A. Notation

We first summarize the notation we use in Table 4. Notice in particular that, following empirical process theory literature, in the proofs we also use $\mathbb{P}$ to denote expectations (interchangeably with E).

In addition, We let $\mathcal{M}_{\mathrm{NMDP}}$ denote the nonparametric model where each distribution is unknown and free. We let $\mathcal{M}_{\mathrm{NMDP}-b}$ denote the submodel of $\mathcal{M}_{\mathrm{NMDP}}$ where π^b is known and fixed. We let $\mathcal{M}_{\mathrm{MDP}}$, $\mathcal{M}_{\mathrm{MDP}-b}$ denote the corresponding models where both the decision process and the behavior policy are restricted to be Markovian.

Table 4. Notation

∇_β	Differentiation with respect to β
$r_t, s_t, a_t,$	Reward, state, action at t
$\mathcal{J}_{r_t}, \mathcal{J}_{s_t}, \mathcal{J}_{a_t}$	History up to time r_t, s_t, a_t, including reward variables
$\mathcal{H}_{s_t}, \mathcal{H}_{a_t}$	History up to time s_t, a_t, excluding reward variables
$\pi_t(a_t\|\mathcal{H}_{s_t}), \pi_t(a_t\|s_t)$	Policy in NMDP and MDP case, respectively
π_t^e, π_t^b	Target and behavior policies at t, respectively
ρ^π	Policy value, $\mathrm{E}_\pi[\sum_{t=0}^T r_t]$
$v_t = v_t(\mathcal{H}_{s_t}), v_t(s_t)$	Value function at t, in $\mathrm{NMDP}, \mathrm{MDP}$ respectively
$q_t = q_t(\mathcal{H}_{a_t}), q_t(s_t, a_t)$	q-function at t, in $\mathrm{NMDP}, \mathrm{MDP}$ respectively
ν_t	Cumulative density ratio $\prod_{k=0}^t \pi_t^e/\pi_t^b$
μ_t	Marginal density ratio $\mathrm{E}[\nu_t \mid s_t, a_t]$
η_t	Instantaneous density ratio π_t^e/π_t^b
Λ	Tangent space
$\mathcal{M}$	A model for the data generating distribution
$\mathcal{M}_{\mathrm{NMDP}}, \mathcal{M}_{\mathrm{NMDP}-b}$	NMDP model with unknown behavior policy, known behavior policy, and parametric q-function, respectively
$\mathcal{M}_{\mathrm{MDP}}, \mathcal{M}_{\mathrm{MDP}-b},$	MDP model with unknown behavior policy, known behavior policy, and parametric q-function, respectively
$C, R_{\max}$	Upper bound of density ratio and reward, respectively
$\prod(A\|B)$	Projection of A onto B
$\bigoplus$	Direct sum
$\|\cdot\|_p$	L^p-norm $\mathrm{E}[f^p]^{1/p}$
$\lessapprox$	Inequality up to constant
$\mathrm{E}_\pi[\cdot], \mathbb{P}_\pi$	Expectation with respect to a sample from a policy π
$\mathrm{E}[\cdot], \mathbb{P}$	Same as above for $\pi = \pi^b$
$\mathrm{E}_n[\cdot], \mathbb{P}_n$	Empirical expectation (based on sample from a behavior policy)
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 $\sqrt{n}(\mathbb{P}_n - \mathbb{P})$
$\mathrm{Asmse}[\cdot]$, $\mathrm{var}[\cdot]$	Asymptotic variance, variance
$\mathcal{N}(a, b)$	Normal distribution with mean a and variance b
$\mathrm{Uni}[a, b]$	Uniform distribution on $[a, b]$
$A_n = o_p(a_n)$	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
Λ_d^α	Hölder space with smoothness α with a dimension d
NMDP	$p_0(s_0)\prod_{t=0}^T \pi_t(a_t \mid \mathcal{H}_{s_t}) p_t(r_t \mid \mathcal{H}_{a_t}) p_t(s_{t+1} \mid \mathcal{H}_{a_t})$

B. Semiparametric Theory

We denote the all of the history $\{\mathcal{H}^{(i)}\}_{i=1}^n$ as $\mathcal{H}^n$, the estimand as $R(F) : \mathcal{M} \to \mathbb{R}$ and the estimator as $\hat{R} : \mathcal{H}^n \to \mathbb{R}$. First, we introduce some definitions.

Definition 1 (One-dimensional submodel and its score function)**.** A one-dimensional submodel of $\mathcal{M}$ that passes through F at 0 is a subset of $\mathcal{M}$ of the form $\{F_\epsilon : \epsilon \in [-a, a]\}$ for some small $a > 0$ s.t. $F_{\epsilon=0} = F$. The score of the submodel F_ϵ at $\theta = 0$ is defined as

$$s(\mathcal{H}) = \frac{\log(\mathrm{d}F_\epsilon/\mathrm{d}\mu)(\mathcal{H})}{\mathrm{d}\epsilon} \mid_{\epsilon=0} .$$

Definition 2 (Tangent space)**.** The tangent space of a model $\mathcal{M}$ at F denoted by $T_{\mathcal{M}}(F)$ is the linear closure of the set of score functions of the all one-dimensional submodels regarding $\mathcal{M}$ that pass through F.

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

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

Definition 4 (Pathwise differentiability)**.** A functional $R(F)$ is pathwise differentiable at F w.r.t the model $\mathcal{M}$ (or w.r.t the tangent space $\mathcal{T}_{\mathcal{M}}(F)$) if there exists a function $D_F(\mathcal{H})$ such that for all submodels $\{F_\epsilon : \epsilon\}$ in $\mathcal{M}$ satisfying $F_{\epsilon=0} = F$ and

$$\frac{dR(F_\epsilon)}{d\epsilon} \mid_{\epsilon=0} = \mathbb{E}[D_F(\mathcal{H})s(\mathcal{H})],$$

where $s(\mathcal{H})$ is a corresponding score function for F_ϵ. The function $D_F(\mathcal{H})$ is called a gradient of $R(F)$ at F w.r.t the model $\mathcal{M}$. The efficient IF (EIF) of $R(F)$ w.r.t the model $\mathcal{M}$ is called a canonical gradient $\tilde{D}_F(\mathcal{H})$, which is the unique gradient of $R(F)$ at F w.r.t the model $\mathcal{M}$ that belongs to the tangent space $\mathcal{T}_{\mathcal{M}}(F)$.

Next, we define regular estimators. Regular estimators means estimators whose limiting distribution is insensitive to local changes to the data generating process. It excludes a well-known Hodge estimator. Here, we denote a submodel with some score function g in a given tangent space $T_{\mathcal{M}}(F)$ as $\{F_{t,g}; t \in [-a, a]\}$.

Definition 5 (Regular estimators)**.** An estimator sequence T_n is called regular at F for $R(F)$ w.r.t the model $\mathcal{M}$ (or w.r.t the tangent space $T_{\mathcal{M}}(F)$), if there exists a probability measure L such that

$$\sqrt{n}\{T_n - R(F_{1/\sqrt{n},g})\} \overset{d(F_{1/\sqrt{n},g})}{\to} L, \text{for every } g \in T_{\mathcal{M}}(F).$$

The following three theorems imply that influence functions of the estimators $\hat{R}(F)$ for $R(F)$ and gradients of $R(F)$ correspond to each other, and how to construct an efficient estimator. These theorems are based on Theorem 3.1 (van der Vaart, 1991).

Theorem 13 (Influence functions are gradients)**.** *Under certain regularity conditions, for $P \in \mathcal{M}$, suppose $\hat{R}(\mathcal{H}^n)$ is a regular estimator of $R(F)$ w.r.t the model $\mathcal{M}$, and that it is asymptotically linear with influence function $D_F(\mathcal{H})$. Then, $R(F)$ is pathwise differentiable at F w.r.t $\mathcal{M}$ and $D_F(\mathcal{H})$ is a gradient of $R(F)$ at F w.r.t $\mathcal{M}$.*

Theorem 14 (Gradients are influence functions)**.** *Under certain regularity conditions, if a $D_F(\mathcal{H})$ is a gradient of $R(F)$ at F w.r.t the model $\mathcal{M}$, there exists an asymptotically linear estimator of $R(F)$ with influence function $D_F(\mathcal{H})$, which is regular w.r.t the model $\mathcal{M}$.*

Corollary 15 (Characterization of efficient influence functions)**.** *The efficient influence function is the projection of any gradient onto the tangent space $\mathcal{T}_{\mathcal{M}}(F)$.*

Note that gradients w.r.t the model $\mathcal{M}$ are not unique if the model $\mathcal{M}$ is not a fully nonparametric model. If the underlying model is fully nonparametric model, the gradient is unique.

Strategy to calculate the EIF With the abovementioned definitions and theorems in mind, our general strategy to compute efficient influence functions is as follows.

1. Calculate some gradient $D_F(\mathcal{H})$ (a candidate of EIF) of the target functional $R(F)$ w.r.t $\mathcal{M}$
2. Calculate the the tangent space $\mathcal{T}_{\mathcal{M}}(F)$ at F
3. Show that some candidate of EIF is orthogonal to the orthogonal tangent space, i.e., the candidate of EIF lies in the tangent space. Then, this implies that a candidate of EIF is actually the EIF.

The other common strategy is calculating some gradient and projection it onto $\mathcal{T}_{\mathcal{M}}(F)$.

