Journal of Machine Learning Research 21 (2020) 1-63 Submitted 10/19; Revised 7/20; Published 7/20

Double Reinforcement Learning for Efficient Off-Policy Evaluation in Markov Decision Processes

Nathan Kallus KALLUS@CORNELL.EDU
Department of Operations Research and Information Engineering and Cornell Tech
Cornell University
New York, NY 10044, USA

Masatoshi Uehara* MU223@CORNELL.EDU
Department of Computer Science and Cornell Tech
Cornell University
New York, NY 10044, USA

Editor: George Konidaris

Abstract

Off-policy evaluation (OPE) in reinforcement learning allows one to evaluate novel decision policies without needing to conduct exploration, which is often costly or otherwise infeasible. We consider for the first time the semiparametric efficiency limits of OPE in Markov decision processes (MDPs), where actions, rewards, and states are memoryless. We show existing OPE estimators may fail to be efficient in this setting. We develop a new estimator based on cross-fold estimation of q-functions and marginalized density ratios, which we term double reinforcement learning (DRL). We show that DRL is efficient when both components are estimated at fourth-root rates and is also doubly robust when only one component is consistent. We investigate these properties empirically and demonstrate the performance benefits due to harnessing memorylessness.

Keywords: Off-policy evaluation, Markov decision processes, Semiparametric efficiency, Double machine learning

1. Introduction

Off-policy evaluation (OPE) is the problem of estimating mean rewards of a given policy (target policy) for a sequential decision-making problem using data generated by the log of another policy (behavior policy). OPE is a key problem in reinforcement learning (RL) (Jiang and Li, 2016; Li et al., 2015; Liu et al., 2018; Mahmood et al., 2014; Munos et al., 2016; Precup et al., 2000; Thomas and Brunskill, 2016) and it finds applications as varied as healthcare (Murphy, 2003) and education (Mandel et al., 2014). Because data can be scarce, it is crucial to use all available data efficiently, while at the same time using flexible, nonparametric estimators that avoid misspecification error.

In this paper, our goal is to obtain an estimator for policy value with minimal asymptotic mean squared error under nonparametric models for the sequential decision process and behavior policy, that is, achieving the semiparametric efficiency bound (Bickel et al., 1998). Toward that end, we explore the efficiency bound and efficient influence function of the

*. Corresponding author

©2020 Nathan Kallus and Masatoshi Uehara.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v21/19-827.html.

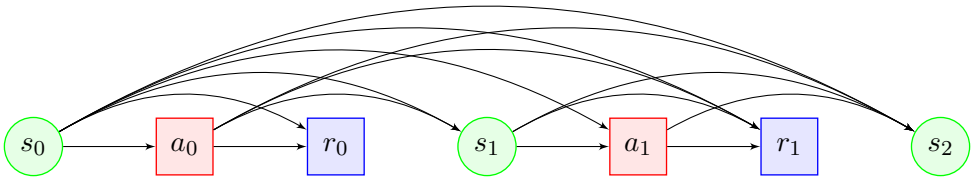

Figure 1: $\mathcal{M}_1$: Non-Markov decision process (NMDP)

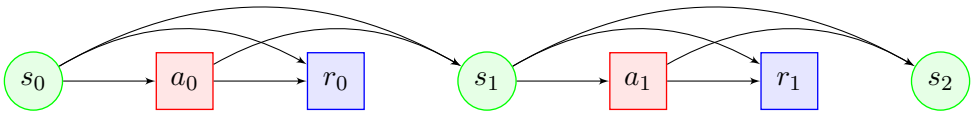

Figure 2: $\mathcal{M}_2$: Markov decision process (MDP)

$$\begin{array}{ccccccc} O(T^2) & = & \text{EffBd}(\mathcal{M}_2) & = & \text{EffBd}(\mathcal{M}_{2b}) & > & \text{EffBd}(\mathcal{M}_{2q}) \\ & & \wedge & & \wedge & & \\ 2^{\Omega(T)} & = & \text{EffBd}(\mathcal{M}_1) & = & \text{EffBd}(\mathcal{M}_{1b}) & > & \text{EffBd}(\mathcal{M}_{1q}) \end{array}$$

Figure 3: Relationship between the semiparametric efficiency bounds in each model, which lower bound achievable mean-squared error. $\mathcal{M}_1$, $\mathcal{M}_2$ are, respectively, NMDP and MDP with unknown behavior policy. $\mathcal{M}_{1b}$, $\mathcal{M}_{2b}$ are with known behavior policy. $\mathcal{M}_{1q}$, $\mathcal{M}_{2q}$ are with parametric assumptions on q-functions. Inequalities are generically strict (Theorem 4), and the MDP bound is generally polynomial in horizon length T while the NMDP bound is generally exponential (see Theorem 5).

target policy value under two models: non-Markov decision processes (NMDP) and Markov decision processes (MDP). The two models are illustrated in Figs. 1 and 2 and defined precisely in Section 1.1. While much work has studied efficient estimation under $\mathcal{M}_1$ (Dudik et al., 2014; Jiang and Li, 2016; Kallus and Uehara, 2019; Thomas and Brunskill, 2016), work on $\mathcal{M}_2$ has been restricted to the parametric, finite-state-finite-action case (Jiang and Li, 2016) and no globally efficient estimators have been proposed. The two models are clearly nested and indeed we obtain that the efficiency bounds are generally strictly ordered (see Fig. 3). In other words, if we correctly leverage the Markov property, we can obtain OPE estimators that are *more efficient* than existing ones. This is quite important, given the practical difficulty of evaluation in long horizons (see, *e.g.*, Gottesman et al., 2019) and given that many RL problems are Markovian. In particular, our results show the NMDP efficiency bound is generally exponential in horizon length so that estimators that target the NMDP model *necessarily* suffer from the curse of horizon, a phenomenon previously

identified only for specific estimators. In contrast, the MDP efficiency bound, which we achieve, is generally polynomial in horizon.

