r_t|g_{R|S,A}(r_t|s_t,a_t)|] = \mathrm{E}_{p^{(\infty)}_{e,\gamma}(s:\theta)\pi_e(a|s)p_\theta(r|s,a)}[r|g_{R|S,A}(r|s,a:\theta)|] \quad \text{(Fubini)}\\ &\le R_{\max}\mathrm{E}_{p^{(\infty)}_{e,\gamma}(s:\theta)\pi_e(a|s)}[\mathrm{E}_{p_\theta(r|s,a)}[|g_{R|S,A}(r|s,a:\theta)| \mid s,a]]\\ &\le R_{\max}\mathrm{E}_{p^{(\infty)}_{e,\gamma}(s:\theta)\pi_e(a|s)}[\int h(r|s,a)\mathrm{d}\mu(r)] < \infty,\end{aligned}$$

where $h(r|s,a)$ is some integrable function. Here we leveraged condition (d) in the definition of a one-dimensional submodel.

The second term (EC.4) is equal to

$$\begin{aligned}
&\lim_{T\to\infty}\gamma\int c_T(\gamma)\sum_{c=1}^{T}\gamma^c\sum_{t=c+1}^{T}\mathrm{E}_{\pi_e,\theta}[\gamma^{t-c-2}r_t|g_{S'|S,A}(s_{t+1}|s_t,a_t:\theta)|]\\
&=\gamma\mathrm{E}_{p_{e,\gamma}^{(\infty)}(s:\theta)\pi_e(a|s)p_\theta(s'|s,a)}[v(s')|g_{S'|S,A}(s'|s,a:\theta)|] \qquad \text{(Fubini)}\\
&\leq(1-\gamma)^{-1}R_{\max}\gamma\mathrm{E}_{p_{e,\gamma}^{(\infty)}(s:\theta)\pi_e(a|s)}[\mathrm{E}_{p_\theta(s'|s,a)}[|g_{S'|S,A}(s'|s,a:\theta)|]]\\
&\leq(1-\gamma)^{-1}R_{\max}\gamma\mathrm{E}_{p_{e,\gamma}^{(\infty)}(s:\theta)\pi_e(a|s)}[\int h(s'|s,a)\mathrm{d}(s')]<\infty,
\end{aligned}$$

where $h(s'|s,a)$ is some integrable function. Combing all together, the exchange of integration and differentiation (EC.2) is justified.

To derive the influence function, we continue the calculation:

$$\begin{aligned}
\nabla\rho^{\pi_e}(\theta)|_{\theta=0}&=\mathrm{E}_{p_{e,\gamma}^{(\infty)}}[rg_{R|S,A}(r|s,a)]+\gamma\mathrm{E}_{p_{e,\gamma}^{(\infty)}}[v(s')g_{S'|S,A}(s'|s,a)]\\
&=\mathrm{E}_{p_{e,\gamma}^{(\infty)}}[\{r-\mathrm{E}[r|s,a]\}g_{R|S,A}(r|s,a)]+\gamma\mathrm{E}_{p_{e,\gamma}^{(\infty)}}[\{v(s')-\mathrm{E}_{p_{e,\gamma}^{(\infty)}}[v(s')|s,a]\}g_{S'|S,A}(s'|s,a)]\\
&=\mathrm{E}_{p_{e,\gamma}^{(\infty)}}\left[\left\{r-\mathrm{E}_{p_{e,\gamma}^{(\infty)}}[r|s,a]+\gamma v(s')-\gamma\mathrm{E}[v(s')|s,a]\right\}g(s,a,r,s')\right]\\
&=\mathrm{E}_{p_{e,\gamma}^{(\infty)}}[\{r+\gamma v(s')-q(s,a)\}g(s,a,r,s')]\\
&=\mathrm{E}_{p_b(s,a,r,s')}[(w(s)\eta(s,a)\{r+\gamma v(s')-q(s,a)\})g(s,a,r,s')].
\end{aligned}$$

Therefore, the following function is a gradient, if its L_2-norm is finite:

$$w(s)\eta(s,a)\{r+\gamma v(s')-q(s,a)\}$$

In addition, if it exists, it belongs to $\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3+\mathcal{T}_4$ since

$$\begin{aligned}
&w(s)\eta(s,a)\{r-\mathrm{E}[r|s,a]\}\in\mathcal{T}_4,\ w(s)\eta(s,a)\{v(s')-\mathrm{E}[v(s')|s,a]\}\in\mathcal{T}_3,\\
&w(s)\eta(s,a)\{r+\gamma v(s')-q(s,a)\}=w(s)\eta(s,a)\{r-\mathrm{E}[r|s,a\}+\gamma\{v(s')-\mathrm{E}[v(s')|s,a]\}\}.
\end{aligned}$$

Thus, it is the EIF. The efficiency bound is therefore

$$\mathrm{E}_{p_b}[w^2(s)\eta^2(a,s)\{r+\gamma v(s')-q(s,a)\}^2].$$

Finally, we show that the EIF is still the same under $\mathcal{M}_{3bs},\mathcal{M}_{3b}$. First, the above EIF under $\mathcal{M}_3$ is still a gradient under $\mathcal{M}_{3bs},\mathcal{M}_{3b}$. Again, what we have to prove is that this function belongs to the tangent space. This is obvious since the tangent space of $\mathcal{M}_{3b}$ is

$$L_2\cap(\mathcal{T}_1+\mathcal{T}_3+\mathcal{T}_4),$$

the tangent space of $\mathcal{M}_{3bs}$ is

$$L_2 \cap (\mathcal{T}_3 + \mathcal{T}_4),$$

and the gradient above belongs to $L_2 \cap (\mathcal{T}_3 + \mathcal{T}_4) \subseteq L_2 \cap (\mathcal{T}_1 + \mathcal{T}_3 + \mathcal{T}_4) \subseteq L_2 \cap (\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 + \mathcal{T}_4)$.

Proof of Theorem 4 In this proof, we write $q_k^{\omega_N}$ as q_k to simplify the notation. Define:

$$\psi(\{\hat{\nu}_k\}, \{\hat{q}_k\}) = \sum_{k=0}^{\omega_N} \hat{\nu}_k r_k - \{\hat{\nu}_{k-1}\hat{q}_k - \hat{\nu}_k \mathrm{E}_{\pi_e}[\hat{q}_k(s_k, a_k)|s_k]\}.$$

The estimator $\hat{\rho}^{\mathcal{M}1}_{\mathrm{DRL}(\mathcal{M}_1)}$ is then given by

$$\frac{N_0}{N}\mathbb{P}_{\mathcal{D}_0}\psi(\{\hat{\nu}_k^{[1]}\}, \{\hat{q}_k^{[1]}\}) + \frac{N_1}{N}\mathbb{P}_{\mathcal{D}_1}\psi(\{\hat{\nu}_k^{[0]}\}, \{\hat{q}_k^{[0]}\}),$$

where $\mathbb{P}_{\mathcal{D}_j}$ is the empirical average over $\mathcal{D}_j$ and N_j is the sample size of each. Then, we have

$$\sqrt{N}(\mathbb{P}_{\mathcal{D}_0}\psi(\{\hat{\nu}_k^{[1]}\}, \{\hat{q}_k^{[1]}\}) - \rho^{\pi_e}_{\omega_N}) = \sqrt{N/N_0}\mathbb{G}_{N_0}[\psi(\{\hat{\nu}_k^{[1]}\}, \{\hat{q}_k^{[1]}\}) - \psi(\{\nu_k\}, \{q_k\})] \tag{EC.5}$$

$$+ \sqrt{N/N_0}\mathbb{G}_{N_0}[\psi(\{\nu_k\}, \{q_k\})] \tag{EC.6}$$

$$+ \sqrt{N}(\mathrm{E}[\psi(\{\hat{\nu}_k^{[1]}\}, \{\hat{q}_k^{[1]}\})|\{\hat{\nu}_k^{[1]}\}, \{\hat{q}_k^{(1)}\}] - \rho^{\pi_e}_{\omega_N}). \tag{EC.7}$$

We analyze each term. To do that, we use the following relation:

$$\begin{aligned}
&\psi(\{\hat{\nu}_k\}, \{\hat{q}_k\}) - \psi(\{\nu_k\}, \{q_k\}) = D_1 + D_2 + D_3, \quad \text{where} \\
&D_1 = \sum_{k=0}^{w_N} (\hat{\nu}_k - \nu_k)(-\hat{q}_k + q_k) + (\hat{\nu}_{k-1} - \nu_{k-1})(-\hat{v}_k + v_k), \\
&D_2 = \sum_{k=0}^{\omega_N} \nu_k(\hat{q}_k - q_k) + \nu_{k-1}(\hat{v}_k - v_k), \\
&D_3 = \sum_{k=0}^{\omega_N} (\hat{\nu}_k - \nu_k)(r_k - q_k + v_{k+1}).
\end{aligned}$$

First, we show the term (EC.5) is $o_p(1)$.