Optimalites The efficiency bound 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 16 (Local Asymptotic Minimax theorem)**.** *Let $R(F)$ be pathwise diffentiable at F w.r.t the model $\mathcal{M}$ with EIF $\tilde{D}_F(\mathcal{H})$. If $T_{\mathcal{M}}(F)$ is a convex cone, for any estimator sequence $\hat{R}(\mathcal{H}^n)$, and subconvex loss function $l : \mathbb{R} \to [0, \infty)$,*

$$\sup_I \lim_{n\to\infty} \sup_{g\in I} \mathrm{E}_{F_{1/\sqrt{n},g}}[l[\sqrt{n}\{\hat{R}(\mathcal{H}^n) - R(F_{1/\sqrt{n},g})\}]] \geq \int l(u)\mathrm{d}N(0, \mathrm{var}_F[\tilde{D}_F(\mathcal{H})])(u),$$

where the first supremum is taken over all finite subsets I of the tangent set.

Corollary 17. *Under the same assumptions of Theorem 16,*

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

where $\|\cdot\|_T$ is a total variation distance.

Other different type of optimality is seen in the following theorem. The following theorem state that an asymptotic variance of every regular estimator sequence $\hat{R}(\mathcal{H}^n)$ with limiting distribution L is bounded below $\mathrm{E}[\tilde{D}_F^2(\mathcal{H})]$ (van der Vaart, 1998, Thm. 25.21).

Theorem 18 (Convolution theorem)**.** *Let $R(F)$ be pathwise differentiable at F w.r.t the model $\mathcal{M}$ with EIF $\tilde{D}_F(\mathcal{H})$. Let $\hat{R}(\mathcal{H}^n)$ be a regular estimator sequence at F w.r.t the tangent space $\mathcal{T}_{\mathcal{M}}(F)$ with limiting distribution L. Then, if the tangent space $T_{\mathcal{M}}(F)$ is a cone, then, the term*

$$\int u^2 \mathrm{d}L(u) - \mathrm{E}[\tilde{D}_F^2(\mathcal{H})]$$

is non-negative.

C. Additional Details from Section 5

Cliff Walking. This RL task is detailed in Example 6.6 in (Sutton, 2018). We consider a board of size 4×12. The horizon was set to $T = 400$. Each time step incurs -1 reward until the goal is reached, at which point it is 0, and stepping off the cliff incurs -100 reward and a reset to the start.

Mountain Car. The RL task is as follows: a car is between two hills in the interval $[-0.7, 0.5]$ and the agent must move back and forth to gain enough power to reach the top of the right hill. The state space comprises position and velocity. There are three discrete actions: (1) forward, (2) backward, and (3) stay-still. The horizon was set to $T = 200$. The reward for each step is -1 until the position 0.5 is reached, at which point it is 0. The state space was continuous; thus, we obtained a 400-dimensional feature expansion using a radial basis function kernel as mentioned.

The Policy π_d. We construct the policy π_d using standard q-learning (Sutton, 2018). For Cliff Walking, we use a q-learning in a tabular manner. Regarding a Mountain Car, we use q-learning based on the same feature expansion as above. We use 4000 sample to learn an optimal policy.

D. Proofs

Proof of Theorem 2.

Efficient influence function under $\mathcal{M}_{\text{NMDP}}$. The entire regular (regular model as defined in Chapter 7 van der Vaart, 1998) parametric submodel under NMDP is

$$\{p_\theta(s_0)p_\theta(a_0|s_0)p_\theta(r_0|\mathcal{H}_{a_0})p_\theta(s_1|\mathcal{H}_{r_0})p_\theta(a_1|\mathcal{H}_{s_1})p_\theta(r_1|\mathcal{H}_{a_1})\cdots p_\theta(r_T|\mathcal{H}_{a_T})\},$$

where it matches with a true pdf at $\theta = 0$.

The score function of the model $\mathcal{M}_{\text{NMDP}}$ is decomposed as

$$g(\mathcal{J}_{s_T}) = \textstyle\sum_{k=0}^{T} g_{s_k|\mathcal{H}_{a_{k-1}}} + \sum_{k=0}^{T} g_{a_k|\mathcal{H}_{s_k}} + \sum_{k=0}^{T} g_{r_k|\mathcal{H}_{a_k}}.$$

We first calculate an influence function for the target functional. Note that this influence function is not unique. We have

$$\begin{aligned}
&\nabla_\theta \mathrm{E}_{\pi^e}\left[\sum_{t=0}^{T} r_t\right] \\
&= \nabla_\theta \left[\int \sum_{t=0}^{T} r_t \left\{\prod_{k=0}^{T} p_\theta(s_k|\mathcal{H}_{r_{k-1}}) p_{\pi^e}(a_k|\mathcal{H}_{s_k}) p_\theta(r_k|\mathcal{H}_{a_k})\right\} \mathrm{d}\mu(\mathcal{J}_{s_T})\right] \\
&= \sum_{c=0}^{T} \{\mathrm{E}_{\pi^e}[\{\mathrm{E}_{\pi^e}(r_c|s_0) - \mathrm{E}_{\pi^e}(r_c)\} g_{s_0}] - \mathrm{E}_{\pi^e}[\{r_c - \mathrm{E}_{\pi^e}(r_c|\mathcal{H}_{a_c})\} g_{r_c|\mathcal{H}_{a_c}}] \\
&- \mathrm{E}_{\pi^e}\left[\left(\mathrm{E}_{\pi^e}\left[\sum_{t=c+1}^{T} r_t|\mathcal{H}_{s_{c+1}}\right] - \mathrm{E}_{\pi^e}\left[\sum_{t=c+1}^{T} r_t|\mathcal{H}_{a_c}\right]\right) g_{s_{c+1}|\mathcal{H}_{a_c}}\right]\} \\
&= \sum_{c=0}^{T} \{\mathrm{E}_{\pi^e}[\{\mathrm{E}_{\pi^e}(r_c|s_0) - \mathrm{E}_{\pi^e}(r_c)\} g(\mathcal{H}_{s_{T+1}})] - \mathrm{E}_{\pi^e}[\{r_c - \mathrm{E}_{\pi^e}(r_c|\mathcal{H}_{a_c})\} g(\mathcal{J}_{s_T})] \\
&- \mathrm{E}_{\pi^e}\left[\left(\mathrm{E}_{\pi^e}\left[\sum_{t=c+1}^{T} r_t|\mathcal{H}_{s_c+1}\right] - \mathrm{E}_{\pi^e}\left[\sum_{t=c+1}^{T} r_t|\mathcal{H}_{a_c}\right]\right) g(\mathcal{J}_{s_T})\right]\} \\
&= \mathrm{E}\left(\left[-\rho^{\pi^e} + \sum_{c=0}^{T} \nu_c r_c - \left\{\nu_c \sum_{t=c}^{T} \mathrm{E}_{\pi_e}(r_t|\mathcal{H}_{a_c}) - \nu_{c-1} \sum_{t=c}^{T} \mathrm{E}_{\pi_e}(r_t|\mathcal{H}_{s_c})\right\}\right] g(\mathcal{J}_{s_T})\right).
\end{aligned}$$

This concludes that the following function is an influence function:

$$\rho_{\text{eff}}^{\text{NMDP}} = -\rho^{\pi^e} + \sum_{c=0}^{T} \nu_c r_c - \left\{\nu_c \sum_{t=c}^{T} \mathrm{E}_{\pi_e}(r_t|\mathcal{H}_{a_c}) - \nu_{c-1} \sum_{t=c}^{T} \mathrm{E}_{\pi_e}(r_t|\mathcal{H}_{s_c})\right\}. \tag{11}$$

Next, we show that this influence function is the efficient influence function. In order to show this, we calculate the tangent space of model $\mathcal{M}_{\text{NMDP}}$. The (nuisance) tangent space of the model $\mathcal{M}_{\text{NMDP}}$ is the product space:

$$\begin{aligned}
&\bigoplus_{0\le t\le T} (A_t \bigoplus B_t \bigoplus C_t), \\
&A_t = \{q(s_t, \mathcal{H}_{a_{t-1}}); \mathrm{E}[q(s_t, \mathcal{H}_{a_{t-1}})|\mathcal{H}_{a_{t-1}}] = 0,\ q \in L^2\}, \\
&B_t = \{q(a_t, \mathcal{H}_{s_t}); \mathrm{E}[q(a_t, \mathcal{H}_{s_t})|\mathcal{H}_{s_t}] = 0,\ q \in L^2\}, \\
&C_t = \{q(r_t, \mathcal{H}_{a_t}); \mathrm{E}[q(r_t, \mathcal{H}_{a_t})|\mathcal{H}_{a_t}] = 0,\ q \in L^2\}.
\end{aligned}$$

The orthogonal space of the tangent space is the product of

$$\bigoplus_{0\leq t\leq T}(A'_t \bigoplus B'_t \bigoplus C'_t) \tag{12}$$

such that

$$\begin{aligned}
A'_t \bigoplus A_t &= A''_t,\ A''_t = \{q(\mathcal{J}_{s_t}); \mathrm{E}[q(\mathcal{J}_{s_t})|\mathcal{J}_{r_{t-1}}] = 0,\ q \in L^2\},\\
B'_t \bigoplus B_t &= B''_t,\ B''_t = \{q(\mathcal{J}_{a_t}); \mathrm{E}[q(\mathcal{J}_{a_t})|\mathcal{J}_{s_t}] = 0,\ q \in L^2\},\\
C'_t \bigoplus C_t &= C''_t,\ C''_t = \{q(\mathcal{J}_{r_t}); \mathrm{E}[q(\mathcal{J}_{r_t})|\mathcal{J}_{a_t}] = 0,\ q \in L^2\}.
\end{aligned}$$

More specifically, we have the following lemma.