We propose the Double Reinforcement Learning (DRL) estimator, which is given by by cross-fold estimation and plug-in of the q- and density ratio functions into the efficient influence function for each model, which we derive for the first time here. The name DRL is inspired by the Double Machine Learning estimation procedure of Chernozhukov et al. (2018), which we leverage, and by our simultaneous use of two learning procedures: learning of q-functions and of density ratios. We show that DRL achieves the semiparametric efficiency bound globally even when these nuisances are estimated at slow fourth-root rates and without restricting to Donsker or bounded entropy classes, enabling the use of flexible machine learning method for the nuisance estimation in the spirit of Chernozhukov et al. (2018); Zheng and van der Laan (2011). Further, we show that DRL is consistent even if only some of the nuisances are consistently estimated, known as double robustness. To the best of our knowledge, this is the first proposed estimator shown to be globally efficient for OPE in MDPs.

The organization of the paper is as follows. In Section 1.1, we define the OPE problem and our models. In Section 1.2 we summarize semiparametric inference theory and in Section 1.3 we review the literature on OPE. In Section 2, we calculate the efficient influence functions and efficiency bounds in each of our models. In Section 3, we propose the DRL estimator and prove its efficiency and robustness in each model, while also reviewing the inefficiency of other estimators. In Section 4, we discuss how to estimate q-functions in an off-policy manner to be used in DRL as well as the efficiency bound under parametric assumptions on the q-function. In Section 5, we demonstrate the benefits of DRL empirically.

A preliminary version of this work appeared as Kallus and Uehara (2020).

1.1. Problem Setup

A (potentially) non-Markov decision process (NMDP) is given by a sequence of state and action spaces $\mathcal{S}_t, \mathcal{A}_t$ for $t = 0, 1, \dots, T$, an initial state distribution $P_{s_0}(s_0)$, transition probabilities $P_{s_t}(s_t \mid \mathcal{H}_{a_{t-1}})$ for $t = 1, \dots, T$, and emission probabilities $P_{r_t}(r_t \mid \mathcal{H}_{a_t})$ for $t = 0, \dots, T$, where $\mathcal{H}_{a_t} = (s_0, a_0, \dots, s_t, a_t)$ is the state-action history up to a_t. A (non-anticipatory) policy is a sequence of action probabilities $\pi_t(a_t \mid \mathcal{H}_{s_t})$, where $\mathcal{H}_{s_t} = (s_0, a_0, \dots, a_{t-1}, s_t)$ is the state-action history up to s_t. Together, an NMDP and a policy define a joint distribution over trajectories $\mathcal{H} = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_T, a_T, r_T)$, given by the product $P_{s_0}(s_0)\pi_0(a_0 \mid \mathcal{H}_{s_0})P_{r_0}(r_0 \mid \mathcal{H}_{a_0})P_{s_1}(s_1 \mid \mathcal{H}_{a_0}) \cdots P_{r_T}(r_T \mid \mathcal{H}_{a_T})$. The dependence structure of such a distribution is illustrated in Fig. 1. We denote this distribution by P_π and expectations in this distribution by E_π to highlight the dependence on π.

A (time-varying) Markov decision process (MDP) is an NMDP where transitions and emissions depend only on the recent state and action and the time index t, $P_{s_t}(s_t \mid \mathcal{H}_{a_{t-1}}) = P_{s_t}(s_t \mid s_{t-1}, a_{t-1})$ and $P_{r_t}(r_t \mid \mathcal{H}_{a_t}) = P_{r_t}(r_t \mid s_t, a_t)$, and where we restrict to policies that depend only on the recent state, $\pi_t(a_t \mid \mathcal{H}_{s_t}) = \pi_t(a_t \mid s_t)$. MDPs have the important property that they are memoryless: given s_t, the trajectory starting at s_t is independent of the past trajectory, so that s_t fully captures the current state of the system. This imposes a stricter dependence structure, which is illustrated in Fig. 2. In particular, connections between variables with different time indices occurs only via s_t.

Our ultimate goal is to estimate the average cumulative reward of a policy,

$$\rho^\pi = \mathrm{E}_\pi \left[\sum_{t=0}^{T} r_t \right].$$

The quality and value functions (q- and v-functions) are defined as the following conditional averages of the cumulative reward to go, respectively:

$$q_t(\mathcal{H}_{a_t}) = \mathrm{E}_\pi \left[\sum_{k=t}^{T} r_k \mid \mathcal{H}_{a_t} \right], \qquad v_t(\mathcal{H}_{s_t}) = \mathrm{E}_\pi \left[\sum_{k=t}^{T} r_k \mid \mathcal{H}_{s_t} \right] = \mathrm{E}_\pi \left[q_t(\mathcal{H}_{a_t}) \mid \mathcal{H}_{s_t} \right].$$

Note that the very last expectation is taken only over $a_t \sim \pi_t(a_t \mid \mathcal{H}_{s_t})$. For MDPs, we have $q_t(\mathcal{H}_{a_t}) = q_t(s_t, a_t)$ and $v_t(\mathcal{H}_{s_t}) = v_t(s_t) = \mathrm{E}_\pi \left[q_t(s_t, a_t) \mid s_t \right]$, where again the last expectation is taken only over $a_t \sim \pi_t(a_t \mid s_t)$. For brevity, we define the random variables $q_t = q_t(\mathcal{H}_{a_t})$, $v_t = v_t(\mathcal{H}_{s_t})$.