The term (EC.5) **is** $o_p(1)$**.** If we can show that for any $\epsilon > 0$,

$$\begin{aligned}
\lim_{n\to\infty} \sqrt{N_0}P[&\mathbb{P}_{\mathcal{D}_0}[\psi(\{\hat{\nu}_k^{[1]}\}, \{\hat{q}_k^{[1]}\}) - \psi(\{\nu_k\}, \{q_k\})] \\
&- \mathrm{E}[\psi(\{\hat{\nu}_k^{[1]}\}, \{\hat{q}_k^{[1]}\}) - \psi(\{\nu_k\}, \{q_k\})|\{\hat{\nu}_k^{[1]}\}, \{\hat{q}_k^{[1]}\}] > \epsilon|\mathcal{D}_1] = 0,
\end{aligned} \tag{EC.8}$$

Then, by bounded convergence theorem, we would have

$$\begin{aligned}
\lim_{n\to\infty} \sqrt{N_0}P[&\mathbb{P}_{\mathcal{D}_0}[\psi(\{\hat{\nu}_k^{[1]}\}, \{\hat{q}_k^{[1]}\}) - \psi(\{\nu_k\}, \{q_k\})] \\
&- \mathrm{E}[\psi(\{\hat{\nu}_k^{[1]}\}, \{\hat{q}_k^{[1]}\}) - \psi(\{\nu_k\}, \{q_k\})|\{\hat{\nu}_k^{[1]}\}, \{\hat{q}_k^{[1]}\}] > \epsilon] = 0,
\end{aligned}$$

yielding the statement.

To show (EC.8), we show that the conditional mean is 0 and conditional variance is $o_p(1)$. The conditional mean is

$$\begin{aligned}&\mathrm{E}[\mathbb{P}_{\mathcal{D}_0}[\psi(\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\})-\psi(\{\nu_k\},\{q_k\})] \\ &\qquad\qquad - \mathrm{E}[\psi(\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\})-\psi(\{\nu_k\},\{q_k\})\mid\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\}]|\mathcal{D}_1] \\ &=0.\end{aligned}$$

Here, we leverage the sample splitting construction, that is, $\hat{\nu}_k^{[1]}$ and $\hat{q}_k^{[1]}$ only depend on $\mathcal{D}_1$. The conditional variance is

$$\begin{aligned}&\mathrm{var}[\sqrt{N_0}\mathbb{P}_{\mathcal{D}_0}[\psi(\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\})-\psi(\{\nu_k\},\{q_k\})]|\mathcal{D}_1] \\ &\quad = \mathrm{E}[D_1^2+D_2^2+D_3^2+2D_1D_2+2D_2D_3+2D_2D_3\mid\mathcal{D}_1] \\ &\quad = \omega_N^2\max\{\mathcal{O}_p((\kappa_N^\nu)^2),\mathcal{O}_p((\kappa_N^q)^2),\mathcal{O}_p(\kappa_N^\nu\kappa_N^q)\}=o_p(1).\end{aligned}$$

Here, we used the convergence rate assumptions (4c) and the relation $\|\hat{v}_k^{[1]}-v_k\|_2\le C\|\hat{q}_k^{[1]}-q_k\|_2$ arising from the fact that the former is the marginalization of the latter over $\pi_{e,k}$ and Jensen's inequality. Then, from Chebyshev's inequality:

$$\begin{aligned}&\sqrt{N_0}P[\mathbb{P}_{\mathcal{D}_0}[\psi(\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\})-\psi(\{\nu_k\},\{q_k\})]-\mathrm{E}[\psi(\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\})-\psi(\{\nu_k\},\{q_k\})|\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\}]>\epsilon|\mathcal{D}_1] \\ &\le\frac{1}{\epsilon^2}\mathrm{var}[\sqrt{N_0}\mathbb{P}_{\mathcal{D}_0}[\psi(\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\})-\psi(\{\nu_k\},\{q_k\})]|\mathcal{D}_1]=o_p(1).\end{aligned}$$

The term Eq. (EC.7) is $o_p(1)$.

$$\begin{aligned}&\sqrt{N}\mathrm{E}[\psi(\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\})-\mathrm{E}[\psi(\{\nu_k\},\{q_k\})]|\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\}] \\ &=\sqrt{N}\mathrm{E}[\sum_{k=0}^{\omega_N}(\hat{\nu}_k^{[1]}-\nu_k)(-\hat{q}_k^{[1]}+q_k)+(\hat{\nu}_{k-1}^{[1]}-\nu_{k-1})(-\hat{v}_k+v_k)|\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\}] \\ &+\sqrt{N}\mathrm{E}[\sum_{k=0}^{\omega_N}\nu_k(\hat{q}_k^{[1]}-q_k)+\nu_{k-1}(\hat{v}_k-v_k)|\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\}] \\ &+\sqrt{N}\mathrm{E}[\sum_{k=0}^{\omega_N}(\hat{\nu}_k^{[1]}-\nu_k)(r_k-q_k+v_{k+1})|\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\}] \\ &=\sqrt{N}\mathrm{E}[\sum_{k=0}^{\omega_N}(\hat{\nu}_k^{[1]}-\nu_k)(-\hat{q}_k^{[1]}+q_k)+(\hat{\nu}_{k-1}^{[1]}-\nu_{k-1})(-\hat{v}_k+v_k)|\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\}] \\ &=\sqrt{N}\sum_{k=0}^{\omega_N}\mathcal{O}(\|\hat{\nu}_k^{[1]}-\nu_k\|_2\|\hat{q}_k^{[1]}-q_k\|_2)=\sqrt{N}\sum_{k=0}^{\omega_N}\kappa_N^\nu\kappa_N^q=o_p(1).\end{aligned}$$

Here, we have used the assumption (4e).

Combining all things Finally, we get

$$\sqrt{N}(\mathbb{P}_{\mathcal{D}_0}\psi(\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\})-\rho^{\pi_e}_{\omega_N})=\sqrt{N/N_0}\mathbb{G}_{\mathcal{D}_0}[\psi(\{\nu_k\},\{q_k\})]+o_p(1).$$

By flipping the role,

$$\sqrt{N}(\mathbb{P}_{\mathcal{D}_1}\psi(\{\hat{\nu}_k^{[0])}\},\{\hat{q}_k^{[0])}\})-\rho^{\pi_e}_{\omega_N})=\sqrt{N/N_1}\mathbb{G}_{\mathcal{D}_1}[\psi(\{\nu_k\},\{q_k\})]+o_p(1).$$

Therefore,

$$\begin{aligned}
&\sqrt{N}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}-\rho^{\pi_e}_{\omega_N})\\
&=N_0/N\times\sqrt{N}\psi(\{\hat{\nu}_k^{[1]}\},\{\hat{q}_k^{[1]}\})-\rho^{\pi_e}_{\omega_N})+N_1/N\times\sqrt{N}(\mathbb{P}_{\mathcal{D}_1}\psi(\{\hat{\nu}_k^{[0]}\},\{\hat{q}_k^{[0]}\})-\rho^{\pi_e}_{\omega_N})\\
&=\sqrt{N_0/N}\mathbb{G}_{N_0}[\psi(\{\nu_k\},\{q_k\})]+\sqrt{N_1/N}\mathbb{G}_{N_1}[\psi(\{\nu_k\},\{q_k\})]+o_p(1)\\
&=\mathbb{G}_N[\psi(\{\nu_k\},\{q_k\})]+o_p(1).
\end{aligned}$$

Finally, from the assumption (4a),(4b)(4d) and CLT, the efficiency is concluded. This estimator is also regular from Theorem EC.1.

Proof of Theorem 5 The proof is similar to that of Theorem 4

Proof of Theorems 6 and 7 First, We provide the proof the cross-fitting version. Then, we provide the proof of the adaptive version.

In the cross-fitting version, the estimator is given by

$$\frac{n_0}{n}\mathbb{P}_{\mathcal{D}_0}\psi(\hat{w}^{[1]},\hat{q}^{[1]})+\frac{n_1}{n}\mathbb{P}_{\mathcal{D}_1}\psi(\hat{w}^{[0]},\hat{q}^{[0]})$$

where $\mathbb{P}_{\mathcal{D}_0}$ is a sample average over a set of samples in one fold, and $\mathbb{P}_{\mathcal{D}_1}$ is a sample average over a set of samples in another fold. Then, we have

$$\sqrt{n}(\mathbb{P}_{\mathcal{D}_0}\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\rho^{\pi_e})=\sqrt{n/n_0}\mathbb{G}_{n_0}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w^{[1]},q^{[1]})] \tag{EC.9}$$

$$+\sqrt{n/n_0}\mathbb{G}_{n_0}[\psi(w^{[1]},q^{[1]})] \tag{EC.10}$$

$$+\sqrt{n}(\mathrm{E}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})\mid\hat{w}^{[1]},\hat{q}^{[1]}]-\rho^{\pi_e}). \tag{EC.11}$$

We analyze each term. Here, we have

$$\begin{aligned}
&\psi(\hat{w},\hat{q})-\psi(w,q)=D_1+D_2+D_3,\\
&\qquad D_1=\{\hat{w}(s)-w(s)\}\eta(s,a)\{r-q(s,a)+\gamma v(s')\},\\
&\qquad D_2=w(s)\eta(s,a)\{q(s,a)-\hat{q}(s,a)+\gamma\hat{v}(s')-\gamma v(s')\}+(1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[v'(s)-v(s)],\\
&\qquad D_3=\{\hat{w}(s)-w(s)\}\eta(s,a)\{q(s,a)-\hat{q}(s,a)+\gamma\hat{v}(s')-\gamma v(s')\}.
\end{aligned}$$

Term (EC.9) **is** $\mathrm{o}_p(1)$ We show that the conditional variance given $\mathcal{D}_1$ is $\mathrm{o}_p(1)$. The rest of the argument is the same as the proof of Theorem 4. The conditional variance is

$$\begin{aligned}\mathrm{var}[\sqrt{n_0}\mathbb{P}_{\mathcal{D}_0}[\psi(\hat{w},\hat{q})-\psi(w,q)]|\mathcal{D}_1] &\leq \mathbb{E}[D_1^2+D_2^2+D_2^3+2D_1D_2+2D_2D_3+2D_1D_3|\mathcal{D}_1]\\ &= \mathrm{o}_p(1).\end{aligned}$$

Here, we use

$$\|\hat{v}(s')-v(s')\|_2 \leq \sqrt{C_{S'}}\|\hat{q}(s,a)-q(s,a)\|_2.$$

from Jensen's inequality.