Lemma 19. *The orthogonal tangent space is represented as*

$$\begin{aligned}
A'_t &= \left\{q(\mathcal{J}_{s_t}) - \mathrm{E}[q(\mathcal{J}_{s_t})|\mathcal{H}_{s_t}]; \mathrm{E}[q(\mathcal{J}_{s_t})|\mathcal{J}_{r_{t-1}}] = 0,\ q \in L^2\right\},\\
B'_t &= \left\{q(\mathcal{J}_{a_t}) - \mathrm{E}[q(\mathcal{J}_{a_t})|\mathcal{H}_{a_t}]; \mathrm{E}[q(\mathcal{J}_{a_t})|\mathcal{J}_{s_t}] = 0,\ q \in L^2\right\},\\
C'_t &= \left\{q(\mathcal{J}_{r_t}) - \mathrm{E}[q(\mathcal{J}_{r_t})|\mathcal{H}_{a_t}, r_t]; \mathrm{E}[q(\mathcal{J}_{r_t})|\mathcal{J}_{a_t}] = 0,\ q \in L^2\right\}.
\end{aligned}$$

Proof. We give a proof for A'_t. Regarding the other cases, it is proved similarly. First, from the definition of the conditional expectation, A'_t and A_t are orthogonal. Thus, what we have to prove is $\mathrm{E}[q(\mathcal{J}_{s_t})|\mathcal{H}_{s_t}]$ is included in A_t. This is proved as follows:

$$\mathrm{E}[\mathrm{E}[q(\mathcal{J}_{s_t})|\mathcal{H}_{s_t}]|\mathcal{H}_{a_{t-1}}] = \mathrm{E}[q(\mathcal{J}_{s_t})|\mathcal{H}_{a_{t-1}}] = \mathrm{E}[\mathrm{E}[q(\mathcal{J}_{s_t})|\mathcal{J}_{r_{t-1}}]|\mathcal{H}_{a_{t-1}}] = 0. \qquad \square$$

If we can prove that the influence function Eq. (11) is orthogonal to the orthogonal tangent space Eq. (12), we can see that the above influence function is actually the efficient influence function. This fact is shown as follows.

Lemma 20. *The derivative Eq.* (11) *is orthogonal to* $\{A'_t\}_{t=0}^{T+1}$, $\{B''_t\}_{t=0}^{T}$, $\{C'_t\}_{t=0}^{T}$

Proof. The influence function is orthogonal to A'_k: for $t(\mathcal{J}_{s_k}) \in A'_k$

$$\begin{aligned}
&\mathrm{E}\left[\left\{-\rho^{\pi^e} + \sum_{c=0}^{T}\nu_c(\mathcal{H}_{a_c})r_c - \left\{\nu_c(\mathcal{H}_{a_c})\sum_{t=c}^{T}\mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{a_c}] - \nu_{c-1}(\mathcal{H}_{a_{c-1}})\sum_{t=c}^{T}\mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{s_c}]\right\}\right\} t(\mathcal{J}_{s_k})\right]\\
&= \mathrm{E}\left[\left\{\sum_{c=k}^{T}\nu_c(\mathcal{H}_{a_c})r_c - \nu_{k-1}\sum_{t=k}^{T}\mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{s_k}]\right\} t(\mathcal{J}_{s_k})\right]\\
&= 0.
\end{aligned}$$

The influence function is orthogonal to B''_k: for $t(\mathcal{J}_{a_k}) \in B''_k$;

$$\begin{aligned}
&\mathrm{E}\left[\left\{-\rho^{\pi^e} + \sum_{c=0}^{T}\nu_c r_c - \left\{\nu_c\sum_{t=c}^{T}\mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{a_c}] - \nu_{c-1}\sum_{t=c}^{T}\mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{s_c}]\right\}\right\} t(\mathcal{J}_{a_k})\right]\\
&= \mathrm{E}\left[\left\{\sum_{c=k}^{T}\nu_c r_c - \left\{\nu_c\sum_{t=c}^{T}\mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{a_c}] - \nu_{c-1}\sum_{t=c}^{T}\mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{s_c}]\right\}\right\} t(\mathcal{J}_{a_k})\right]\\
&= \mathrm{E}\left[\left\{\left\{\sum_{c=k}^{T}\nu_c r_c\right\} - \left\{\nu_k\sum_{t=k}^{T}\mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{a_k}]\right\}\right\} t(\mathcal{J}_{a_k})\right]\\
&= \mathrm{E}\left[\left\{\left\{\nu_k\sum_{t=k}^{T}\mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{a_k}]\right\} - \left\{\nu_k\sum_{t=k}^{T}\mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{a_k}]\right\}\right\} t(\mathcal{J}_{a_k})\right] = 0.
\end{aligned}$$

The influence function is orthogonal to C'_k: for $t(\mathcal{J}_{r_k}) \in C'_k$;

$$
\begin{aligned}
&\mathrm{E}\left[\left\{-\rho^{\pi^e} + \sum_{c=0}^{T} \nu_c(\mathcal{H}_{a_c}) r_c - \left\{\nu_c(\mathcal{H}_{a_c}) \sum_{t=c}^{T} \mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{a_c}] - \nu_{c-1}(\mathcal{H}_{a_{c-1}}) \sum_{t=c}^{T} \mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{s_c}]\right\}\right\} t(\mathcal{J}_{r_k})\right] \\
&= \mathrm{E}\left[\left\{\sum_{c=k}^{T} \nu_c(\mathcal{H}_{a_c}) r_c\right\} t(\mathcal{J}_{r_k})\right] \\
&= \mathrm{E}\left[\left\{\left\{\nu_{k-1} \sum_{t=k}^{T} \mathrm{E}_{\pi^e}[r_t|\mathcal{J}_{r_k}]\right\}\right\} t(\mathcal{J}_{r_k})\right] = \mathrm{E}\left[\left\{\nu_{k-1} \sum_{t=k}^{T} \mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{a_k}, r_k]\right\} t(\mathcal{J}_{r_k})\right] \\
&= \mathrm{E}\left[\left\{\left\{\nu_{k-1} \sum_{t=k}^{T} \mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{a_k}, r_k]\right\}\right\} \mathrm{E}[t(\mathcal{J}_{r_k})|\mathcal{H}_{a_k}, r_k]\right] = 0.
\end{aligned}
$$

□

This concludes the proof for NMDP.

Efficient influence function under $\mathcal{M}_{\mathrm{NMDP}-b}$. Next, we show that the efficiency bound is still the same even if we know the target policy. To show that, we derive an orthogonal space of the tangent space of the regular parametric submodel:

$$
\{p_\theta(s_0)p(a_0|s_0)p_\theta(r_0|\mathcal{H}_{a_0})p_\theta(s_1|\mathcal{H}_{r_0})p(a_1|\mathcal{H}_{s_1})p_\theta(r_1|\mathcal{H}_{a_1})\cdots p_\theta(r_T|\mathcal{H}_{a_T})\},
$$

where $p(a_t|\mathcal{H}_{s_t})$ is fixed at π_t^b. This is equal to

$$
\bigoplus_{0\leq t\leq T}(A'_t \bigoplus B''_t \bigoplus C'_t) \tag{13}
$$

This space Eq. (13) is orthogonal to the obtained efficient influence function under NMDP. Therefore, the efficient influence function under $\mathcal{M}_{\mathrm{NMDP}-b}$ is the same as the one under $\mathcal{M}_{\mathrm{NMDP}}$.

Efficiency bound. We use a law of total variance (Bowsher & Swain, 2012) to compute the variance of the efficient influence function.

$$
\begin{aligned}
&\mathrm{var}\left[\sum_{t=0}^{T}(\nu_t r_t - (\nu_t q_t - \nu_{t-1} v_t))\right] \\
&= \sum_{t=0}^{T+1} \mathrm{E}\left[\mathrm{var}\left(\mathrm{E}\left[\nu_{t-1}r_{t-1} + \sum_{k=0}^{T}(\nu_k r_k - \{\nu_k q_k - \nu_{k-1} v_k\})\,|\mathcal{J}_{a_t}\right]|\mathcal{J}_{a_{t-1}}\right)\right] \\
&= \sum_{t=0}^{T+1} \mathrm{E}\left[\mathrm{var}\left(\mathrm{E}\left[\nu_{t-1}r_{t-1} + \sum_{k=t}^{T}(\nu_k r_k - \{\nu_k q_k - \nu_{k-1} v_k\})\,|\mathcal{J}_{a_t}\right]|\mathcal{J}_{a_{t-1}}\right)\right] \\
&= \sum_{t=0}^{T+1} \mathrm{E}\left[\mathrm{var}\left(\mathrm{E}\left[\nu_{t-1}r_{t-1} + \left(\sum_{k=t}^{T}\nu_k r_k\right) - \{\nu_t q_t - \nu_{t-1} v_t\}|\mathcal{J}_{a_t}\right]|\mathcal{J}_{a_{t-1}}\right]\right) \\
&= \sum_{t=0}^{T+1} \mathrm{E}\left[\nu_{t-1}^2 \mathrm{var}\left(r_{t-1} + v_t(\mathcal{H}_{s_t})|\mathcal{H}_{a_{t-1}}\right)\right].
\end{aligned}
$$

Here, we used $\mathrm{E}[\sum_{k=t}^{T}\nu_k r_k|\mathcal{J}_{a_k}] = \nu_k q_k$. □

Proof of Theorem 3.