The off-policy evaluation (OPE) problem is to estimate the average cumulative reward of a given (known) target evaluation policy, π^e, using n observations of trajectories $\mathcal{D} = \{\mathcal{H}^{(1)}, \dots, \mathcal{H}^{(n)}\}$ independently generated by the distribution P_{π^b} induced by using another policy, π^b, in the same decision process. This latter policy, π^b, is called the behavior policy and it may be known or unknown.

A model for the data generating process P_π of $\mathcal{D}$ is given by the set of products $P_{s_0}(s_0)\pi_0^b(a_0 \mid \mathcal{H}_{s_0})P_{r_0}(r_0 \mid \mathcal{H}_{a_0})P_{s_1}(s_1 \mid \mathcal{H}_{a_0}) \cdots P_{r_T}(r_T \mid \mathcal{H}_{a_T})$ over some possible values for each probability distribution in the product We let $\mathcal{M}_1$ denote the nonparametric model where each distribution is unknown and free. We let $\mathcal{M}_{1b}$ denote the submodel of $\mathcal{M}_1$ where π^b is known and fixed. We let $\mathcal{M}_{1q}$ denote any submodel of $\mathcal{M}_1$ where the functions $q_t(\mathcal{H}_{a_t})$ are restricted parametrically for $t = 0, \dots, T$. We let $\mathcal{M}_2$, $\mathcal{M}_{2b}$, $\mathcal{M}_{2q}$ denote the corresponding models where both the decision process and the behavior policy are restricted to be Markovian. Since π^e is given, the parameter of interest, ρ^{π^e}, is a function of just the part that specifies the decision process (initial state, transition, and emission probabilities).

To streamline notation, when no subscript is denoted, all expectations $\mathrm{E}[\cdot]$ and variances $\mathrm{var}[\cdot]$ are taken with respect to the behavior policy, π^b. At the same time, all v- and q-functions are for the target policy, π^e. The L^p-norm is defined as $\|g\|_p = \mathrm{E}[|g|^p]^{1/p}$. For any function of trajectories, we define its empirical average as

$$\mathrm{E}_n[f(\mathcal{H})] = n^{-1} \textstyle\sum_{i=1}^{n} f(\mathcal{H}^{(i)}).$$

We denote the density ratio at time t between the target and behavior policy by

$$\eta_t(\mathcal{H}_{a_t}) = \frac{\pi_t^e(a_t \mid \mathcal{H}_{s_t})}{\pi_t^b(a_t \mid \mathcal{H}_{s_t})}.$$

We denote the cumulative density ratio up to time t and the marginal density ratio at time t by, respectively,

$$\lambda_t(\mathcal{H}_{a_t}) = \prod_{k=0}^{t} \eta_t(\mathcal{H}_{a_k}), \qquad \mu_t(s_t, a_t) = \frac{p_{\pi_t^e}(s_t, a_t)}{p_{\pi_t^b}(s_t, a_t)},$$

where $p_{\pi_t}(s_t, a_t)$ denotes the *marginal* distribution of s_t, a_t under P_π. Note that under $\mathcal{M}_2$, $\eta_t(\mathcal{H}_{a_t}) = \eta_t(a_t, s_t)$. Again, for brevity we define the variables $\eta_t = \eta_t(\mathcal{H}_{a_t})$, $\lambda_t = \lambda_t(\mathcal{H}_{a_t})$, $\mu_t = \mu_t(s_t, a_t)$.

We will often assume the following:

Assumption 1 (Sequential overlap)**.** The density ratios η_t, μ_t satisfy $0 \leq \eta_t \leq C, 0 \leq \mu_t \leq C'$ for all $t = 0, \ldots, T$.

Assumption 2 (Bounded rewards)**.** The reward r_t satisfies $0 \leq r_t \leq R_{\max}$ for all $t = 0, \ldots, T$.

Assumption 1 requires that every action supported by the evaluation policy is also supported by the behavior policy, else the evaluation policy may induce state-action combinations that we cannot possibly ever see in the data. The assumption is standard in causal inference. Assumption 2 focuses on bounded rewards, which are common in reinforcement learning. Both assumptions can be relaxed to L_p-norm bounds on the above variables instead of boundedness (see Remark 9).

1.2. Summary of Semiparametric Inference

We briefly review semiparametric inference theory as it pertains to the relevance of our results. We provide a more complete review in Appendix B.1, while providing an accessible casual introduction here sufficient for the reader to understand the nature of our efficiency results. For a complete textbook presentation, we refer the reader to Bickel et al. (1998); Kosorok (2008); Tsiatis (2006); van der Laan and Robins (2003); van der Vaart (1998).

Suppose we have a model $\mathcal{M}$ for the generating process of the iid data $\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(n)}$, that is, a (potentially nonparametric) set of possible distributions for $\mathcal{H}^{(i)}$ that also contains the true distribution $F \in \mathcal{M}$ that generated the data. Consider a (scalar) parameter of interest $R : \mathcal{M} \to \mathbb{R}$. Given an estimator $\hat{R}$ (or rather a sequence of estimators), its limiting law is the distribution limit of $\sqrt{n}(\hat{R} - R(F))$, and the *asymptotic mean-squared error* (AMSE) is the second moment of the limiting law, which in turn lower bounds the scaled limit of the mean-squared error (MSE), $\lim n\mathrm{E}[(\hat{R} - R(F))^2]$, by the portmanteau lemma.