Term (EC.11) **is** $\mathrm{o}_p(1)$

$$\begin{aligned}|\mathbb{E}[D_1+D_2+D_3|\mathcal{D}_1]| &\leq \|\hat{w}-w\|_2\|\eta\|_2\|\hat{q}-q\|_2+\|\hat{w}-w\|_2\|\eta\|_2\|\hat{v}(s')-v(s)\|_2\\ &\leq \|\hat{w}-w\|_2\|\eta\|_2\|\hat{q}-q\|_2+\sqrt{\gamma C'_S}\|\hat{w}-w\|_2\|\eta\|_2\|\hat{q}-q\|_2\\ &= \kappa_n^w\kappa_n^q = \mathrm{o}_p(n^{-1/2}).\end{aligned}$$

Proving efficiency

$$\begin{aligned}&\sqrt{n}(\mathbb{P}_{\mathcal{D}_0}\psi(\hat{w}^{[1]},\hat{q}^{[1]})+\mathbb{P}_{\mathcal{D}_1}\psi(\hat{w}^{[0]},\hat{q}^{[0]})-\rho^{\pi_e})\\ &=\sqrt{n/n_0}\mathbb{G}_{n_0}[\psi(w,q)]+\sqrt{n/n_1}\mathbb{G}_{n_1}[\psi(w,q)]+\mathrm{o}_p(n^{-1/2})\\ &=\mathbb{G}_n[\psi(w,q)]+\mathrm{o}_p(n^{-1/2}).\end{aligned}$$

Finally, from CLT, the final statement is concluded.

Proof of adaptive version The adaptive version is similarly proved.

$$\sqrt{n}(\mathbb{P}_n\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\rho^{\pi_e}) = \mathbb{G}_n[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w^{[1]},q^{[1]})] \tag{EC.12}$$

$$+\mathbb{G}_n[\psi(w^{[1]},q^{[1]})] \tag{EC.13}$$

$$+\sqrt{n}(\mathrm{E}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})\mid \hat{w}^{[1]},\hat{q}^{[1]}]-\rho^{\pi_e}). \tag{EC.14}$$

The third term is $\mathrm{o}_p(1)$ following the similar logic as before. We prove the first term is $\mathrm{o}_p(1)$. From van der Vaart (1998, Lemma 19.24), what we need to prove is

$$\mathbb{F}=\{\psi(w,q): w\in\mathcal{F}_w, q\in\mathcal{F}_q\}$$

belongs to a Donsker class. If we can prove the uniform entropy integral:

$$\int_0^1\sqrt{\log\sup_U \mathcal{N}(\tau,\mathbb{F},L_2(U))d(\tau)},$$

is finite, from van der Vaart (1998, Theorem 19.14), the class $\mathbb{F}$ belongs to a Donsker class. Here,

$$\begin{aligned}\log\sup_U \mathcal{N}(\tau,\mathbb{F},L_2(U)) &\leq \log\sup_U \mathcal{N}(\tau,\mathbb{F},L_2(U))\\ &\leq C\{\log\mathcal{N}(\tau,\mathcal{F}_w,L_\infty(U))+\log\sup_U\mathcal{N}(\tau,\mathcal{F}_q,L_\infty(\cdot))\}\\ &\leq C(1/\tau)^\beta < C(1/\tau)^2.\end{aligned}$$

From the second line to the third line, we use Uehara et al. (2021, Leemma 9). Thus, the uniform entropy integral is finite, which leads to the conclusion that first term (EC.12) is $\mathrm{o}_p(1)$.

Proof of Theorem 8 We provide the proof the cross-fitting version. The estimator is given by

$$\frac{n_0}{n}\mathbb{P}_{\mathcal{D}_0}\psi(\hat{w}^{[1]},\hat{q}^{[1]})+\frac{n_1}{n}\mathbb{P}_{\mathcal{D}_1}\psi(\hat{w}^{[0]},\hat{q}^{[0]})$$

where $\mathbb{P}_{\mathcal{D}_0}$ is a sample average over a set of samples in one fold, and $\mathbb{P}_{\mathcal{D}_1}$ is a sample average over a set of samples in another fold. Then, we have

$$\mathbb{P}_{\mathcal{D}_0}\psi(\hat{w}^{[1]},\hat{q}^{[1]}) = \sqrt{1/n_0}\mathbb{G}_{n_0}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w^\dagger,q^\dagger)] \tag{EC.15}$$

$$+\mathrm{E}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})\mid \hat{w}^{[1]},\hat{q}^{[1]}]-\mathrm{E}[\psi(w^\dagger,q^\dagger)] \tag{EC.16}$$

$$+\mathbb{P}_{\mathcal{D}_0}\psi(w^\dagger,q^\dagger). \tag{EC.17}$$

Then, we have

$$\begin{aligned}\mathrm{E}[\psi(\hat{w},\hat{q})|\hat{w},\hat{q}] &= \mathrm{E}[\{\hat{w}(s)-w^\dagger(s)\}\eta(s,a)\{r-q^\dagger(s,a)+\gamma v^\dagger(s')\}]+\\ &+\mathrm{E}[w^\dagger(s)\eta(s,a)\{q^\dagger(s,a)-\hat{q}(s,a)+\gamma\hat{v}(s')-\gamma v^\dagger(s')\}]+(1-\gamma)\mathrm{E}_{p_e^{(0)}}[v'(s)-v^\dagger(s)]\\ &+\mathrm{E}[\{\hat{w}(s)-w^\dagger(s)\}\eta(s,a)\{q^\dagger(s,a)-\hat{q}(s,a)+\gamma\hat{v}(s')-\gamma v^\dagger(s')\}]\\ &+\mathrm{E}[\{\hat{v}(s')-v^\dagger(s')-\hat{q}(s,a)+q^\dagger(s,a)\}|\hat{w},\hat{q}]\\ &+\mathrm{E}[\psi(w^\dagger,q^\dagger)|\hat{w},\hat{q}]\\ &=\mathcal{O}(\|\hat{w}(s)-w^\dagger(s)\|_2,\|\hat{q}(s,a)-q^\dagger(s,a)\|_2,\|\hat{v}(s')-v^\dagger(s')\|_2)+\mathrm{E}[\psi(w^\dagger,q^\dagger)]\\ &=\mathrm{o}_p(1)+\mathrm{E}[\psi(w^\dagger,q^\dagger)].\end{aligned}$$

Thus, the term (EC.16) is $\mathrm{o}_p(1)$. Thus, the remaining part is proving when at least one model is well-specified, the term $\mathbb{P}_{\mathcal{D}_0}\psi(w^\dagger,q^\dagger)$ is consistent.

q-model is well-specified . Consider the case where $q^\dagger(s,a)=q(s,a)$;

$$\begin{aligned}&\mathrm{E}[\psi(w^\dagger,q^\dagger)] = \mathrm{E}[(1-\gamma)\mathrm{E}_{p_e^{(0)}}[v(s)]+w^\dagger(s)\eta(s,a)\{r+\gamma v(s')-q(s,a)\}]\\ &=(1-\gamma)\mathrm{E}_{p_{\pi_e}^{(0)}}[v(s)]=\rho^{\pi_e}.\end{aligned}$$

This implies when the q-model is consistent, the estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ is also consistent.

w-model is well-specified. Consider the case where $w^\dagger(s)=w(s)$;

$$
\begin{aligned}
\mathrm{E}[\psi(w^\dagger,q^\dagger)] &= \mathrm{E}[(1-\gamma)\mathrm{E}_{p_e^{(0)}}[v^\dagger(s)] + w(s)\eta(s,a)\{r+\gamma v^\dagger(s')-q^\dagger(s,a)\}] \\
&= (1-\gamma)\mathrm{E}_{p_e^{(0)}}[v^\dagger(s)] + \mathrm{E}[w(s)\eta(s,a)r] + \mathrm{E}[w(s)\eta(s,a)\gamma v^\dagger(s')] \qquad \text{(EC.18)} \\
&- \mathrm{E}[w(s)\eta(s,a)q^\dagger(s,a)] \\
&= \mathrm{E}[w(s)\eta(s,a)r] + \mathrm{E}[w(s)v^\dagger(s)] - \mathrm{E}[w(s)\eta(s,a)q^\dagger(s,a)] \qquad \text{(EC.19)} \\
&= \mathrm{E}[w(s)\eta(s,a)r] = \rho^{\pi_e}. \qquad \text{(EC.20)}
\end{aligned}
$$

From (EC.18) to (EC.19), we use a result

$$(1-\gamma)\mathrm{E}_{p_{\pi_e}^{(0)}}[v^\dagger(s)] + \mathrm{E}[w(s)\eta(s,a)\gamma v^\dagger(s')] = \mathrm{E}[w(s)v^\dagger(s)]$$

from Lemma 1. From (EC.19) to (EC.20), we use a result

$$\mathrm{E}[w(s)\eta(s,a)q^\dagger(s,a)] = \mathrm{E}[w(s)\mathrm{E}[\eta(s,a)q^\dagger(s,a)|s]] = \mathrm{E}[w(s)v^\dagger(s)].$$

This implies when the ratio model is correct, the estimator $\hat\rho_{\mathrm{DRL}(\mathcal{M}_3)}$ is also consistent.