Efficient influence function under $\mathcal{M}_{\mathrm{MDP}}$. The entire regular parametric submodel is

$$
\{p_\theta(s_0)p_\theta(a_0|s_0)p_\theta(r_0|s_0,a_0)p_\theta(s_1|s_0,a_0)p_\theta(a_1|s_1)p_\theta(r_1|s_1,a_1)\cdots p_\theta(r_T|s_T,a_T)\}.
$$

The score function of the parametric submodel is

$$g(\mathcal{J}_{s_T}) = \sum_{k=0}^{T} g_{s_k|s_{k-1},a_{k-1}} + \sum_{k=0}^{T} g_{a_{k+1}|s_k} + \sum_{k=0}^{T} g_{r_k|s_k,a_k}.$$

We first calculate the influence function of the target function. Note that this influence function is not only influence function. We have

$$\begin{aligned}
&\nabla_\theta \mathrm{E}_{\pi^e}[\sum_{t=0}^{T} r_t] \\
&= \nabla_\theta \int \sum_{t=0}^{T} r_t \left\{ \prod_{t=0}^{T} p_\theta(s_k|a_{k-1}, s_{k-1}) p_{\pi_k^e}(a_k|s_k) p_\theta(r_k|a_k, s_k) \right\} \mathrm{d}\mu(\mathcal{J}_{s_T}) \\
&= \sum_{c=0}^{T} \{ \mathrm{E}_{\pi^e}\left[(\mathrm{E}_{\pi^e}[r_c|s_0] - \mathrm{E}_{\pi^e}[r_c]) g_{s_0} \right] - \mathrm{E}_{\pi^e}[(r_c - \mathrm{E}_{\pi^e}[r_c|s_c, a_c]) g_{r_c|s_c,a_c}] \\
&- \mathrm{E}_{\pi^e}\left[\left(\mathrm{E}_{\pi^e}[\sum_{c=t+1}^{T} r_t|s_{c+1}] - \mathrm{E}_{\pi^e}[\sum_{c=t+1}^{T} r_t|s_c, a_c] \right) g_{s_{c+1}|s_c,a_c} \right] \} \\
&= \sum_{c=0}^{T} \{ \mathrm{E}[(\mathrm{E}[r_c|s_0] - \mathrm{E}_{\pi^e}[r_c]) g] - \mathrm{E}\left[\frac{p_{\pi^e}(s_c, a_c)}{p_{\pi^b}(s_c, a_c)} (r_c - \mathrm{E}[r_c|s_c, a_c]) g \right] \\
&- \mathrm{E}\left[\frac{p_{\pi^e}(s_c, a_c)}{p_{\pi^b}(s_c, a_c)} (\mathrm{E}[\sum_{t=c+1}^{T} r_t|s_{c+1}] - \mathrm{E}[\sum_{t=c+1}^{T} r_t|s_c, a_c]) g \right] \} \\
&= \mathrm{E}\left[\left[-\rho^{\pi^e} + \sum_{c=0}^{T} \frac{p_{\pi^e}(s_c, a_c)}{p_{\pi^b}(s_c, a_c)} r_c - \left\{ \frac{p_{\pi^e}(s_c, a_c)}{p_{\pi^b}(s_c, a_c)} \sum_{t=c}^{T} \mathrm{E}_{\pi_e}[r_t|s_c, a_c] - \frac{p_{\pi^e}(s_{c-1}, a_{c-1})}{p_{\pi^b}(s_{c-1}, a_{c-1})} \sum_{t=c}^{T} \mathrm{E}_{\pi_e}[r_t|s_c] \right\} \right] g(\mathcal{J}_{s_T}) \right]
\end{aligned}$$

Therefore, the following function is an influence function;

$$-\rho^{\pi_c^e} + \sum_{c=0}^{T} \frac{p_{\pi_c^e}(s_c, a_c)}{p_{\pi_c^b}(s_c, a_c)} r_c - \left\{ \frac{p_{\pi_c^e}(s_c, a_c)}{p_{\pi_c^b}(s_c, a_c)} \sum_{t=c}^{T} \mathrm{E}_{\pi_e}[r_t|s_c, a_c] - \frac{p_{\pi_c^e}(s_{c-1}, a_{c-1})}{p_{\pi_c^b}(s_{c-1}, a_{c-1})} \sum_{t=c}^{T} \mathrm{E}_{\pi_e}[r_t|s_c] \right\}. \tag{14}$$

We will show this influence function is the efficient influence function.

In order to show this, we calculate the tangent space of model $\mathcal{M}_{\mathrm{MDP}}$. The tangent space of the model $\mathcal{M}_{\mathrm{MDP}}$ is the product space;

$$\begin{aligned}
&\bigoplus_{0 \le t \le T} (A_t \bigoplus B_t \bigoplus C_t), \\
&A_t = \{q(s_t, s_{t-1}, a_{t-1}); \mathrm{E}[q(s_t, s_{t-1}, a_{t-1})|s_{t-1}, a_{t-1}] = 0,\ q \in L^2\}, \\
&B_t = \{q(a_t, s_t); \mathrm{E}[q(a_t, s_t)|s_t] = 0,\ q \in L^2\}, \\
&C_t = \{q(r_t, s_t, a_t); \mathrm{E}[q(r_t, s_t, a_t)|s_t, a_t] = 0,\ q \in L^2\}.
\end{aligned}$$

The orthogonal space of the tangent space is the product of

$$\bigoplus_{0 \le t \le T} (A_t' \bigoplus B_t' \bigoplus C_t') \tag{15}$$

such that

$$\begin{aligned}
A_t' \bigoplus A_t = A_t'',\ A_t'' = \{q(\mathcal{J}_{s_t}); \mathrm{E}[q(\mathcal{J}_{s_t})|\mathcal{J}_{r_{t-1}}] = 0,\ q \in L^2\}, \\
B_t' \bigoplus B_t = B_t'',\ B_t'' = \{q(\mathcal{J}_{a_t}); \mathrm{E}[q(\mathcal{J}_{a_t})|\mathcal{J}_{s_t}] = 0,\ q \in L^2\},
\end{aligned}$$

$$C_t' \bigoplus C_t = C_t'',\ C_t'' = \{q(\mathcal{J}_{r_t}); \mathrm{E}[q(\mathcal{J}_{r_t})|\mathcal{J}_{a_t}] = 0,\ q \in L^2\}.$$

More specifically, the orthogonal tangent space is represented as

$$\begin{aligned}
A_t' &= \left\{q(\mathcal{J}_{s_t}) - \mathrm{E}[q(\mathcal{J}_{s_t})|s_t, a_{t-1}, s_{t-1}]; \mathrm{E}[q(\mathcal{J}_{s_t})|\mathcal{J}_{r_{t-1}}] = 0,\ q \in L^2\right\},\\
B_t' &= \left\{q(\mathcal{J}_{a_t}) - \mathrm{E}[q(\mathcal{J}_{r_t})|s_t, a_t]; \mathrm{E}[q(\mathcal{J}_{a_t})|\mathcal{J}_{s_t}] = 0,\ q \in L^2\right\},\\
C_t' &= \left\{q(\mathcal{J}_{r_t}) - \mathrm{E}[q(\mathcal{J}_{r_t})|r_t, s_t, a_t]; \mathrm{E}[q(\mathcal{J}_{r_t})|\mathcal{J}_{a_t}] = 0,\ q \in L^2\right\}.
\end{aligned}$$

If we can prove that the influence function Eq. (14) is orthogonal to the orthogonal tangent space Eq. (15), we can see that the above influence function is actually the efficient influence function. This fact is shown as follows.

Lemma 21. *The derivative Eq.* (14) *is orthogonal to* $\{A_t'\}_{t=0}^{T+1}, \{B_t''\}_{t=0}^{T}, \{C_t'\}_{t=0}^{T}$.

Proof. First, the influence function Eq. (14) is orthogonal to A_k'; for $t(\mathcal{J}_{s_k}) \in A_k'$

$$\begin{aligned}
&\mathrm{E}\left[\left\{v_0 + \sum_{t=0}^{T} \mu_t(s_t, a_t)(r_t + v_{t+1} - q_t)\right\} t(\mathcal{J}_{s_k})\right]\\
&= \mathrm{E}\left[\left\{\sum_{t=k-1}^{T} \mu_t(s_t, a_t)(r_t + v_{t+1} - q_t)\right\} t(\mathcal{J}_{s_k})\right]\\
&= \mathrm{E}\left[\mu_{k-1}(s_{k-1}, a_{k-1})(r_{k-1} + v_k - q_{k-1}) t(\mathcal{J}_{s_k})\right]\\
&= \mathrm{E}\left[\mu_{k-1}(s_{k-1}, a_{k-1}) v_k t(\mathcal{J}_{s_k})\right]\\
&= \mathrm{E}\left[\mu_{k-1}(s_{k-1}, a_{k-1}) v_k \mathrm{E}[t(\mathcal{J}_{s_k})|s_k, a_{k-1}, s_{k-1}]\right] = 0
\end{aligned}$$

Second, the influence function Eq. (14) is orthogonal to B_k''; for $t(\mathcal{J}_{a_k}) \in B_k''$

$$\begin{aligned}
&\mathrm{E}\left[\left\{v_0 + \sum_{t=0}^{T} \mu_t(s_t, a_t)(r_t + v_{t+1} - q_t)\right\} t(\mathcal{J}_{a_k})\right]\\
&= \mathrm{E}\left[\left\{\sum_{t=k}^{T} \mu_t(s_t, a_t)(r_t + v_{t+1} - q_t)\right\} t(\mathcal{J}_{a_k})\right] = 0.
\end{aligned}$$

Third, the influence function Eq. (14) is orthogonal to C_k'; for $t(\mathcal{J}_{r_k}) \in C_k'$

$$\begin{aligned}
&\mathrm{E}\left[\left\{v_0 + \sum_{t=0}^{T} \mu_t(s_t, a_t)(r_t + v_{t+1} - q_t)\right\} t(\mathcal{J}_{r_k})\right]\\
&= \mathrm{E}\left[\left\{\sum_{t=k}^{T} \mu_t(s_t, a_t)(r_t + v_{t+1} - q_t)\right\} t(\mathcal{J}_{r_k})\right]\\
&= \mathrm{E}\left[\{\mu_k(s_k, a_k)(r_k + v_{k+1} - q_k)\} t(\mathcal{J}_{r_k})\right]\\
&= \mathrm{E}\left[\{\mu_k(s_k, a_k)(\mathrm{E}[r_k + \mathrm{E}[v_{k+1}|\mathcal{J}_{r,k}] - q_k)\} t(\mathcal{J}_{r_k})\right]\\
&= \mathrm{E}\left[\{\mu_k(s_k, a_k)(r_k + \mathrm{E}[v_{k+1}|s_k] - q_k)\} \mathrm{E}[t(\mathcal{J}_{r_k})|s_k, a_k, r_k]\right] = 0.
\end{aligned}$$