Every gradient of $R(\cdot)$ at $F \in \mathcal{M}$ (for paths in the model $\mathcal{M}$) is an F-measurable (scalar) random variable, that is, $\phi(\mathcal{H})$ with $\mathcal{H} \sim F$ for some function $\phi(\cdot)$. Each such function is called an *influence function*, and the influence function $\phi_{\text{eff}}(\cdot)$ with the smallest L^2 norm is is called the *efficient influence function* because

$$\text{EffBd}(\mathcal{M}) = \mathrm{E}_{\mathcal{H} \sim F}\left[\phi^2_{\text{eff}}(\mathcal{H})\right]$$

bounds below the AMSE of *any* estimator that is regular with respect to the model $\mathcal{M}$.[1] Regular estimators are roughly those that have risk that is invariant to vanishing perturbations to the data generating process F (that remain inside the model $\mathcal{M}$), which is a desirable property else the estimator may be unreasonably sensitive to undetectable

1. Note that $\text{EffBd}(\mathcal{M})$ depends on the estimand $R(\cdot)$, the model $\mathcal{M}$, and the instance F in that model. We emphasize foremost the dependence on the model to highlight the differences when we change the model from NMDP to MDP, while our estimand is always the target policy value.

changes.[2] Essentially, regular estimators with respect to $\mathcal{M}$ are those that would work for estimating $R(F)$ for *any* instance $F \in \mathcal{M}$. Thus, this lower bound applies *per-instance* for any estimator that, in a sense, works in the model $\mathcal{M}$. Note we have $\mathrm{EffBd}(\mathcal{M}') \leq \mathrm{EffBd}(\mathcal{M})$ whenever $F \in \mathcal{M}' \subset \mathcal{M}$, and that these may be different even though $F \in \mathcal{M}'$, as the set of estimators that work in $\mathcal{M}'$ is potentially larger. For example, the lower bound in the NMDP model is still larger than (or equal to) the bound in MDP model, even if considered at a specific instance that happens to be an MDP.

There are several further interpretations of this lower bound. By the portmanteau lemma, the lower bound on limiting law also means that $\mathrm{EffBd}(\mathcal{M})$ lower bounds the limit of the MSE for any regular $\hat{R}$, namely

$$\liminf_{n\to\infty} n\mathrm{E}[(\hat{R} - R(F))^2] \geq \mathrm{EffBd}(\mathcal{M}).$$

Moreover, standard results (*e.g.*, van der Vaart, 1998, Thm. 25.21) establish that the lower bound also applies to *all* estimators (not just regular ones) in a local minimax fashion: for *any* estimator, n times the worst-case MSE in a $1/\sqrt{n}$-sized $\mathcal{M}$-neighborhood around F has a limit infimum of at least $\mathrm{EffBd}(\mathcal{M})$. Here the ambient model $\mathcal{M}$ is relevant in determining the bound as the local worst-case neighborhoods are restricted to remain inside the model. When $\mathcal{M}$ is a fully parametric model the semiparametric efficiency bound is actually the same as the Cramér-Rao bound. In fact, the semiparametric efficiency bound corresponds to the supremum of the Cramér-Rao bounds over all regular parametric submodels $F \in \mathcal{M}_{\mathrm{para}} \subset \mathcal{M}$. Thus, it also describes the best-achievable behavior by nonparametric estimators that work in *every* parametric submodel.

In these senses, $\mathrm{EffBd}(\mathcal{M})$, known as the semiparametric efficiency bound, lower bounds the achievable MSE in estimating R on the model $\mathcal{M}$. If we can find an estimator whose limiting law has zero mean and variance $\mathrm{EffBd}(\mathcal{M})$ then it must have the smallest-possible (asymptotic) MSE, and such estimators are known as (asymptotically) *efficient*. Moreover, all efficient regular estimators must satisfy

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

that is, they are asymptotically linear with efficient influence function ϕ_{eff}. This suggests an estimation strategy: try to approximate $\hat{\psi}(\mathcal{H}) \approx \phi_{\mathrm{eff}}(\mathcal{H}) + R(F)$ and use $\hat{R} = \frac{1}{n}\sum_{i=1}^{n} \hat{\psi}(\mathcal{H}^{(i)})$. Done appropriately, this can provide an efficient estimate. Therefore, deriving the efficient influence function is important both for computing the semiparametric efficiency bound and for coming up with good estimators.

Note that the efficiency notion of optimality is different from (non-local, non-asymptotic) minimax optimality (*e.g.*, as used by Wang et al., 2017 for non-sequential OPE with $H = 0$), which considers the worst-case behavior against a large, fixed class of instances rather than considering behavior locally at F. While efficiency provides a much finer analysis of the precise behavior at the particular instance generating the data, it is also only asymptotic. Nonetheless, we will also establish finite-sample guarantees for our efficient estimators, where the leading terms will be controlled by $\mathrm{EffBd}(\mathcal{M})$.

2. See Appendix B.1 and van der Vaart, 1998, Ch. 25 for precise definitions of path derivatives and regular estimators.

1.3. Summary of Literature on OPE

OPE is a central problem in both RL and in closely related problems such as dynamic treatment regimes (DTR; Murphy et al., 2001). While the NMDP model $\mathcal{M}_1$ is commonly the one assumed in the causal inference literature in the context of marginal structural model estimation (Robins, 2000; Robins et al., 2000) and DTRs (Chakraborty and Moodie, 2013; Murphy et al., 2001; Zhang et al., 2013),[3] in RL one often assumes that the MDP model $\mathcal{M}_2$ holds. Nonetheless, with some exceptions that we review below, OPE methods in RL have largely not leveraged the additional independence structure of $\mathcal{M}_2$ to improve estimation, and in particular, the effect of this structure on efficiency has not previously been studied and no efficient evaluation method has been proposed.