Proof of Theorem 9 We only provide the proof the cross-fitting version. The adaptive version is similarly proved. The estimator is given by

$$\frac{n_0}{n}\mathbb{P}_{\mathcal{D}_0}\psi(\hat w^{[1]},\hat q^{[1]}) + \frac{n_1}{n}\mathbb{P}_{\mathcal{D}_1}\psi(\hat w^{[0]},\hat q^{[0]}).$$

We have the following decomposition:

$$
\begin{aligned}
\sqrt{n}(\mathbb{P}_{\mathcal{D}_0}\psi(\hat w^{[1]},\hat q^{[1]}) - \rho^{\pi_e}) &= \sqrt{n/n_0}\{\mathbb{G}_{n_0}\psi(\hat w^{[1]},\hat q^{[1]}) - \mathbb{G}_{n_0}\psi(w^\dagger,q^\dagger)\}+ \qquad \text{(EC.21)} \\
&+ \sqrt{n}(\mathrm{E}[\psi(\hat w^{[1]},\hat q^{[1]})|\hat w^{[1]},\hat q^{[1]}] - \mathrm{E}[\psi(w^\dagger,q^\dagger)]) \qquad \text{(EC.22)} \\
&+ \sqrt{n}(\mathrm{E}[\psi(w^\dagger,q^\dagger)] - \rho^{\pi_e}) \qquad \text{(EC.23)} \\
&+ \sqrt{n/n_0}\mathbb{G}_{n_0}\psi(w^\dagger,q^\dagger).
\end{aligned}
$$

The term (EC.21) is $\mathrm{o}_p(1)$ as in the proof of Theorem 6. The term (EC.22) is $\mathcal{O}_p(1)$ as in the proof of Theorem 8:

$$
\begin{aligned}
\mathrm{E}[\psi(\hat w^{[1]},\hat q^{[1]})|\hat w^{[1]},\hat q^{[1]}] &= \mathcal{O}(\|\hat w^{[1]}(s)-w^\dagger(s)\|_2, \|\hat q^{[1]}(s,a)-q^\dagger(s,a)\|_2, \|\hat v^{[1]}(s')-v^\dagger(s')\|_2) + \mathrm{E}[\psi(w^\dagger,q^\dagger)] \\
&= \mathcal{O}_p(n^{-1/2}) + \mathrm{E}[\psi(w^\dagger,q^\dagger)].
\end{aligned}
$$

The term (EC.23) is 0 following the argument in the proof of Theorem 8. Then, we obtain $\sqrt{n}(\hat\rho_{\mathrm{DRL}(\mathcal{M}_3)} - \rho^{\pi_e}) = \mathbb{G}_n\psi(w^\dagger,q^\dagger) + \mathcal{O}_p(1) = \mathcal{O}_p(1)$.

Proof of Theorem 10 For the ease of the notation, we assume $N=1$. The extension to general N is straightforward.

Under the assumptions, $\psi(s,a,r,s';w,q)$ is bounded. Invoking Ibragimov and Linnik (1971, Theorem 18.5.4), we obtain that $\sqrt{T}\,(\mathbb{P}_T[\psi(s,a,r,s';w,q)]-\rho^{\pi_e})$ converges to a normal distribution with zero mean and the variance:

$$\mathrm{var}[\phi_{\mathrm{eff}}(s_0,a_0,r_0,s_1)]+2\sum_{i=1}^{\infty}\mathrm{cov}[\phi_{\mathrm{eff}}(s_0,a_0,r_0,s_1),\phi_{\mathrm{eff}}(s_i,a_i,r_i,s_{i+1})],$$

where ϕ_{eff} is as in Theorem 3. The first term is $\mathrm{EB}(\mathcal{M}_3)$. The second term is zero because

$$\begin{aligned}&\mathrm{E}[\phi_{\mathrm{eff}}(s_0,a_0,r_0,s_1)\phi_{\mathrm{eff}}(s_i,a_i,r_i,s_{i+1})]\\&=\mathrm{E}[\phi_{\mathrm{eff}}(s_0,a_0,r_0,s_1)\mathrm{E}[\phi_{\mathrm{eff}}(s_i,a_i,r_i,s_{i+1})\mid s_{i-1},a_{i-1},r_{i-1},s_i]]=0.\end{aligned}$$

Proof of Theorem 11 Using the short-hand $\psi(w',q')$ for $\psi(s,a,r,s';w',q')$ and $N_j=|\mathcal{D}_j|$, the estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}3)}$ is given by

$$\frac{N_0}{N}\mathbb{P}_{\mathcal{D}_0}\mathbb{P}_T\psi(\hat{w}^{[1]},\hat{q}^{[1]})+\frac{N_1}{N}\mathbb{P}_{\mathcal{D}_1}\mathbb{P}_T\psi(\hat{w}^{[0]},\hat{q}^{[[0]}),$$

where $\mathbb{P}_{\mathcal{D}_j}$ is the empirical average on the samples in $\mathcal{D}_j$.

Then, we have

$$\begin{aligned}\sqrt{NT}(\mathbb{P}_{\mathcal{D}_0}\mathbb{P}_T\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\rho^{\pi_e})&=\sqrt{N/N_0}\mathbb{G}_{\mathcal{D}_0,T}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)] &\text{(EC.24)}\\&+\sqrt{N/N_0}\mathbb{G}_{\mathcal{D}_0,T}[\psi(w,q)]\\&+\sqrt{NT}(\mathrm{E}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})|\hat{w}^{[1]},\hat{q}^{[1]}]-\rho^{\pi_e}), &\text{(EC.25)}\end{aligned}$$

where $\mathbb{G}_{\mathcal{D}_0,T}$ is an empirical process defined over the all sample in the first fold. Here, we have

$$\begin{aligned}&\mathbb{P}_T\psi(\hat{w},\hat{q})-\mathbb{P}_T\psi(w,q)\\&=\mathbb{P}_T[\{\hat{w}(s)-w(s)\}\eta(s,a)\{r-q(s,a)+\gamma v(s')\}]\\&\quad+\mathbb{P}_T[w(s)\eta(s,a)\{q(s,a)-\hat{q}(s,a)+\gamma\hat{v}(s')-\gamma v(s')\}]+(1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[v'(s)-v(s)]\\&\quad+\mathbb{P}_T[\{\hat{w}(s)-w(s)\}\eta(s,a)\{q(s,a)-\hat{q}(s,a)+\gamma\hat{v}(s')-\gamma v(s')\}].\end{aligned}$$

We analyze each term. First, we show the term Eq. (EC.24) is $o_p(1)$.

***The term* (EC.24) *is* $o_p(1)$** Consider the case $N_0=1$. The case with $N_0>1$ similarly holds. If we can show that for any $\epsilon>0$,

$$\begin{aligned}\lim_{T\to\infty}\sqrt{T}P[\mathbb{P}_T[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)]& &\text{(EC.26)}\\-\mathrm{E}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)|\hat{w}^{[1]},\hat{q}^{[1]}]>\epsilon|\mathcal{D}_1]=0,\end{aligned}$$

then, by bounded convergence theorem, we would have

$$\lim_{T\to\infty}\sqrt{T}P[\mathbb{P}_T[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)] \\ -\mathrm{E}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)|\hat{w}^{[1]},\hat{q}^{[1]}]>\epsilon]=0,$$

which yields the statement.

To show Eq. (EC.26), we show that the conditional mean is 0 and conditional variance is $o_p(1)$. The conditional mean is

$$\mathrm{E}[\mathbb{P}_T[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)|\hat{w}^{[1]},\hat{q}^{[1]}] \\ -\mathbb{P}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)]|\mathcal{D}_1]=0.$$

Here, we leverage the sample-splitting construction, that is, $\hat{w}^{[1]}$ and $\hat{q}^{[1]}$ only depend on $\mathcal{D}_1$, and $\mathcal{D}_1$, $\mathcal{D}_0$ are independent. The conditional variance is

$$\begin{aligned}&\mathrm{var}[\sqrt{T}\mathbb{P}_T[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)]|\mathcal{D}_1]=\mathrm{E}\left[T\left\{\mathbb{P}_T[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)]\right\}^2|\mathcal{D}_1\right]\\&=\frac{1}{T}\left[\sum_{i=0}^{T}\max\{\mathcal{O}_p((\kappa_N^w)^2),\mathcal{O}_p((\kappa_N^q)^2)\}+2\sum_{i<j}^{T}\rho_{\|i-j\|}\max\{\mathcal{O}_p((\kappa_N^w)^2),\mathcal{O}_p((\kappa_N^q)^2)\}\right]\\&=o_p(1).\end{aligned}$$