□

Efficient influence function under $\mathcal{M}_{\mathrm{MDP}-b}$. In Lemma 21, we check that the $\phi_{\mathrm{eff}}^{\mathrm{MDP}}$ is orthogonal to $B"_t$. This concludes the proof noting that the orthogonal tangent space of $\mathcal{M}_{\mathrm{MDP}-b}$ is

$$\bigoplus_{0 \le t \le T} (A_t' \bigoplus B"_t \bigoplus C_t').$$

Efficiency bound. To show an efficiency bound, we use a law of total variance (Bowsher & Swain, 2012). Recall that We can also easily derive this variance form using another equivalent form of efficient influence function.

$$
\begin{aligned}
&\text{var}\left[v_0 + \sum_{t=0}^{T} \mu_t(s_t, a_t)(r_t + v_{t+1} - q_t)\right] \\
&= \sum_{t=0}^{T+1} \text{E}\left[\text{var}\left[\text{E}\left[v_0 + \sum_{t=0}^{T} \mu_k(s_k, a_k)(r_k + v_{k+1} - q_k)|\mathcal{J}_{a_t}\right]|\mathcal{J}_{a_{t-1}}\right]\right] \\
&= \sum_{t=0}^{T+1} \text{E}\left[\text{var}\left[\text{E}\left[\sum_{k=t-1}^{T} \mu_k(s_k, a_k)(r_k + v_{k+1} - q_k)|\mathcal{J}_{a_t}\right]|\mathcal{J}_{a_{t-1}}\right]\right] \\
&= \sum_{t=0}^{T+1} \text{E}\left[\text{var}\left[\text{E}\left[\mu_{t-1}(s_{t-1}, a_{t-1})(r_{t-1} + v_t - q_{t-1})|\mathcal{J}_{a_t}\right]|\mathcal{J}_{a_{t-1}}\right]\right] \\
&= \sum_{t=0}^{T+1} \text{E}\left[\text{var}\left[\mu_{t-1}(s_{t-1}, a_{t-1})(r_{t-1} + v_t - q_{t-1})|\mathcal{J}_{a_{t-1}}\right]\right] \\
&= \sum_{t=0}^{T+1} \text{E}\left[\mu_{t-1}^2(s_{t-1}, a_{t-1})\text{var}\left[(r_{t-1} + v_t)|\mathcal{J}_{a_{t-1}}\right]\right].
\end{aligned}
$$

□

Proof of Theorem 4. From Jensen's inequality,

$$
\begin{aligned}
&\sum_{t=0}^{T+1} \text{E}\left[\nu_{t-1}^2 \text{var}\left\{r_t + v_t(s_t)|s_{t-1}, a_{t-1}\right\}\right] = \sum_{t=0}^{T+1} \text{E}\left[\text{E}(\nu_{t-1}^2|s_{t-1}, a_{t-1})\text{var}\left\{r_t + v_t(s_t)|s_{t-1}, a_{t-1}\right\}\right] \\
&\geq \sum_{t=0}^{T+1} \text{E}\left[\text{E}(\nu_{t-1}|s_{t-1}, a_{t-1})^2\text{var}\left\{r_t + v_t(s_t)|s_{t-1}, a_{t-1}\right\}\right] = \sum_{t=0}^{T+1} \text{E}\left[\mu_{t-1}^2 \text{var}\left\{r_t + v_t(s_t)|s_{t-1}, a_{t-1}\right\}\right].
\end{aligned}
$$

When ν_{t-1}^2 is not constant given s_{t-1}, a_{t-1} and $\text{var}\left\{r_t + v_t(s_t)|s_{t-1}, a_{t-1}\right\} \neq 0$, the inequality is strict. □

Proof of Theorem 5. By changing the limits of summation and letting $r_{-1} = 0$, $\lambda_0 = 1$, we can write the efficiency bound under NMDP as

$$
\begin{aligned}
\sum_{t=0}^{T+1} \text{E}\left[\lambda_{t-1}^2 \text{var}\left\{r_{t-1} + v_t(\mathcal{H}_{s_t}) \mid \mathcal{H}_{a_{t-1}}\right\}\right] &\leq C^{T+1} \sum_{t=0}^{T+1} \text{E}\left[\lambda_{t-1}\text{var}\left\{r_{t-1} + v_t(\mathcal{H}_{s_t})|\mathcal{H}_{a_{t-1}}\right\}\right] \\
&= C^{T+1} \sum_{t=0}^{T+1} \text{E}_{\pi^e}\left[\text{var}_{\pi^e}\left\{r_{t-1} + v_t(\mathcal{H}_{s_t})|\mathcal{H}_{a_{t-1}}\right\}\right] \\
&= C^{T+1} \sum_{t=0}^{T+1} \text{E}_{\pi^e}\left[\text{var}\left\{r_{t-1} + v_t(\mathcal{H}_{s_t})|\mathcal{H}_{a_{t-1}}\right\}\right] \\
&= C^{T+1}\text{var}[\sum_{t=0}^{T+1} r_{t-1}] \\
&\leq C^{T+1}(T+1)^2 R_{\max}^2.
\end{aligned}
$$

The last equality follows by the law of total variance.

Similarly, the efficiency bound under MDP is

$$
\sum_{t=0}^{T+1} \text{E}\left[\mu_{t-1}^2 \text{var}\left\{r_{t-1} + v_t(s_t) \mid s_{t-1}, a_{t-1}\right\}\right] \leq C' \sum_{t=0}^{T+1} \text{E}\left[\mu_{t-1}\text{var}\left\{r_{t-1} + v_t(s_t) \mid s_{t-1}, a_{t-1}\right\}\right]
$$

$$
\begin{aligned}
&= C' \sum_{t=0}^{T+1} \mathrm{E}_{\pi^e} \left[\mathrm{var} \left\{ r_{t-1} + v_t(s_t) \mid s_{t-1}, a_{t-1} \right\} \right] \\
&= C' \sum_{t=0}^{T+1} \mathrm{E}_{\pi^e} \left[\mathrm{var}_{\pi^e} \left\{ r_{t-1} + v_t(s_t) \mid s_{t-1}, a_{t-1} \right\} \right] \\
&= C' \mathrm{var}[\sum_{t=0}^{T+1} r_{t-1}] \\
&\leq C'(T+1)^2 R_{\max}^2.
\end{aligned}
$$

The last equality again follows by the law of total variance.

Finally, for the NMDP lower bound we have by Jensen's inequality

$$
\begin{aligned}
\sum_{t=0}^{T+1} \mathrm{E} \left[\lambda_{t-1}^2 \mathrm{var} \left\{ r_{t-1} + v_t(\mathcal{H}_{s_t}) \mid \mathcal{H}_{a_{t-1}} \right\} \right] &= \sum_{t=0}^{T+1} \mathrm{E}_{\pi^e} \left[\lambda_{t-1} \mathrm{var} \left\{ r_{t-1} + v_t(\mathcal{H}_{s_t}) \mid \mathcal{H}_{a_{t-1}} \right\} \right] \\
&\geq \sum_{t=0}^{T+1} \exp \mathrm{E}_{\pi^e} \left[\log(\lambda_{t-1} \mathrm{var} \left\{ r_{t-1} + v_t(\mathcal{H}_{s_t}) \mid \mathcal{H}_{a_{t-1}} \right\}) \right] \\
&\geq \sum_{t=0}^{T+1} \exp(\mathrm{E}_{\pi^e} \left[\log(\lambda_{t-1}) \right] + \mathrm{E}_{\pi^e} \left[\mathrm{var} \left\{ r_{t-1} + v_t(\mathcal{H}_{s_t}) \mid \mathcal{H}_{a_{t-1}} \right\}) \right]) \\
&\geq \sum_{t=0}^{T+1} \exp(t \log C_{\min} + \log V_{\min}^2) \\
&\geq V_{\min}^2 C_{\min}^{T+1}.
\end{aligned}
$$

□

Proof of Theorem 6. Define $\phi(\{\hat{\nu}_k\}, \{\hat{q}_k\})$ as:

$$\textstyle\sum_{k=0}^{T} \hat{\nu}_k r_k - \{\hat{\nu}_{k-1} \hat{q}_k - \hat{\nu}_k \mathrm{E}_{\pi^e}[\hat{q}_k(\mathcal{H}_{a_k}) | \mathcal{H}_{s_k}]\}.$$

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

$$\tfrac{n_0}{n} \mathbb{P}_{n_0} \phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) + \tfrac{n_1}{n} \mathbb{P}_{n_1} \phi(\{\hat{\nu}_k^{(0)}\}, \{\hat{q}_k^{(0)}\}),$$

where $\mathbb{P}_{n_0}$ is an empirical approximation based on a set of samples such that $J = 0$, $\mathbb{P}_{n_1}$ is an empirical approximation based on a set of samples such that $J = 1$. Then, we have

$$
\begin{aligned}
\sqrt{n}(\mathbb{P}_{n_0} \phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) &= \sqrt{n/n_0} \mathbb{G}_{n_0}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k\}, \{q_k\})] && (16) \\
&+ \sqrt{n/n_0} \mathbb{G}_{n_0}[\phi(\{\nu_k^{(1)}\}, \{q_k^{(1)}\})] && (17) \\
&+ \sqrt{n}(\mathrm{E}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) | \{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] - \rho^{\pi^e}). && (18)
\end{aligned}
$$

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

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

$$D_3 = \sum_{k=0}^{T} (\hat{\nu}_k - \nu_k)(r_k - q_k + v_{k+1}).$$

First, we show the term Eq. (16) is $o_p(1)$.