Methods for OPE can be roughly categorized into three types. The first approach is the *direct method* (DM), wherein we directly estimate the q-function and use it to directly estimate the value of the target evaluation policy. For example, one can use model-based estimates (Mannor et al., 2007) or estimate the q-function directly using fitted LSTDQ or more general q-iteration (Antos et al., 2008; Lagoudakis and Parr, 2004; Le et al., 2019) (we further review estimation of q-functions in Section 4. Once we have an estimate $\hat{q}_0$ of the first q-function, the DM estimate is simply $\hat{\rho}^{\pi^e}_{\text{DM}} = \mathrm{E}_n\left[\mathrm{E}_{\pi^e}\left[\hat{q}_0(s_0, a_0) \mid s_0\right]\right]$, where the inner expectation is simply over $a_0 \sim \pi^e(\cdot \mid s_0)$ and is thus computable as a sum or integral over a known measure and the outer expectation is simply an average over the n observations of s_0. Recall we define all q-functions to be with respect to π^e. For DM, we can leverage the structure of $\mathcal{M}_2$ by simply restricting q-functions to be Markovian. However, DM can fail to be efficient even under $\mathcal{M}_1$ unless q-functions are parametric (and correctly specified) or extremely smooth (as shown by Hahn, 1998 but only in the $T = 0$ case). DM is also not robust in that, if q-functions are inconsistently estimated, the estimate will be inconsistent.

The second approach is *importance sampling* (IS), which averages the data weighted by the density ratio of the evaluation and behavior policies. Given estimates $\hat{\lambda}_t$ of the cumulative density ratios (or, letting $\hat{\lambda}_t = \lambda_t$ if the behavior policy is known), the IS estimate is simply $\hat{\rho}^{\pi^e}_{\text{IS}} = \mathrm{E}_n\left[\sum_{t=0}^T \hat{\lambda}_t r_t\right]$. (An alternative but higher-variance IS estimator is $\mathrm{E}_n\left[\hat{\lambda}_T \sum_{t=0}^T r_t\right]$.) When behavior policy is known, IS is unbiased and consistent, but its variance tends to be large due to extreme weights. In particular, under $\mathcal{M}_1$, IS with $\hat{\lambda}_t = \lambda_t$ is known to be inefficient (Hirano et al., 2003), which implies it must be inefficient under $\mathcal{M}_2$ as well. A common variant of IS is the self-normalized estimate $\sum_{t=0}^T \frac{\mathrm{E}_n\left[\hat{\lambda}_t r_t\right]}{\mathrm{E}_n\left[\hat{\lambda}_t\right]}$ (Swaminathan and Joachims, 2015), which trades off some bias for variance but does not make IS efficient.

The third approach is the *doubly robust* (DR) method, which combines DM and IS and is given by adding the estimated q-function as a control variate (Dudik et al., 2014; Jiang and Li, 2016; Scharfstein et al., 1999). The DR estimate has the form $\hat{\rho}^{\pi^e}_{\text{DR}} = \mathrm{E}_n\left[\sum_{t=0}^T \left(\hat{\lambda}_t(r_t - \hat{q}_t) + \hat{\lambda}_{t-1}\mathrm{E}_{\pi^e}\left[\hat{q}_t | s_t\right]\right)\right]$.

DR is colloquially known to be efficient under $\mathcal{M}_1$ but no precise result is available. When state and action spaces are finite, the model $\mathcal{M}_1$ is necessarily parametric, and,

3. OPE is equivalent to estimating the total treatment effect of a DTR in a causal inference setting. Although we do not explicitly use counterfactual notation (either potential outcomes or *do*-calculus), if we assume the usual sequential ignorability conditions (Ertefaie and Strawderman, 2018; Luckett et al., 2018; Murphy et al., 2001), the estimands we consider are the same and our results immediately apply.

under this parametric model, Jiang and Li (2016) study the Cramér-Rao lower bound and observe that an infeasible DR estimator that uses oracle nuisance values instead of estimates, $\hat{q}_t = q_t$ and $\hat{\lambda}_t = \lambda_t$, would achieve the bound. For completeness, we derive precisely the more general semiparametric efficiency bound under $\mathcal{M}_1$ (Theorem 1) and show that two (feasible) variants of the standard DR estimate are semiparametrically efficient, either using sample splitting with a rate condition (Theorem 6) or without sample splitting with a Donsker condition (Theorem 9). Jiang and Li (2016) also study parametric Cramér-Rao lower bounds under finite action and state space in the MDP model $\mathcal{M}_2$, but no efficient estimator, whether parametric or nonparametric, has been proposed. See also Remarks 2 and 5. There is a significant gap to deriving the semiparametric bound, which generalizes these results to more general action and state spaces and nonparametric models. More importantly, our derivation yields the efficient influence function, which provides a way to construct an efficient estimator under $\mathcal{M}_2$.

Many variations of DR have been proposed. Thomas and Brunskill (2016) propose both a self-normalized variant of DR and a variant blending DR with DM when density ratios are extreme. Farajtabar et al. (2018) propose to optimize the choice of $\hat{q}_t$ to minimize a variance estimate for DR rather than use a plug-in value. Kallus and Uehara (2019) propose a DR estimator that achieves local efficiency, has certain stability properties enjoyed by self-normalized IS, and at the same time is guaranteed to have asymptotic MSE that is never worse than both DR, IS, and self-normalized IS.

However, all of the aforementioned IS and DR estimators do not exploit MDP structure and, in particular, will *fail* to be efficient under $\mathcal{M}_2$. Recently, in the finite-state-space setting Xie et al. (2019) studied an IS-type estimator that exploits MDP structure by replacing density ratios with marginalized density ratios, estimated by a recursive formula. However, this estimator is also not efficient, even in the finite tabular setting. Remark 4 of Xie et al. (2019) points out the inefficiency of the estimator proposed therein.