Then, from Chebyshev's inequality:

$$\begin{aligned}&\sqrt{T}P[\mathbb{P}_T[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)]-\mathrm{E}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)|\hat{w}^{[1]},\hat{q}^{[1]}]>\epsilon|\mathcal{D}_1]\\&\leq\frac{1}{\epsilon^2}\mathrm{var}[\sqrt{T}\mathbb{P}_T[\psi(\hat{w}^{[1]},\hat{q}^{[1]})-\psi(w,q)]|\mathcal{D}_1]=o_p(1).\end{aligned}$$

Second Term We show

$$\sqrt{NT}(\mathrm{E}[\psi(\hat{w}^{[1]},\hat{q}^{[1]})|\hat{w}^{[1]},\hat{q}^{[1]}]-\rho^{\pi_e})=\mathrm{o}_p(1).$$

Assume $N_0=1$ for simplicity. The case with $N_0>1$ similarly holds. Noting $\mathrm{E}[(\hat{q}(s,a)-q(s,a))^2]\geq\mathrm{E}[(\hat{v}(s')-v(s'))^2]$ from Jensen's inequality, we have

$$\begin{aligned}&|\sqrt{T}\mathrm{E}[\psi(s,a,r,s';\hat{w}^{[1]},\hat{q}^{[1]})-\psi(s,a,r,s';w,q)|\hat{w}^{[1]},\hat{q}^{[1]}]|\\=&|\sqrt{T}\mathrm{E}[(\hat{w}(s)-w(s))\eta(s,a)(r-q(s,a)+\gamma v(s'))|\hat{w}^{[1]},\hat{q}^{[1]}]+\\&\sqrt{T}\mathrm{E}[w(s)\eta(s,a)\{q(s,a)-\hat{q}(s,a)+\gamma\hat{v}(s')-\gamma v(s')\}+(1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[\hat{v}(s)-v(s)]||\hat{w},\hat{q}]+\\&\sqrt{T}\mathrm{E}[\{\hat{w}(s)-w(s)\}\eta(s,a)\{q(s,a)-\hat{q}(s,a)+\gamma\hat{v}(s')-\gamma v(s')\}|\hat{w}^{[1]},\hat{q}^{[1]}]|\\=&|\sqrt{T}\mathrm{E}[\{\hat{w}(s)-w(s)\}\eta(s,a)\{q(s,a)-\hat{q}(s,a)+\gamma\hat{v}(s')-\gamma v(s')\}|\hat{w}^{[1]},\hat{q}^{[1]}]|\\\leq&\sqrt{T}\|\hat{w}^{[1]}(s)-w(s)\|_2\|\eta(s,a)\|_2\|q(s,a)-\hat{q}^{[1]}(s,a)\|_2=o_p(1).\end{aligned}$$

Combining all results Combining all result,

$$\sqrt{NT}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}-\rho^{\pi_e})=\mathbb{G}_{N,T}\psi(w,q)+\mathrm{o}_p(1).$$

Then, from Eq. (EC.1) and Ibragimov and Linnik (1971, Theorem 18.5.4) the statement is concluded as in Theorem 10.

Proof of Theorem 12 We focus on $N=1$, which is most relevant for cross-time-fitting; also $N>1$ easily follows similarly. Using the short-hand $\psi(w',q')$ for $\psi(s,a,r,s';w',q')$ and $T_j=|\mathcal{T}_j|$, the estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}3)}$ is given by

$$\frac{T_0}{T}\mathbb{P}_{\mathcal{T}_0}\psi(\hat{w}^{[2]},\hat{q}^{[2]})+\frac{T_1}{T}\mathbb{P}_{\mathcal{T}_1}\psi(\hat{\nu}^{[3]},\hat{q}^{[3]})+\frac{T_2}{T}\mathbb{P}_{\mathcal{T}_2}\psi(\hat{\nu}^{[0]},\hat{q}^{[0]})+\frac{T_3}{T}\mathbb{P}_{\mathcal{T}_3}\psi(\hat{\nu}^{[1]},\hat{q}^{[1]}),$$

where $\mathbb{P}_{\mathcal{T}_j}$ is the empirical average on the samples in $\mathcal{T}_j$.

The proof now proceeds as in Theorem 11. In particular, we have to deal with the dependence across folds more carefully.

First, we show the analysis of the stochastic equicontinuity term. To prove it, we must leverage the fact that $\mathcal{T}_0$ and $\mathcal{T}_1$ are separated by $T/4$ time steps, rather than being independent, unlike Theorem 11.

First part: $\mathbb{G}_{\mathcal{T}_0}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)]=\mathrm{o}_p(1)$

Proof If we can show that for any $\epsilon>0$,

$$\begin{aligned}\lim_{T\to\infty}\sqrt{T_0}P[\mathbb{P}_{\mathcal{T}_0}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)]&\\ -\mathrm{E}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)|\hat{w}^{[2]},\hat{q}^{[2]}]>\epsilon|\mathcal{T}_2]=0,&\end{aligned}\tag{EC.27}$$

Then, by bounded convergence theorem, we would have

$$\begin{aligned}\lim_{T\to\infty}\sqrt{T_0}P[\mathbb{P}_{\mathcal{T}_0}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)]&\\ -\mathrm{E}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)|\hat{w}^{[2]},\hat{q}^{[2]}]>\epsilon]=0,&\end{aligned}$$

yielding the statement.

To show (EC.27), we show that the conditional mean given $\mathcal{T}_2$ is $o_p(1)$ and conditional variance given $\mathcal{T}_2$ is $o_p(1)$. The conditional mean given $\mathcal{T}_2$ is

$$\begin{aligned}&|\mathbb{E}[\mathbb{P}_{\mathcal{T}_0}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)]-\mathbb{E}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)]|\mathcal{T}_2]|\\ &=|\mathbb{E}[\mathbb{P}_{\mathcal{T}_0}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)]|\mathcal{T}_2]-\mathbb{E}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)\mid\hat{w}^{[2]},\hat{q}^{[2]}]| &&\text{(EC.28)}\\ &\leq|\mathbb{P}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)|\mid\hat{w}^{[2]},\hat{q}^{[2]}]-\mathbb{E}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)\mid\hat{w}^{[2]},\hat{q}^{[2]}]|+\mathrm{o}_p(1) &&\text{(EC.29)}\\ &=o_p(1).\end{aligned}$$

Here, we leverage the sample splitting construction, that is, $\hat{w}_k^{[2]}$ and $\hat{q}_k^{[2]}$ only depend on $\mathcal{T}_2$, and the correlation between the train data and test data set is small. To go from Eq. (EC.28) to Eq. (EC.29), we use the following argument. First, note $\alpha_m \le \phi_m$ by Eq. (EC.1). In addition, from the definition of α-mixing and its moment inequality, for any bounded function $f(x)$ with a first moment, based on Davidson (1994, Theorem 14.2), we have

$$\|\mathbb{E}[f(s_t,a_t,s_{t+1},r_t)\mid s_0,a_0,s_1,r_0]-\mathbb{E}[f(s_t,a_t,s_{t+1},r_t)]\|_1 \le 6\|f(s_t,a_t,s_{t+1},r_t)\|_1.$$

Note that since α-mixing is time-reversible (Davidson 1994, Page 209), the following also holds:

$$\|\mathbb{E}[f(s_0,a_0,s_1,r_1)\mid s_t,a_t,s_{t+1},r_t]-\mathbb{E}[f(s_0,a_0,s_1,r_1)]\|_1 \le 6\|f(s_0,a_0,s_1,r_1)\|_1.$$

By applying this moment inequality to Eq. (EC.28) and noting ψ is a bounded function we have

$$\begin{aligned}
&|\mathbb{E}[\mathbb{P}_{\mathcal{T}_0}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)]\mid \mathcal{T}_2]-\mathbb{E}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)|\hat{w}^{[2]},\hat{q}^{[2]}]| \\
&=|\frac{1}{T_0}\sum_{i\in T_0}\int\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]})-\psi(z_i;w,q)\}p(z_i\mid\mathcal{T}_2)\mathrm{d}(z_i)-\int\left\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]})-\psi(z_i;w,q)\right\}p(z_i)\mathrm{d}(z_i)| \\
&\le \frac{1}{T_0}\sum_{i\in T_0}|\int\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]})-\psi(z_i;w,q)\}p(z_i\mid\mathcal{T}_2)\mathrm{d}(z_i)-\int\left\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]})-\psi(z_i;w,q)\right\}p(z_i)\mathrm{d}(z_i)| \\
&\le 6\int|\left\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]})-\psi(z_i;w,q)\right\}|p(z_i)\mathrm{d}(z_i) \qquad (\alpha\text{-mixing moment inequality}) \\
&=6\mathbb{E}[|\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)|\mid\hat{w}^{[2]},\hat{q}^{[2]}]=\mathrm{o}_p(1).
\end{aligned}$$

Then,

$$\mathbb{E}[\mathbb{P}_{\mathcal{T}_0}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)]\mid\mathcal{T}_2]=\mathbb{E}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)|\hat{w}^{[2]},\hat{q}^{[2]}]+\mathrm{o}_p(1).$$