Lemma 22. *The term Eq. (16) is* $o_p(1)$.

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

$$\lim_{n\to\infty} \sqrt{n_0} P[\mathbb{P}_{n_0}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k\}, \{q_k\})] \\ - \mathrm{E}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k\}, \{q_k\})|\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon|\mathcal{D}_1] = 0, \tag{19}$$

Then, by bounded convergence theorem, we would have

$$\lim_{n\to\infty} \sqrt{n_0} P[\mathbb{P}_{n_0}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k\}, \{q_k\})] \\ - \mathrm{E}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k\}, \{q_k\})|\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon] = 0,$$

yielding the statement.

To show Eq. (19), we show that the conditional mean is 0 and conditional variance is $o_p(1)$. The conditional mean is

$$\mathrm{E}[\mathbb{P}_{n_0}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k\}, \{q_k\})|\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] \\ - \mathbb{P}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k\}, \{q_k\})]|\mathcal{D}_1] = 0.$$

Here, we leveraged 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

$$\begin{aligned}
&\mathrm{var}[\sqrt{n_0}\mathbb{P}_{n_0}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k\}, \{q_k\})]|\mathcal{D}_1] \\
&\quad = \mathrm{E}[\mathrm{E}[D_1^2 + D_2^2 + D_3^2 + 2D_1D_2 + 2D_2D_3 + 2D_2D_3|\{\hat{q}_k^{(1)}\}, \{\nu_k^{(1)}\}]|\mathcal{D}_1] \\
&\quad = o_p(1).
\end{aligned}$$

Here, we used the convergence rate assumption and the relation $\|\hat{v}_k^{(1)} - v_k\|_2 < \|\hat{q}_k^{(1)} - q_k\|_2$ arising from the fact that the former is the marginalization of the latter over π_k^e. Then, from Chebyshev's inequality:

$$\begin{aligned}
&\sqrt{n_0} P[\mathbb{P}_{n_0}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k\}, \{q_k\})] - \mathrm{E}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) \\
&\qquad - \phi(\{\nu_k\}, \{q_k\})|\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon|\mathcal{D}_1] \\
&\leq \frac{1}{\epsilon^2} \mathrm{var}[\sqrt{n_0}\mathbb{P}_{n_0}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k\}, \{q_k\})]|\mathcal{D}_1] = o_p(1).
\end{aligned}$$

□

Lemma 23. *The term Eq. (18) is* $o_p(1)$.

Proof.

$$\begin{aligned}
&\sqrt{n}\mathrm{E}[\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \mathrm{E}[\phi(\{\nu_k\}, \{q_k\})]|\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T} (\hat{\nu}_k^{(1)} - \nu_k)(-\hat{q}_k^{(1)} + q_k) + (\hat{\nu}_{k-1}^{(1)} - \nu_{k-1})(-\hat{v}_k + v_k)|\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T} \nu_k(\hat{q}_k^{(1)} - q_k) + \nu_{k-1}(\hat{v}_k - v_k)|\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T} (\hat{\nu}_k^{(1)} - \nu_k)(r_k - q_k + v_{k+1})|\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}]
\end{aligned}$$

$$= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\nu}_k^{(1)} - \nu_k)(-\hat{q}_k^{(1)} + q_k) + (\hat{\nu}_{k-1}^{(1)} - \nu_{k-1})(-\hat{v}_k + v_k)|\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}]$$
$$= \sqrt{n}\sum_{k=0}^{T}\mathcal{O}(\|\hat{\nu}_k^{(1)} - \nu_k\|_2\|\hat{q}_k^{(1)} - q_k\|_2) = \sqrt{n}\sum_{k=0}^{T}o_p(n^{-\alpha_{t,1}})o_p(n^{-\alpha_{t,2}}) = o_p(1). \quad \square$$

Finally, we get

$$\sqrt{n}(\mathbb{P}_{n_0}\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) = \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\nu_k\}, \{q_k\})] + o_p(1).$$

Therefore,

$$\begin{aligned}
&\sqrt{n}(\hat{\rho}^{\pi^e}_{\mathrm{DRL(NMDP)}} - \rho^{\pi^e}) \\
&= n_0/n\sqrt{n}\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) + n_1/n\sqrt{n}(\mathbb{P}_{n_1}\phi(\{\hat{\nu}_k^{(0)}\}, \{\hat{q}_k^{(0)}\}) - \rho^{\pi^e}) \\
&= \sqrt{n_0/n}\mathbb{G}_{n_0}[\phi(\{\nu_k\}, \{q_k\})] + \sqrt{n_1/n}\mathbb{G}_{n_1}[\phi(\{\nu_k\}, \{q_k\})] + o_p(1) \\
&= \mathbb{G}_n[\phi(\{\nu_k\}, \{q_k\})] + o_p(1),
\end{aligned}$$

concluding the proof by showing the influence function of $\hat{\rho}^{\pi^e}_{\mathrm{DRL(NMDP)}}$ is the efficient one. $\square$

Proof of Theorem 8. We define $\phi(\{\hat{\nu}_k\}, \{\hat{q}_k\})$ as:

$$\textstyle\sum_{k=0}^{T}\hat{\nu}_k r_k - \hat{\nu}_{k-1}\{\hat{\eta}_k\hat{q}_k - \mathrm{E}_{\pi^e}[\hat{q}_k(\mathcal{H}_{a_k})|\mathcal{H}_{s_k}]\}.$$

The estimator $\hat{\rho}^{\pi^e}_{\mathrm{DR}}$ is given by $\mathbb{P}_n\phi(\{\hat{\nu}_k\}, \{\hat{q}_k\})$. Then, we have

$$\begin{aligned}
\sqrt{n}(\mathbb{P}_n\phi(\{\hat{\nu}_k\}, \{\hat{q}_k\}) - \rho^{\pi^e}) &= \mathbb{G}_n[\phi(\{\hat{\nu}_k\}, \{\hat{q}_k\}) - \phi(\{\nu_k\}, \{q_k\})] && (20)\\
&\quad + \mathbb{G}_n[\phi(\{\nu_k\}, \{q_k\})] && (21)\\
&\quad + \sqrt{n}(\mathrm{E}[\phi(\{\hat{\nu}_k\}, \{\hat{q}_k\})|\{\hat{\nu}_k\}, \{\hat{q}_k\}] - \rho^{\pi^e}). && (22)
\end{aligned}$$

If we can prove that the term Eq. (20) is $\mathrm{o}_p(1)$, the statement is concluded as in the proof of Theorem 6. We proceed to prove this.

First, we show $\phi(\{\hat{\nu}_k\}, \{\hat{q}_t\}) - \phi(\{\nu_t\}, \{q_t\})$ belongs to a Donsker class. The transformation

$$(\{\nu_k\}, \{q_k\}) \mapsto \textstyle\sum_{k=0}^{T}\nu_k r_k - \{\nu_k q_k - \nu_{k-1}\mathrm{E}_{\pi^e}[q_k(\mathcal{H}_{a_k})|\mathcal{H}_{s_k}]\}$$

is a Lipschitz function. Therefore, by Example 19.20 in van der Vaart (1998), $\phi(\{\hat{\nu}_k\}, \{\hat{q}_k\}) - \phi(\{\nu_k\}, \{q_k\})$ is an also Donsker class. In addition, we can also show that

$$\|\phi(\{\hat{\nu}_k\}, \{\hat{q}_k\}) - \phi(\{\nu_k\}, \{q_k\})\|_2 = \mathrm{o}_p(1),$$

as in Lemma 22. Therefore, from Lemma 19.24 in (van der Vaart, 1998), the term Eq. (20) is $\mathrm{o}_p(1)$, concluding the proof. $\square$

Proof of Theorem 9. We use the following doubly robust structure

$$\begin{aligned}
&\mathrm{E}\left[\sum_{k=0}^{T}\nu_k r_k - \{\nu_k q_k - \nu_{k-1}\mathrm{E}_{\pi^e}(q_k|\mathcal{H}_{s_k})\}\right] \\
&= \mathrm{E}\{\mathrm{E}_{\pi^e}(q_0|s_0)\} + \mathrm{E}\left[\sum_{k=0}^{T}\nu_k\{r_k - q_k + \mathrm{E}_{\pi^e}(q_k|\mathcal{H}_{s_{k+1}})\}\right] = \rho^{\pi^e}.
\end{aligned}$$

Then, as in the proof of Theorem 6,

$$\sqrt{n}(\mathbb{P}_{n_0}\phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e})$$

$$
\begin{aligned}
&= \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k^\dagger\},\{q_k^\dagger\})] + \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\nu_k^\dagger\},\{q_k^\dagger\})] \\
&+ \sqrt{n/n_0}(\mathrm{E}[\phi(\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\})|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] - \mathrm{E}[\phi(\{\nu_k^\dagger\},\{q_k^\dagger\})]) + \sqrt{n}(\mathrm{E}[\phi(\{\nu_k^\dagger\},\{q_k^\dagger\})] - \rho^{\pi^e}) \\
&= \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\nu_k^\dagger\},\{q_k^\dagger\})] + \sqrt{n}(\mathrm{E}[\phi(\{\nu_k^\dagger\},\{q_k^\dagger\})] - \rho^{\pi^e}) + o_p(1).
\end{aligned}
$$

Here, we used

$$
\sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \phi(\{\nu_k^\dagger\},\{q_k^\dagger\})] = \mathrm{o}_p(1)
$$

from Lemma 22 and

$$
\sqrt{n/n_0}(\mathrm{E}[\phi(\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\})|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] - \mathrm{E}[\phi(\{\nu_k^\dagger\},\{q_k^\dagger\})]) = \mathcal{O}_p(1),
$$

which we prove below as in Lemma 23.