2. Semiparametric Inference for Off-Policy Evaluation

In this section, we derive the efficiency bounds and efficient influence functions for ρ^{π^e} under the models $\mathcal{M}_1$, $\mathcal{M}_{1b}$, $\mathcal{M}_2$, and $\mathcal{M}_{2b}$. Recall that the former two models are NMDP and the latter two are MDP and that the second and fourth assume a known behavior policy.

2.1. Semiparametric Efficiency in Non-Markov Decision Processes

First, we consider the NMDP models $\mathcal{M}_1$ and $\mathcal{M}_{1b}$. We do this mostly for the sake of completeness since, while the influence function we derive below for the NMDP model appears as a central object in the structure of various previously proposed doubly-robust OPE estimators for RL (*e.g.*, among others, Dudik et al., 2014; Farajtabar et al., 2018; Jiang and Li, 2016; Kallus and Uehara, 2019; Thomas and Brunskill, 2016), these do not establish it as the efficient influence function in the NMDP model or derive the semiparametric efficiency bound, with the exception of the concurrent Bibaut et al. (2019). We note that in contrast, the influence function we derive for the MDP model in the next section appears to be novel and leads to new, more efficient estimators.

Theorem 1 (Efficiency bound under $\mathcal{M}_1$)**.** *The efficient influence function of ρ^{π^e} under $\mathcal{M}_1$ is*

$$\phi_{\text{eff}}^{\mathcal{M}_1}(\mathcal{H}) = -\rho^{\pi^e} + \sum_{t=0}^{T} \left(\lambda_t \left(r_t - q_t\right) + \lambda_{t-1} v_t\right). \tag{1}$$

The semiparametric efficiency bound under $\mathcal{M}_1$ is

$$\text{EffBd}(\mathcal{M}_1) = \text{var}(v_0) + \sum_{t=0}^{T} \text{E}\left[\lambda_t^2 \text{var}\left(r_t + v_{t+1} \mid \mathcal{H}_{a_t}\right)\right], \tag{2}$$

where $v_{T+1} = 0$.

Under $\mathcal{M}_{1b}$, the efficient influence function and bound are the same.

Note that we do not assume Assumptions 1 and 2 in the above. The quantity $\text{EffBd}(\mathcal{M}_1)$ may or may not be finite. An infinite efficiency bound would imply the impossibility of consistent $\sqrt{n}$ estimation. Below in Theorem 5 we show how to bound $\text{EffBd}(\mathcal{M}_1)$ under Assumptions 1 and 2.

Remark 1. The efficient influence function and bound are both the same whether we know the behavior policy or not. Intuitively, this happens because the estimand ρ^{π_e} does not in fact depend on behavior policy part of the data generating distribution, P_{π^b}, but only on the decision process parts (initial state, transition, and emission probabilities). This phenomenon mirrors the situation with knowledge of the propensity score in average treatment effect estimation in causal inference noted by Hahn (1998).

Remark 2. When the action and state spaces are discrete, $\mathcal{M}_1$ is necessarily a parametric model. In this discrete-space parametric model and with $r_t = 0$ for $t \leq T-1$, Theorem 2 of Jiang and Li (2016) derives the Cramér-Rao lower bound for estimating ρ^{π^e}. Because the semiparametric efficiency bound is the same as the Cramér-Rao lower bound for parametric models, the bound coincides with ours in this special discrete setting. Theorem 1 and the related result in Bibaut et al. (2019) are more general, establishing the limit on estimation in non-discrete, nonparametric settings and, moreover, establishes that the efficient influence function coincides with the structure of many doubly-robust OPE estimators used in RL (see references above).

Remark 3. The efficient influence function $\phi_{\text{eff}}^{\mathcal{M}_1}$ has the oft-noted doubly robust structure. Specifically,

$$\begin{aligned}
\rho^{\pi^e} + \text{E}\left[\phi_{\text{eff}}^{\mathcal{M}_1}(\mathcal{H})\right] &= \underbrace{\text{E}\left[\sum_{t=0}^{T} \lambda_t r_t\right]}_{=\rho^{\pi^e}} + \underbrace{\text{E}\left[\sum_{t=0}^{T} \left(-\lambda_t q_t + \lambda_{t-1} v_t\right)\right]}_{=0} \\
&= \underbrace{\text{E}\left[v_0\right]}_{=\rho^{\pi^e}} + \underbrace{\text{E}\left[\sum_{t=0}^{T} \lambda_t (r_t - q_t + v_{t+1})\right]}_{=0}.
\end{aligned}$$

The first term in each line corresponds to a sequential IS estimator and direct method (DM), respectively. The second term in each line is a control variate, which remain mean zero even if we plug in different (*i.e.*, wrong) q- and v-functions or density ratios, respectively. In this sense, it is sufficient to estimate only one part of these for consistent OPE. We will leverage this in Theorem 10 to achieve double robustness for DRL.