Besides, the conditional variance is

$$\begin{aligned}
&\mathrm{var}[\sqrt{T_0}\mathbb{P}_{\mathcal{T}_0}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)]|\mathcal{T}_2]\le \mathrm{E}[T_0\{\mathbb{P}_{\mathcal{T}_0}[\psi(\hat{w}^{[2]},\hat{q}^{[2]})-\psi(w,q)]\}^2|\mathcal{T}_2] \\
&\le\frac{1}{T_0}\sum_{i,j\in\mathcal{T}_0}\int|\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]})-\psi(z_i;w,q)\}\{\psi(z_j:\hat{w}^{[2]},\hat{q}^{[2]})-\psi(z_j;w,q)\}p(z_i,z_j\mid\mathcal{T}_2)\mathrm{d}(z_i,z_j)| && \text{(EC.30)}\\
&=\frac{1}{T_0}\sum_{i,j\in\mathcal{T}_0}\int|\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]})-\psi(z_i;w,q)\}\{\psi(z_j:\hat{w}^{[2]},\hat{q}^{[2]})-\psi(z_j;w,q)\}p(z_i,z_j)\mathrm{d}(z_i,z_j)|+\mathrm{o}_p(1) && \text{(EC.31)}\\
&=\frac{1}{T_0}\left[\sum_{i=0}^{T_0}\max\{\mathcal{O}_p((\kappa_N^w)^2),\mathcal{O}_p((\kappa_N^q)^2)\}+2\sum_{i<j}^{T_0}\rho_{\|i-j\|}\max\{\mathcal{O}_p((\kappa_N^w)^2),\mathcal{O}_p((\kappa_N^q)^2)\}\right]+o_p(1) && \text{(EC.32)}\\
&=o_p(1). && \text{(EC.33)}
\end{aligned}$$

Here, we used a α-mixing condition and its moment inequality from (EC.30) to (EC.31). Then, we used a ρ-mixing condition based on $\rho_t\le 2\sqrt{\phi_t}=2/t^{1+\epsilon}$ (see Eq. (EC.1)) from (EC.32) to (EC.33).

Finally, based on the obtained conditional mean and conditional variance, from Chebyshev's inequality, the rest of the proof is concluded.

Second part We prove

$$\mathrm{E}[\mathbb{P}_{\mathcal{T}_0}[\psi(s,a,r,s';\hat{w}^{[2]},\hat{q}^{[2]})] \mid \hat{w}^{[2]},\hat{q}^{[2]}] - \mathrm{E}[\psi(s,a,r,s';w,q)] = \mathrm{o}_p(T^{-1/2}).$$

Note from Davidson (1994, Theorem 14.2),

$$\|\mathbb{E}[f(s_0,a_0,s_1,r_1) \mid s_t,a_t,s_{t+1},r_t] - \mathbb{E}[f(s_0,a_0,s_1,r_1)]\|_1 \le 6\alpha_{T_0}^{1/2}\|f(s_0,a_0,s_1,r_1)\|_2.$$

Then,

$$\begin{aligned}
&|\mathbb{E}[\mathbb{P}_{\mathcal{T}_0}[\psi(\hat{w}^{[2]},\hat{q}^{[2]}) - \psi(w,q)] \mid \mathcal{T}_2] - \mathbb{E}[\psi(\hat{w}^{[2]},\hat{q}^{[2]}) - \psi(w,q)|\hat{w}^{[2]},\hat{q}^{[2]}]| \\
&= \left|\mathbb{E}\left[\frac{1}{T_0}\sum_{i\in T_0}\int\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]}) - \psi(z_i;w,q)\}p(z_i \mid \mathcal{T}_2)\mathrm{d}(z_i) - \int\left\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]}) - \psi(z_i;w,q)\right\}p(z_i)\mathrm{d}(z_i)\right]\right| \\
&\le \frac{1}{T_0}\sum_{i\in T_0}\mathbb{E}[|\int\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]}) - \psi(z_i;w,q)\}p(z_i \mid \mathcal{T}_2)\mathrm{d}(z_i) - \int\left\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]}) - \psi(z_i;w,q)\right\}p(z_i)\mathrm{d}(z_i)|] \\
&\le 6\int|\left\{\psi(z_i;\hat{w}^{[2]},\hat{q}^{[2]}) - \psi(z_i;w,q)\right\}|p(z_i)\mathrm{d}(z_i) \qquad (\alpha\text{-mixing moment inequality}) \\
&= 6\alpha_{T_0}^{1/2}\mathbb{E}[|\psi(\hat{w}^{[2]},\hat{q}^{[2]}) - \psi(w,q)|^2 \mid \hat{w}^{[2]},\hat{q}^{[2]}] = \mathrm{o}_p(1/\sqrt{T_0}). \qquad (\alpha_{T_0} \le \phi_{T_0})
\end{aligned}$$

To sum up, we have

$$\mathbb{E}[\mathbb{P}_{\mathcal{T}_0}[\psi(\hat{w}^{[2]},\hat{q}^{[2]}) - \psi(w,q)] \mid \mathcal{T}_2] = \mathbb{E}[\psi(\hat{w}^{[2]},\hat{q}^{[2]}) - \psi(w,q)|\hat{w}^{[2]},\hat{q}^{[2]}] + \mathrm{o}_p(T_0^{-1/2}). \tag{EC.34}$$

Using this above, we have

$$\begin{aligned}
&\sqrt{T}\mathrm{E}[\mathbb{P}_{\mathcal{T}_0}[\psi(s,a,r,s';\hat{w}^{[2]},\hat{q}^{[2]})] \mid \hat{w}^{[2]},\hat{q}^{[2]}] \\
&= \sqrt{T}\mathrm{E}[\mathrm{E}[\mathbb{P}_{\mathcal{T}_0}[\psi(s,a,r,s';\hat{w}^{[2]},\hat{q}^{[2]})] \mid \hat{w}^{[2]},\hat{q}^{[2]},\mathcal{T}_2] \mid \hat{w}^{[2]},\hat{q}^{[2]}] \\
&= \sqrt{T}\mathrm{E}[\mathrm{E}[\mathbb{P}_{\mathcal{T}_0}[\psi(s,a,r,s';\hat{w}^{[2]},\hat{q}^{[2]})] \mid \mathcal{T}_2] \mid \hat{w}^{[2]},\hat{q}^{[2]}] \\
&= \sqrt{T}\mathrm{E}[\mathrm{E}[\mathbb{P}_{\mathcal{T}_0}[\psi(s,a,r,s';\hat{w}^{[2]},\hat{q}^{[2]})] \mid \hat{w}^{[2]},\hat{q}^{[2]}] \mid \hat{w}^{[2]},\hat{q}^{[2]}] + \mathrm{o}_p(1) \qquad (\text{Use Eq. (EC.34)}) \\
&= \sqrt{T}\mathrm{E}[\psi(s,a,r,s';\hat{w}^{[2]},\hat{q}^{[2]}) \mid \hat{w}^{[2]},\hat{q}^{[2]}] + \mathrm{o}_p(1) \\
&= \sqrt{T}\mathrm{E}[\psi(s,a,r,s';w,q)] + \mathrm{o}_p(1). \qquad (\text{The same as the proof of Theorem 11})
\end{aligned}$$

Summary Combining, we have

$$\sqrt{T}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} - \rho^{\pi_e}) = \mathbb{G}_T\psi(w,q) + \mathrm{o}_p(1).$$

The statement is concluded as in Theorem 10.

Proof of Theorem 13 For the ease of notation we assume $N=1$. The extension to general N is straightforward. We use the same shorthand as in the proofs of Theorems 11 and 12.

We have the following decomposition:

$$\sqrt{T}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} - \rho^{\pi_e}) = \mathbb{G}_T\psi(\hat{w},\hat{q}) - \mathbb{G}_T\psi(w,q) + \mathbb{G}_T\psi(w,q) + \sqrt{T}(\mathrm{E}[\psi(\hat{w},\hat{q})|\hat{w},\hat{q}] - \rho^{\pi_e}). \tag{EC.35}$$

We again analyze each term.

First term: $\mathbb{G}_T\psi(\hat{w},\hat{q})-\mathbb{G}_T\psi(w,q)=\mathrm{o}_p(1)$. From Theorem 11.24 of Kosorok (2008) based on our assumptions, $\mathbb{G}_T\xrightarrow{d}\mathbf{H}$ in $L^\infty(p_b)$, where $\mathbf{H}$ is a tight mean zero Gaussian process with some covariance. By (13d), we have $\|\psi(\hat{w},\hat{q})-\psi(w,q)\|_2=\mathrm{o}_p(1)$. Then by Lemma 18.5 of van der Vaart (1998), the statement is concluded.

Second term: $\sqrt{T}(\mathrm{E}[\psi(\hat{w},\hat{q})|\hat{w},\hat{q}]-\rho^{\pi_e})=\mathrm{o}_p(1)$. The derivation is done as in the proof of Theorem 11.

Summary Combining, we have

$$\sqrt{T}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}-\rho^{\pi_e})=\mathbb{G}_T\psi(w,q)+\mathrm{o}_p(1).$$

The statement is concluded as in Theorem 10.