Lemma 24.

$$
\sqrt{n/n_0}(\mathrm{E}[\phi(\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\})|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] - \mathrm{E}[\phi(\{\nu_k^\dagger\},\{q_k^\dagger\})]) = \mathcal{O}_p(1).
$$

Proof. First, consider the case where $\nu_k = \nu_k^\dagger$.

$$
\begin{aligned}
&\sqrt{n}\mathrm{E}[\phi(\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \mathrm{E}[\phi(\{\nu_k\},\{q_k^\dagger\})]|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\nu}_k^{(1)} - \nu_k)(-\hat{q}_k^{(1)} + q_k^\dagger) + (\hat{\nu}_{k-1}^{(1)} - \nu_{k-1})(-\hat{v}_k + v_k^\dagger)|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}\nu_k(\hat{q}_k^{(1)} - q_k^\dagger) + \nu_{k-1}(\hat{v}_k - v_k^\dagger)|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\nu}_k^{(1)} - \nu_k)(r_k - q_k^\dagger + v_{k+1}^\dagger)|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\nu}_k^{(1)} - \nu_k)(-\hat{q}_k^{(1)} + q_k^\dagger) + (\hat{\nu}_{k-1}^{(1)} - \nu_{k-1})(-\hat{v}_k + v_k^\dagger)|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\nu}_k^{(1)} - \nu_k)(r_k - q_k^\dagger + v_{k+1}^\dagger)|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\sum_{k=0}^{T}\mathcal{O}(\|\hat{\nu}_k^{(1)} - \nu_k\|_2\|\hat{q}_k^{(1)} - q_k^\dagger\|_2 + \|\hat{\nu}_k^{(1)} - \nu_k\|^2) \\
&= \sqrt{n}\sum_{k=0}^{T}\left\{\mathcal{O}_p(n^{-\alpha_{t,1}})o_p(n^{-\alpha_{t,2}}) + \mathcal{O}_p(n^{-\alpha_{t,1}})\right\} = \mathcal{O}_p(1).
\end{aligned}
$$

Next, consider the case where $q_k = q_k^\dagger$:

$$
\begin{aligned}
&\sqrt{n}\mathrm{E}[\phi(\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \mathrm{E}[\phi(\{\nu_k^\dagger\},\{q_k\})]|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\nu}_k^{(1)} - \nu_k^\dagger)(-\hat{q}_k^{(1)} + q_k) + (\hat{\nu}_{k-1}^{(1)} - \nu_{k-1}^\dagger)(-\hat{v}_k + v_k)|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}\nu_k^\dagger(\hat{q}_k^{(1)} - q_k) + \nu_{k-1}^\dagger(\hat{v}_k - v_k)|\{\hat{\nu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\sum_{k=0}^{T}\mathcal{O}(\|\hat{\nu}_k^{(1)} - \nu_k^\dagger\|_2\|\hat{q}_k^{(1)} - q_k\|_2 + \|\hat{q}_k^{(1)} - q_k\|_2)
\end{aligned}
$$

$$= \sqrt{n} \sum_{k=0}^{T} \left\{ \mathcal{O}_p(n^{-\alpha_{t,1}}) o_p(n^{-\alpha_{t,2}}) + \mathcal{O}_p(n^{-\alpha_{t,2}}) \right\} = \mathcal{O}_p(1). \qquad \square$$

Using the above result, we prove the statement for each case below.

ν-model is well-specified. First, consider the case when $\nu_k^\dagger = \nu_k$:

$$\begin{aligned} \mathrm{E}[\phi(\{\nu_k\}, \{q_k^\dagger\})] &= \mathrm{E}[\sum_{k=0}^{T} [\nu_k r_k - \{\nu_k q_k^\dagger(\mathcal{H}_{a_k}) - \nu_{k-1} \mathrm{E}_{\pi^e}[q_k^\dagger(\mathcal{H}_{a_k})|s_k]] \\ &= \mathrm{E}[\sum_{k=0}^{T} \nu_k r_k] = \rho^{\pi^e}. \end{aligned}$$

Then,

$$\sqrt{n}(\mathbb{P}_{n_0} \phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) = \sqrt{n/n_0} \mathbb{G}_{n_0}[\phi(\{\nu_k\}, \{q_k^\dagger\})] + \mathcal{O}_p(1).$$

Therefore,

$$\begin{aligned} \sqrt{n}(\hat{\rho}^{\pi^e}_{\mathrm{DRL(NMDP)}} - \rho^{\pi^e}) &= \sqrt{n_0/n} \mathbb{G}_{n_0}[\phi(\{\nu_k\}, \{q_k^\dagger\})] + \sqrt{n_1/n} \mathbb{G}_{n_1}[\phi(\{\nu_k\}, \{q_k^\dagger\})] + \mathcal{O}_p(1) \\ &= \sqrt{n} \mathbb{G}_n[\phi(\{\nu_k\}, \{q_k^\dagger\})] + \mathcal{O}_p(1) = \mathcal{O}_p(1). \end{aligned}$$

which shows $\hat{\rho}^{\pi^e}_{\mathrm{DRL(NMDP)}}$ is $\sqrt{n}$-consistent when the model for the behavior policy is well-specified.

q-model is well-specified. Next, consider the case where $q_k^\dagger = q_k$.

$$\begin{aligned} &\mathrm{E}[\phi(\{\nu_k^\dagger\}, \{q_k\})] = \mathrm{E}\left[\mathrm{E}_{\pi^e}[q_0(\mathcal{H}_{a_0})|s_0] + \sum_{k=0}^{T} \nu_k^\dagger \{r_k - q_k(\mathcal{H}_{a_k}) + \mathrm{E}_{\pi^e}[q_k(\mathcal{H}_{a_k})|s_{k+1}]\} \right] \\ &= \mathrm{E}\left[\mathrm{E}_{\pi^e}[q_0(\mathcal{H}_{a_0})|s_0] \right] = \rho^{\pi^e}. \end{aligned}$$

Then,

$$\sqrt{n}(\mathbb{P}_{n_0} \phi(\{\hat{\nu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) = \sqrt{n/n_0} \mathbb{G}_{n_0}[\phi(\{\nu_k^\dagger\}, \{q_k\})] + \mathcal{O}_p(1).$$

Therefore,

$$\begin{aligned} \sqrt{n}(\hat{\rho}^{\pi^e}_{\mathrm{DRL(NMDP)}} - \rho^{\pi^e}) &= \sqrt{n_0/n} \mathbb{G}_{n_0}[\phi(\{\nu_k^\dagger\}, \{q_k\})] + \sqrt{n_1/n} \mathbb{G}_{n_1}[\phi(\{\nu_k^\dagger\}, \{q_k\})] + \mathcal{O}_p(1) \\ &= \sqrt{n} \mathbb{G}_n[\phi(\{\nu_k^\dagger\}, \{q_k\})] + \mathcal{O}_p(1) = \mathcal{O}_p(1). \end{aligned}$$

which shows $\hat{\rho}^{\pi^e}_{\mathrm{DRL(NMDP)}}$ is $\sqrt{n}$-consistent when the model for the q-function is well-specified. $\square$

Proof of Lemma 10. We have

$$\begin{aligned} &\left\| \prod_{0=t}^{k} \frac{\pi_t^e}{\hat{\pi}_t^b} - \nu_k \right\|_2 \leq \left\| \sum_{i=0}^{T} \left(\prod_{t=0}^{i} \frac{\pi_t^e}{\hat{\pi}_t^b} \prod_{t=i}^{k} \frac{\pi_t^e}{\pi_t^b} - \prod_{t=0}^{i-1} \frac{\pi_t^e}{\hat{\pi}_t^b} \prod_{t=i}^{k} \frac{\pi_t^e}{\pi_t^b} \right) \right\|_2 \\ &\leq \sum_{t=0}^{k} \mathcal{O}(\|\pi_t^e/\hat{\pi}_t^b - \eta_t\|_2) \\ &= \mathcal{O}_p(n^{-(\alpha)/(\alpha + d_{\mathcal{H}_{s_k}})}). \end{aligned}$$

$\square$

Proof of Theorem 11. Define $\phi(\{\hat{\mu}_k\}, \{\hat{q}_k\})$ as:

$$\textstyle\sum_{k=0}^{T} \hat{\mu}_k r_k - \hat{\nu}_{k-1}\{\hat{\eta}_k \hat{q}_k - \mathrm{E}_{\pi^e}[\hat{q}_k(\mathcal{H}_{a_k})|\mathcal{H}_{s_k}]\}.$$

The estimator $\hat{\rho}_{\mathrm{DRL(MDP)}}$ is given by

$$\tfrac{n_0}{n}\mathbb{P}_{n_0}\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) + \tfrac{n_1}{n}\mathbb{P}_{n_1}\phi(\{\hat{\mu}_k^{(0)}\}, \{\hat{q}_k^{(0)}\}).$$

Then, we have

$$\begin{aligned}\sqrt{n}(\mathbb{P}_{n_0}\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) &= \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})] &(23)\\ &+ \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\mu_k\}, \{q_k\})] &(24)\\ &+ \sqrt{n}(\mathrm{E}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\})|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] - \rho^{\pi^e}). &(25)\end{aligned}$$

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

$$\begin{aligned}&\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\}) = D_1 + D_2 + D_3, \quad \text{where}\\ &D_1 = \sum_{k=0}^{T}(\hat{\mu}_k^{(1)} - \mu_k)(-\hat{q}_k^{(1)} + q_k) + (\hat{\mu}_{k-1}^{(1)} - \mu_{k-1})(-\hat{v}_k + v_k),\\ &D_2 = \sum_{k=0}^{T}\mu_k(\hat{q}_k^{(1)} - q_k) + \mu_{k-1}(\hat{v}_k - v_k),\\ &D_3 = \sum_{k=0}^{T}(\hat{\mu}_k^{(1)} - \mu_k)(r_k - q_k + v_{k+1}).\end{aligned}$$

First, we show the term Eq. (23) is $o_p(1)$.