Remark 4 (Dependence on Rewards and the Completely Unrestricted Model)**.** Our definition of NMDP does not allow for transitions, rewards, and policies to depend on past rewards, as is indeed standard in RL. Nonetheless, we can easily also consider the alternative (and larger) model where we do allow these dependences, $\mathcal{M}_{1r}$. Defining $\mathcal{J}_{a_j} = (s_0, a_0, r_0, \cdots, r_{j-1}, s_j, a_j)$, and similarly $\mathcal{J}_{r_j}$ and $\mathcal{J}_{s_j}$, the joint distribution over trajectories $\mathcal{H} = (s_0, \dots, r_T)$ in this model is given by the product $P_{s_0}(s_0)\pi_0(a_0 \mid \mathcal{J}_{s_0})P_{r_0}(r_0 \mid \mathcal{J}_{a_0})P_{s_1}(s_1 \mid \mathcal{J}_{r_0})\pi_1(a_0 \mid \mathcal{J}_{s_1}) \cdots P_{r_T}(r_T \mid \mathcal{J}_{a_T})$, where each density is free and unrestricted. This in fact means that the model includes *any* joint density on the trajectory $\mathcal{H}$ and is completely unrestricted because such a sequential factorization can *always* be generically done to any density, that is, the model $\mathcal{M}_{1r}$ is the model containing *all* joint densities over the trajectory $\mathcal{H}$. Redefining q- and v-functions in this model as $q_t(\mathcal{J}_{a_t}) = \mathrm{E}[\sum_{j=t}^{T} r_j \mid \mathcal{J}_{a_t}]$ and $v_t(\mathcal{J}_{s_t}) = \mathrm{E}[\sum_{j=t}^{T} r_j \mid \mathcal{J}_{s_t}]$, respectively, we can compute the efficient influence function and efficiency bound in this model in a similar way to Theorem 1.

Theorem 2 (Efficiency bounder under $\mathcal{M}_{1r}$)**.** *The efficient influence function of $\rho^{\pi^{\mathrm{e}}}$ under $\mathcal{M}_{1r}$ is*

$$-\rho^{\pi^{\mathrm{e}}} + \sum_{t=0}^{T} \left(\lambda_t(r_t - q_t) + \lambda_{t-1} v_t\right).$$

The semiparametric efficiency bound under $\mathcal{M}_{1r}$ is

$$\mathrm{var}(v_0) + \sum_{t=0}^{T} \mathrm{E}[\lambda_t^2 \mathrm{var}(r_t + v_{t+1} | \mathcal{J}_{a_t})].$$

2.2. Semiparametric Efficiency in Markov Decision Processes

Next, we derive the efficiency bound and efficient influence function for ρ^{π^e} under the models $\mathcal{M}_2$ and $\mathcal{M}_{2b}$, *i.e.*, when restricting to MDP structure. To our knowledge, not only have these never before been derived, the influence function we derive has also not appeared in any existing OPE estimators in RL. We recall that under $\mathcal{M}_2$, we have $q_t = q_t(s_t, a_t)$ and $v_t = v_t(s_t)$.

Theorem 3 (Efficiency bound under $\mathcal{M}_2$)**.** *The efficient influence function of ρ^{π^e} under $\mathcal{M}_2$ is*

$$\phi_{\mathrm{eff}}^{\mathcal{M}_2}(\mathcal{H}) = -\rho^{\pi^e} + \sum_{t=0}^{T} \left(\mu_t \left(r_t - q_t\right) + \mu_{t-1} v_t\right). \tag{3}$$

The semiparametric efficiency bound under $\mathcal{M}_2$ is

$$\mathrm{EffBd}(\mathcal{M}_2) = \mathrm{var}(v_0) + \sum_{t=0}^{T} \mathrm{E}\left[\mu_t^2 \mathrm{var}\left(r_t + v_{t+1} \mid s_t, a_t\right)\right], \tag{4}$$

where $v_{T+1} = 0$.

Under $\mathcal{M}_{2b}$*, the efficient influence function and bound are the same.*

Again, we do not assume Assumptions 1 and 2 in the above. Below in Theorem 5 we show how to bound $\text{EffBd}(\mathcal{M}_2)$ under Assumptions 1 and 2.

Remark 5. Again, when the action and state spaces are discrete, $\mathcal{M}_2$ is necessarily a parametric model. In this discrete-space parametric model and with $r_t = 0$ for $t \leq T-1$, Theorem 3 of Jiang and Li (2016) derives the Cramér-Rao lower bound, which must (and does) coincide with ours in this setting. Again, our result is more general, covering nonparametric models and estimators, and, importantly, derives the efficient influence function, which we will use to construct the first globally efficient estimator for ρ^{π^e} under $\mathcal{M}_2$.

Remark 6. The difference between the efficient influence functions in the NMDP and MDP models, $\phi_{\text{eff}}^{\mathcal{M}_1}$ and $\phi_{\text{eff}}^{\mathcal{M}_2}$, is that (a) the cumulative density ratio λ_t is replaced with the marginalized density ratio μ_t and (b) that q- and v-functions only depend on recent state and action rather than full past trajectory. Note that the latter difference is slightly hidden in our notation: in $\phi_{\text{eff}}^{\mathcal{M}_1}$, q_t refers to $q_t(\mathcal{H}_{a_t})$, while in $\phi_{\text{eff}}^{\mathcal{M}_2}$, q_t refers to the much simpler $q_t(s_t, a_t)$.

Although the efficient influence function in Theorem 3 is derived *de-novo* in the proof, which is the most direct route to a rigorous derivation, we can also use the geometry of influence functions to understand the result relative to Theorem 1. The efficient influence function is always given by projecting the influence function of any regular asymptotic linear estimator onto the tangent space (Tsiatis, 2006, Thm. 4.3). Under $\mathcal{M}_2$, the function $\phi_{\text{eff}}^{\mathcal{M}_1}(\mathcal{H})$ from Theorem 1 can be shown to still be an influence function of some regular asymptotic linear estimator in $\mathcal{M}_2$. Projecting it onto the tangent space in $\mathcal{M}_2$, where we have imposed the independence of past and future trajectories given intermediate state, can be seen to exactly correspond to the above marginalization over the past trajectory, explaining this structure of $\phi_{\text{eff}}^{\mathcal{M}_2}(\mathcal{H})$.