Proof of Theorem 14 We provide the proof of cross-fitting version. The adaptive version is similarly proved. The estimator is given by

$$\frac{n_0}{n}\mathbb{P}_{\mathcal{D}_0}\psi^2_{eff}(\hat{w}^{[1]},\hat{q}^{[1]})+\frac{n_1}{n}\mathbb{P}_{\mathcal{D}_1}\psi^2_{eff}(\hat{w}^{[0]},\hat{q}^{[0]})$$

where $\mathbb{P}_{\mathcal{D}_0}$ is a sample average over a set of samples in one fold, and $\mathbb{P}_{\mathcal{D}_1}$ is a sample average over a set of samples in another fold. Then, we have

$$(\mathbb{P}_{\mathcal{D}_0}\psi^2_{eff}(\hat{w}^{[1]},\hat{q}^{[1]})-\mathbb{E}[\psi^2_{eff}(w,q)])=\mathbb{P}_{\mathcal{D}_0}[\psi^2_{eff}(\hat{w}^{[1]},\hat{q}^{[1]})-\psi^2_{eff}(w,q)] \tag{EC.36}$$
$$+\sqrt{n/n_0}\mathbb{G}_{n_0}[\psi^2_{eff}(w,q)] \tag{EC.37}$$
$$+(\mathrm{E}[\psi^2_{eff}(\hat{w}^{[1]},\hat{q}^{[1]})\mid\hat{w}^{[1]},\hat{q}^{[1]}]-\mathbb{E}[\psi^2_{eff}(w,q)]). \tag{EC.38}$$

We analyze each term. Here, we have

$$\begin{aligned}|\psi^2_{eff}(\hat{w},\hat{q})-\psi^2_{eff}(w,q)|&=|\hat{w}^2\{r-\hat{q}(s,a)+\gamma\hat{v}(s')\}^2-w^2\{r-q(s,a)+\gamma v(s')\}^2|\\&\lesssim\max\{|\hat{w}-w|,|\hat{q}-q|,|\hat{v}-v|\}\end{aligned}$$

for some constant C.

Term (EC.36) **is** $\mathcal{O}_p(1/\sqrt{n})$ The conditional mean given $\mathcal{D}_1$ is 0. The conditional variance given $\mathcal{D}_1$ is $\mathcal{O}_p(1/n)$. The rest of the argument is the same as the proof of Theorem 6.

Term (EC.37) **is** $\mathcal{O}_p(1/\sqrt{n})$ This is obvious.

Term (EC.38) **is** $\mathrm{o}_p(1)$

$$|\mathbb{E}[\psi^2_{eff}(\hat{w},\hat{q})-\psi^2_{eff}(w,q)\mid\mathcal{D}_1]|\leq C\max\{\|\hat{w}-w\|_2,\|\hat{q}-q\|_2,\|\hat{v}-v\|_2\}=\mathrm{o}_p(1).$$

Proof of Lemma 1 Define

$$\delta(g,s')=\gamma\int p(s'|s)g(s)\mathrm{d}\lambda(s)-g(s')+(1-\gamma)p^{(0)}_{\pi_e}(s'),$$

where $g(s)$ is any function and $p(s'|s)$ is a marginal distribution of $p(s'|s,a)\pi_e(a|s)$. Then,

$$
\begin{aligned}
&L(w,f_w)\\
&=\mathrm{E}_{p_b}[\{\gamma w(s)\eta(s,a)f_w(s')-w(s)f_w(s)\}]+(1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[f_w(s)]\\
&=\mathrm{E}_{p^{(\infty)}_{e,\gamma}}[(p^{(\infty)}_{e,\gamma}(s)/p_b(s))^{-1}\gamma w(s)\eta(s,a)f_w(s')]-\mathrm{E}_{p^{(\infty)}_{e,\gamma}}[(p^{(\infty)}_{e,\gamma}(s)/p_b(s))^{-1}w(s)f_w(s)]\\
&+(1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[f_w(s)]\\
&=\int \delta(g,s')f_w(s')\mathrm{d}\lambda(s'),
\end{aligned}
$$

where we have $g(s)=p^{(\infty)}_{e,\gamma}(s)\{p^{(\infty)}_{e,\gamma}(s)/p_b(s)\}^{-1}w(s)$.

From the above, when $g(s)=p^{(\infty)}_{e,\gamma}(s)$, i.e., when $w(s)=p^{(\infty)}_{e,\gamma}(s)/p_b(s)$, $L(w,f_w)=0$.

Conversely, $L(w,f_w)=0,\forall f_w\in L_2(S)$ means $\delta(g,s')=0$ from Riesz representation theorem. From the assumption, this means $w(s)=p^{(\infty)}_{e,\gamma}(s)/p_b(s)$.

Proof of Theorem 15

Consistency We have $\hat{\beta}_{f_w}\xrightarrow{p}\beta^*$. We use van der Vaart (1998, Theorem 5.7). We need to check two conditions:

$$
\begin{aligned}
&\sup_{\beta\in\Theta_\beta}|(\mathbb{P}_N-\mathbb{P})[\Delta(s,a,s:\beta)]|\xrightarrow{p}0,\\
&\inf_{\|\beta-\beta_*\|>\epsilon}\|L(w(s:\beta,f_w))\|>0
\end{aligned}
$$

for any $\epsilon>0$. The first condition is proved by Davidson (1994, Section 21.4) noting the first-order derivative of the map $\Theta_\beta\ni\beta\mapsto\Delta(s,a,s:\beta)$ is uniformly bounded. The second condition is proved by noting that

$$L(w(s:\beta),f_w)=0,\beta\in\Theta_\beta\iff\beta=\beta^*,$$

$\Theta_\beta\ni\beta\mapsto L(w(s:\beta)$ is continuous, and Θ_β is a compact space.

Calculation of asymptotic variance We calculate the asymptotic variance for general $f_w(s)$. We prove that the asymptotic MSE is given by

$$\mathrm{E}[\nabla_{\beta^\top}\Delta(s,a,s';\beta)]^{-1}\mathrm{E}[\Delta^2(s,a,s';\beta)]\{\mathrm{E}[\nabla_{\beta^\top}\Delta(s,a,s';\beta)]^{-1}\}^\top|_{\beta^*}. \tag{EC.39}$$

For simplicity, we assume that β is one-dimensional. Using mean value theorem, we have

$$\sqrt{n}(\hat{\beta}_{f_w}-\beta^*)=-\mathbb{P}_n[\nabla_{\beta^\top}\Delta(s,a,s';\beta)]^{-1}|_{\beta^\dagger}\sqrt{n}\mathbb{P}_n[\Delta(s,a,s';\beta)]|_{\beta^*}, \tag{EC.40}$$

where $\beta^\dagger$ is a value between $\hat{\beta}$ and β^*. The first term in right hand side of (EC.40) has the following property:

$$\mathbb{P}_n[\nabla_{\beta^\top}\Delta(s,a,s';\beta)]|_{\beta^\dagger}\xrightarrow{p}\mathrm{E}[\nabla_{\beta^\top}\Delta(s,a,s';\beta)]|_{\beta^*}. \tag{EC.41}$$

This is proved by an uniform convergence condition from the fact the second-order derivative of the map $\Theta_\beta \ni \beta \mapsto \Delta(s,a,s';\beta)$ is uniformly bounded, and $\beta^\dagger \stackrel{p}{\to} \beta^*$.

Next, we calculate the second term in right hand size of (EC.40). By CLT, we have

$$\sqrt{n}\mathbb{P}_n[\Delta(s,a,s';\beta)]|_{\beta^*} \stackrel{d}{\to} \mathcal{N}(0, \mathrm{var}[\Delta(s,a,s';\beta)]).$$

Finally, from Slutsky's theorem the asymptotic variance is

$$\mathrm{E}[\nabla_{\beta^\top}\Delta(s,a,s';\beta)]^{-1}\mathrm{var}[\Delta(s,a,s';\beta)]\{\mathrm{E}[\nabla_{\beta^\top}\Delta(s,a,s';\beta)]^\top\}^{-1}|_{\beta^*}.$$

Proof of Theorem 16

$$\begin{aligned}\sqrt{n}(\hat\rho_{\mathrm{EIS}} - \rho^{\pi_e}) &= \mathbb{G}_n[w(s;\hat\beta_{f_w})\eta(s,a)r] - \mathbb{G}_n[w(s)\eta(s,a)r] \\ &\quad + \mathbb{G}_n[w(s)\eta(s,a)r] + \\ &\quad + \sqrt{n}(\mathrm{E}[w(s;\hat\beta_{f_w})\eta(s,a)r|\hat\beta_{f_w}] - \rho^{\pi_e}).\end{aligned}$$

Here, from the standard argument,

$$\begin{aligned}&\sqrt{n}(\mathrm{E}[w(s;\hat\beta_{f_w})\eta(s,a)r] - \rho^{\pi_e}) \\ &= \sqrt{n}(\mathrm{E}[w(s;\hat\beta_{f_w})\eta(s,a)r] - \mathrm{E}[w(s;\beta^*)\eta(s,a)r]) \\ &= \sqrt{n}\{\mathrm{E}[\nabla w(s;\beta^*)\eta(s,a)r]|_{\beta^*}(\hat\beta_{f_w}-\beta^*) + 0.5(\hat\beta_{f_w}-\beta^*)^\top\mathbb{E}[\nabla_{\beta\beta^\top}\nabla w(s;\beta)\eta(s,a)r]|_{\beta^\dagger}(\hat\beta_{f_w}-\beta^*)\} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(Taylor expansion)} \\ &= \mathrm{E}[\nabla_{\beta^\top}w(s;\beta)\eta(s,a)r]\mathrm{E}[\nabla_{\beta^\top}\Delta(s,a,s';\beta)]^{-1}\mathbb{G}_n[\Delta(s,a,s';\beta)]|_{\beta^*} + \mathrm{o}_p(1).\end{aligned}$$