Lemma 25. *The term Eq. (23) is* $o_p(1)$.

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

$$\begin{aligned}\lim_{n\to\infty}\sqrt{n_0}P[\mathbb{P}_{n_0}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k^{(1)}\}, \{q_k^{(1)}\})] &\\ - \mathrm{E}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon|\mathcal{D}_1] = 0.&\end{aligned} \tag{26}$$

Then, by bounded convergence theorem, we have

$$\begin{aligned}\lim_{n\to\infty}\sqrt{n_0}P[\mathbb{P}_{n_0}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})] &\\ - \mathrm{E}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon] = 0,&\end{aligned}$$

yielding the statement.

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

$$\begin{aligned}\mathrm{E}[\mathbb{P}_{n_0}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] -&\\ \mathbb{P}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})]|\mathcal{D}_1] = 0.&\end{aligned}$$

Here, we used a sample splitting construction, that is, $\hat{\mu}_k^{(1)}$ and $\hat{q}_k^{(1)}$ only depend on $\mathcal{D}_1$. The conditional variance is

$$\begin{aligned}&\mathrm{var}[\sqrt{n_0}\mathbb{P}_{n_0}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})]|\mathcal{D}_1]\\ =\ &\mathrm{E}[\mathrm{E}[D_1^2 + D_2^2 + D_3^2 + 2D_1D_2 + 2D_2D_3 + 2D_2D_3|\{\hat{q}_k^{(1)}\}, \{\mu_k^{(1)}\}]|\mathcal{D}_1]\\ =\ &o_p(1).\end{aligned}$$

Here, we used the convergence rate assumption and the relation $\|\hat{v}_k^{(1)} - v_k\|_2 \leq \|\hat{q}_k^{(1)} - q_k\|_2$. Then, from Chebyshev's inequality;

$$\begin{aligned}
&\sqrt{n_0} P[\mathbb{P}_{n_0}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})] - \mathrm{E}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \\
&\phi(\{\mu_k\}, \{q_k\})|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon|\mathcal{D}_1] \\
&\leq \frac{1}{\epsilon^2}\mathrm{var}[\sqrt{n_0}\mathbb{P}_{n_0}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})]|\mathcal{D}_1] = o_p(1).
\end{aligned}$$

□

Lemma 26. *The term Eq. (25) is* $o_p(1)$.

Proof.

$$\begin{aligned}
&\sqrt{n}\mathrm{E}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \mathrm{E}[\phi(\{\mu_k\}, \{q_k\})]|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\mu}_k^{(1)} - \mu_k)(-\hat{q}_k^{(1)} + q_k) + (\hat{\mu}_{k-1}^{(1)} - \mu_{k-1})(-\hat{v}_k + v_k)|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}\mu_k(\hat{q}_k^{(1)} - q_k) + \mu_{k-1}(\hat{v}_k - v_k)|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\mu}_k^{(1)} - \mu_k)(r_k - q_k + v_{k+1})|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\mu}_k^{(1)} - \mu_k)(-\hat{q}_k^{(1)} + q_k) + (\hat{\mu}_{k-1}^{(1)} - \mu_{k-1})(-\hat{v}_k + v_k)|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\sum_{k=0}^{T}\mathcal{O}(\|\hat{\mu}_k^{(1)} - \mu_k\|_2\|\hat{q}_k^{(1)} - q_k\|_2) = \sqrt{n}\sum_{t=0}^{T} o_p(n^{-\alpha_{t,1}})o_p(n^{-\alpha_{t,2}}) = o_p(1).
\end{aligned}$$

□

Finally, we get

$$\sqrt{n}(\mathbb{P}_{n_0}\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) = \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\mu_k\}, \{q_k\})] + o_p(1).$$

Therefore,

$$\begin{aligned}
&\sqrt{n}(\hat{\rho}^{\pi^e}_{\mathrm{DRL(MDP)}} - \rho^{\pi^e}) \\
&= n_0/n\sqrt{n}\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) + n_1/n\sqrt{n}(\mathbb{P}_{n_1}\phi(\{\hat{\mu}_k^{(0)}\}, \{\hat{q}_k^{(0)}\}) - \rho^{\pi^e}) \\
&= \sqrt{n_0/n}\mathbb{G}_{n_0}[\phi(\{\mu_k\}, \{q_k\})] + \sqrt{n_1/n}\mathbb{G}_{n_1}[\phi(\{\mu_k\}, \{q_k\})] + o_p(1) \\
&= \mathbb{G}_n[\phi(\{\mu_k\}, \{q_k\})] + o_p(1),
\end{aligned}$$

concluding the proof by showing the influence function of $\hat{\rho}^{\pi^e}_{\mathrm{DRL(MDP)}}$ is the efficient one. □

Proof of Theorem 12. We use the following doubly robust structure

$$\begin{aligned}
&\mathrm{E}\left[\sum_{k=0}^{T}\mu_k r_k - \{\mu_k q_k - \mu_{k-1}\mathrm{E}_{\pi^e}(q_k|s_k)\}\right] \\
&= \mathrm{E}[\mathrm{E}_{\pi^e}(q_0|s_0)]] + \mathrm{E}\left[\sum_{k=0}^{T}\mu_k\{r_k - q_k + \mathrm{E}_{\pi^e}(q_k|s_{k+1})\}\right] = \rho^{\pi^e}.
\end{aligned}$$

Then, as in the proof of Theorem 6,

$$\sqrt{n}(\mathbb{P}_{n_0}\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e})$$

$$
\begin{aligned}
&= \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k^\dagger\}, \{q_k^\dagger\})] + \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\mu_k^\dagger\}, \{q_k^\dagger\})] \\
&\quad + \sqrt{n/n_0}(\mathrm{E}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}); \{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] \\
&\quad - \mathrm{E}[\phi(\{\mu_k^\dagger\}, \{q_k^\dagger\})]) + \sqrt{n}(\mathrm{E}[\phi(\{\mu_k^\dagger\}, \{q_k^\dagger\})] - \rho^{\pi^e}) \\
&= \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\mu_k^\dagger\}, \{q_k^\dagger\})] + \sqrt{n}(\mathrm{E}[\phi(\{\mu_k^\dagger\}, \{q_k^\dagger\})] - \rho^{\pi^e}) + \mathcal{O}_p(1).
\end{aligned}
$$

We proceed by considering each case.

μ-model is well-specified. First, consider the case when $\mu_k^\dagger = \mu_k$:

$$
\begin{aligned}
&\mathrm{E}[\phi(\{\mu_k\}, \{q_k^\dagger\})] = \mathrm{E}[\sum_{k=0}^{T}[\mu_k r_k - \{\mu_k q_k^\dagger(s_k, a_k) - \mu_{k-1}\mathrm{E}_{\pi^e}[q_k^\dagger(s_k, a_k)|s_k]] \\
&= \mathrm{E}[\sum_{k=0}^{T} \mu_k r_k] = \rho^{\pi^e}.
\end{aligned}
$$

Then,

$$
\sqrt{n}(\mathbb{P}_{n_0}\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) = \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\mu_k\}, \{q_k^\dagger\})] + \mathcal{O}_p(1).
$$

Therefore,

$$
\begin{aligned}
\sqrt{n}(\hat{\rho}_T^{\mathcal{M}2} - \rho^{\pi^e}) &= \sqrt{n_0/n}\mathbb{G}_{n_0}[\phi(\{\mu_k\}, \{q_k^\dagger\})] + \sqrt{n_1/n}\mathbb{G}_{n_1}[\phi(\{\mu_k\}, \{q_k^\dagger\})] + o_p(1) \\
&= \sqrt{n}\mathbb{G}_n[\phi(\{\mu_k\}, \{q_k^\dagger\})] + \mathcal{O}_p(1) = \mathcal{O}_p(1),
\end{aligned}
$$

which shows $\hat{\rho}^{\pi^e}_{\mathrm{DRL(MDP)}}$ is $\sqrt{n}$-consistent when the model for the μ-function is well-specified.

q-model is well-specified. Next, consider the case where $q_k^\dagger = q_k$:

$$
\begin{aligned}
&\mathrm{E}[\phi(\{\mu_k^\dagger\}, \{q_k\})] = \mathrm{E}\left[\mathrm{E}_{\pi^e}[q(s_k, a_k)|s_0] + \sum_{k=0}^{T} \mu_k^\dagger\{r_k - q_k(s_k, a_k) + \mathrm{E}_{\pi^e}[q_k(s_k, a_k)|s_{k+1}]\}\right] \\
&= \mathrm{E}\left[\mathrm{E}_{\pi^e}[q_0(s_0, a_0)|s_0]\right] = \rho^{\pi^e}.
\end{aligned}
$$

We have

$$
\sqrt{n}(\mathbb{P}_{n_0}\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) = \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\mu_k^\dagger\}, \{q_k\})] + \mathcal{O}_p(1).
$$

Therefore,

$$
\begin{aligned}
\sqrt{n}(\hat{\rho}_T^{\mathcal{M}2} - \rho^{\pi^e}) &= \sqrt{n/n_0}\mathbb{G}_{n_0}[\phi(\{\mu_k^\dagger\}, \{q_k\})] + \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\mu_k^\dagger\}, \{q_k\})] + \mathcal{O}_p(1) \\
&= \sqrt{n}\mathbb{G}_n[\phi(\{\mu_k^\dagger\}, \{q_k\})] + \mathcal{O}_p(1) = \mathcal{O}_p(1),
\end{aligned}
$$

which shows $\hat{\rho}^{\pi^e}_{\mathrm{DRL(MDP)}}$ is $\sqrt{n}$-consistent when the model for the q-function is well-specified. □