Combining all together, we have

$$\sqrt{n}(\hat\rho_{\mathrm{EIS}} - \rho^{\pi_e}) = \mathbb{G}_n[w(s;\beta)\eta(s,a)r + \mathrm{E}[\nabla_{\beta^\top}w(s;\beta)\eta(s,a)r]\mathrm{E}[\nabla_{\beta^\top}\Delta(s,a,s';\beta)]^{-1}\Delta(s,a,s';\beta)]|_{\beta^*} + \mathrm{o}_p(1).$$

Proof of Theorems 17 and 18 The proof is almost the same as the proof of Theorem 15. First, the asymptotic variance of $\hat\beta_{f_q}$ is

$$\mathbb{E}[f_q(s,a)\nabla_{\beta^\top}e_q(s,a,r,s';\beta)]^{-1}\mathbb{E}[f_q(s,a)f_q^\top(s,a)e_q^2(s,a,r,s';\beta)]\mathbb{E}[\nabla_\beta e_q(s,a,r,s';\beta)f_q^\top(s,a)]^{-1}|_{\beta^*}$$

This is equal to

$$\mathbb{E}[f_q(s,a)\nabla_{\beta^\top}m_q(s,a:\beta)]^{-1}\mathbb{E}[f_q(s,a)f_q^\top(s,a)v_q(s,a)]\mathbb{E}[\nabla_\beta m_q(s,a:\beta)f_q^\top(s,a)]^{-1}|_{\beta^*}$$

Then, from Tripathi (1999), the lower bound is

$$\mathbb{E}[\nabla_\beta m(s,a;\beta)v_q^{-1}(s,a;\beta)\nabla_{\beta^\top}m(s,a;\beta)]|_{\beta^*}.$$

The asymptotic variance of the direct method is

$$(1-\gamma)^2\mathbb{E}_{d_0}[\nabla_{\beta^\top}q(s,\pi_e)]\mathbb{E}\left[\frac{\otimes\{\nabla_\beta\mathbb{E}[\gamma q(s',\pi;\beta) - q(s,a;\beta)|s,a]\}}{\mathrm{var}[r+\gamma q(s',\pi_e)|s,a]}\right]^{-1}\mathbb{E}_{d_0}[\nabla_\beta q(s,\pi_e)].$$

Proof of Lemma 2 Recall that the asymptotic variance of the direct method is

$$(1-\gamma)^2\mathbb{E}_{d_0}[\nabla_{\beta^\top} q(s,\pi_e)]\mathbb{E}\left[\frac{\otimes\{\nabla_\beta \mathbb{E}[\gamma q(s',\pi;\beta)-q(s,a;\beta)|s,a]\}}{\mathrm{var}[r+\gamma q(s',\pi_e)|s,a]}\right]^{-1}\mathbb{E}_{d_0}[\nabla_\beta q(s,\pi_e)]. \tag{EC.42}$$

In addition, we have

$$(1-\gamma)\mathbb{E}_{d_0}[q(s_0,\pi_e)] = \mathbb{E}[-\gamma w(s,a)q(s',\pi_e)+w(s,a)q(s,a)].$$

By differentiating this equation,

$$(1-\gamma)\nabla_\beta \mathbb{E}_{d_0}[q(s,\pi_e)] = \mathbb{E}[w(s,a)\nabla_\beta \mathbb{E}[-\gamma q(s',\pi_e)+q(s,a)|s,a]].$$

According to CS-inequality, this immediately means that (EC.42) is smaller than the efficient bound under the nonparametric model:

$$\mathbb{E}[w(s,a)^2\mathrm{var}[r+\gamma q(s',\pi_e)|s,a]].$$

Appendix D: Additional Experimental Results

Here, we provide additional results from the experiment in Section 8 with $\alpha = 0.4, 0.8$. The results are given in Figs. EC.1 to EC.6.

References

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1998) *Efficient and Adaptive Estimation for Semiparametric Models* (Springer).

Bradley RC (2005) Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys* 2:107–144.

Davidson JEH (1994) *Stochastic limit theory :an introduction for econometricians.* Advanced texts in econometrics (Oxford).

Ibragimov IA, Linnik YV (1971) *Independent and stationary sequences of random variables* (The Netherlands: Walters-Noordhof).

Kallus N, Uehara M (2020) Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research* 21(167):1–63.

Kosorok MR (2008) *Introduction to Empirical Processes and Semiparametric Inference.* Springer Series in Statistics (New York, NY: Springer New York).

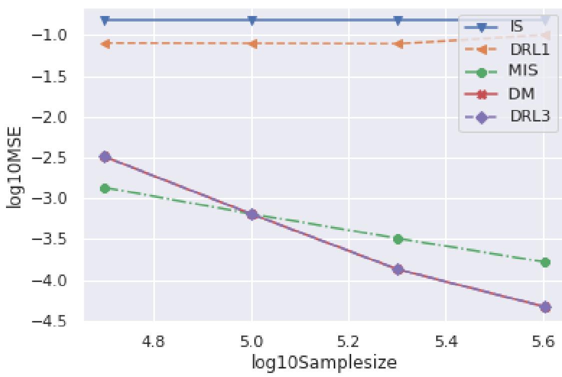

Figure EC.1 Setting (1) with $\alpha = 0.4$

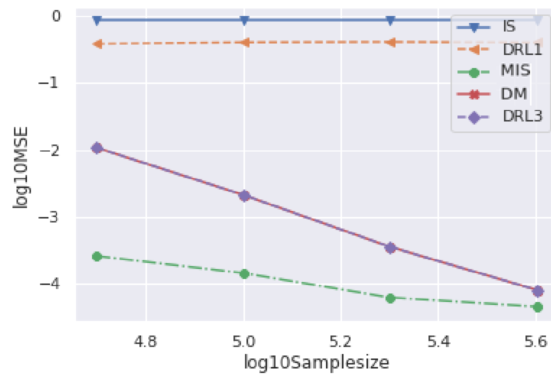

Figure EC.2 Setting (1) with $\alpha = 0.8$

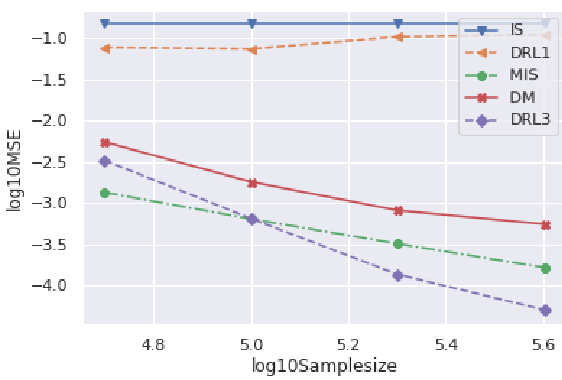

Figure EC.3 Setting (2) with $\alpha = 0.4$

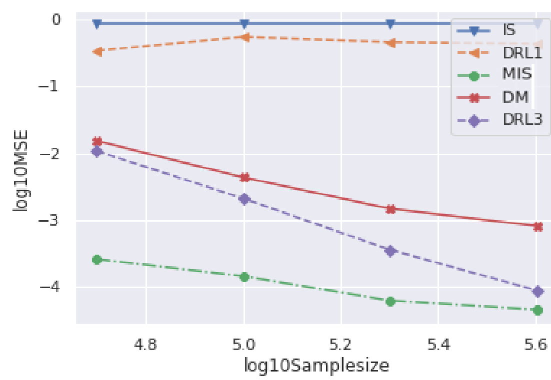

Figure EC.4 Setting (2) with $\alpha = 0.8$

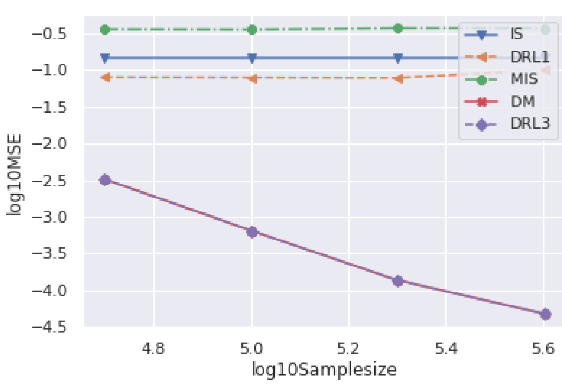

Figure EC.5 Setting (3) with $\alpha = 0.4$

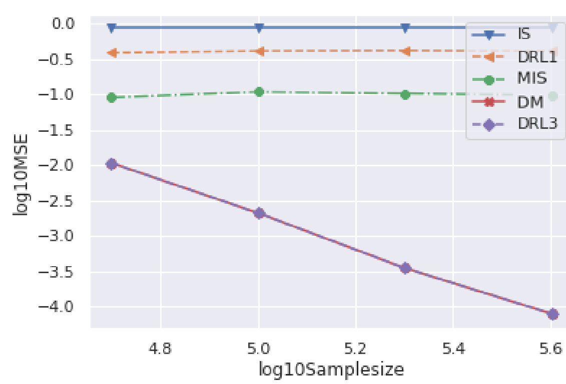

Figure EC.6 Setting (3) with $\alpha = 0.8$

Tripathi G (1999) A matrix extension of the Cauchy-Schwarz inequality. *Economics Letters* 63:1–3.

Uehara M, Imaizumi M, Jiang N, Kallus N, Sun W, Xie T (2021) Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*

.

van der Vaart AW (1991) On differentiable functionals. *Ann. Statist.* 19:178–204.

van der Vaart AW (1998) *Asymptotic statistics* (Cambridge, UK: Cambridge University Press).