Remark 7. The efficient influence function $\phi_{\text{eff}}^{\mathcal{M}_2}(\mathcal{H})$ also has a doubly robust structure. Specifically,

$$\begin{aligned}\rho^{\pi^e} + \mathrm{E}\left[\phi_{\text{eff}}^{\mathcal{M}_2}(\mathcal{H})\right] &= \underbrace{\mathrm{E}\left[\sum_{t=0}^{T} \mu_t r_t\right]}_{=\rho^{\pi^e}} + \underbrace{\mathrm{E}\left[\sum_{t=0}^{T} (-\mu_t q_t + \mu_{t-1} v_t)\right]}_{=0} \\ &= \underbrace{\mathrm{E}\left[v_0\right]}_{=\rho^{\pi^e}} + \underbrace{\mathrm{E}\left[\sum_{t=0}^{T} \mu_t (r_t - q_t + v_{t+1})\right]}_{=0}.\end{aligned}$$

The first term on the first line corresponds to the Marginalized Importance Sampling (MIS) estimator (Xie et al., 2019). The first term on the second line corresponds to the DM estimator. The second term on each line corresponds to control variate terms. We will leverage this in Theorem 16 to achieve double robustness for DRL.

By comparing the efficiency bounds of Theorem 1 and Theorem 3 and using Jensen's inequality, we can see that the Markov assumption reduces the efficiency bound, usually strictly so.

Theorem 4. *If $P_{\pi^b} \in \mathcal{M}_2$ (i.e., the underlying distribution is an MDP), then*

$$\mathrm{EffBd}(\mathcal{M}_2) \leq \mathrm{EffBd}(\mathcal{M}_1).$$

Moreover, the inequality is strict if there exists $t \leq T$ such that both λ_{t-1} and $r_{t-1} + v_t$ are not constant given s_{t-1}, a_{t-1}.

Beyond being sorted, we can actually see that EffBd(MDP) is generally polynomial in T while EffBd(NMDP) is generally exponential in T. This shows that the curse of horizon is *inevitable* in NMDP. While previously it was just shown to be a limitation specifically of IS and DR estimators (Liu et al., 2018; Xie et al., 2019), this result shows it *must* plague any estimator that targets the NMDP model and that it is insurmountable without leveraging additional structure that further narrows the model. That EffBd(MDP) is generally polynomial in T shows that we can potentially overcome this by efficiently leveraging MDP structure, which is exactly what our novel DRL estimator will do.

Theorem 5. *Under Assumptions 1 and 2,*

$$\mathrm{EffBd}(\mathrm{MDP}) \leq C' R_{\max}^2 (T+1)^2,$$
$$\mathrm{EffBd}(\mathrm{NMDP}) \leq C^{T+1} R_{\max}^2 (T+1)^2.$$

If $\mathrm{E}_{\pi^e}[\log(\eta_t)] \geq C_{\min}$ and $\mathrm{E}_{\pi^e}\left[\log(\mathrm{var}(r_t + v_{t+1} \mid \mathcal{H}_{a_t}))\right] \geq \log(V_{\min}^2)$ then

$$\mathrm{EffBd}(\mathrm{NMDP}) \geq C_{\min}^{T+1} V_{\min}^2.$$

Note that $\mathrm{E}_{\pi^e}[\log(\eta_t)] = \mathrm{E}_{\pi^e}\left[\mathrm{KL}(\pi_t^e(\cdot \mid s_t) \,||\, \pi_t^b(\cdot \mid s_t))\right]$ is the KL divergence between the distributions over actions induced by π^e and π^b, averaged over the states visited by π^e, and that $\mathrm{E}_{\pi^e}\left[\log(\mathrm{var}(r_t + v_{t+1} \mid \mathcal{H}_{a_t}))\right] = \mathrm{E}\left[\log(\mathrm{var}(r_t + v_{t+1} \mid \mathcal{H}_{a_t}))\right]$.

The lower bound on EffBd(NMDP) shows that the curse of horizon is inevitable. The condition on η_t simply means that the evaluation and behavior policies are not becoming arbitrarily similar as t grows (on-policy evaluation does not suffer from curse of horizon). The condition on $r_t + v_{t+1}$ essentially ensures that rewards are not trivially constant. The upper bound on EffBd(MDP) shows that, in contrast, variance that is polynomial in T is possible in the MDP model.

Remark 8 (Consistency of EffBd(NMDP) and EffBd(NMDP))**.** NMDP models may be trivially transformed into MDP models by letting the state variable be the whole history $\mathcal{H}_{s_t}$. Then, the trajectory becomes $\{\mathcal{H}_{s_0}, a_0, r_0, \mathcal{H}_{s_1}, a_1, r_1, \cdots\}$ and the efficiency bound under this transformed MDP matches the efficiency bound under the original NMDP since

$$\mu_t(\mathcal{H}_{s_t}, a_t) = \frac{p_{\pi_t^e}(\mathcal{H}_{s_t}, a_t)}{p_{\pi_t^b}(\mathcal{H}_{s_t}, a_t)} = \prod_{k=0}^{t} \eta_k(\mathcal{H}_{a_k}) = \lambda_t(\mathcal{H}_{a_t}).$$

3. Efficient Estimation Using Double Reinforcement Learning

In this section, we construct the DRL estimator and then study its properties in the various models. In particular, we show that DRL is globally efficient under very mild assumptions. In the NMDP model, these assumptions are generally weaker than needed for efficiency of previous estimators. In the MDP model, this provides the first globally efficiency estimator for OPE. We further show that DRL enjoys certain double robustness properties when some nuisances are inconsistently estimated.

DRL is a meta-estimator; it takes in as input estimators for q-functions and density ratios and combines them in a particular manner that ensures efficiency even when the input estimators may not be well behaved. This is achieved by following the cross-fold sample-splitting strategy developed by Chernozhukov et al. (2018); Klaassen (1987); Zheng and van der Laan (2011). We proceed by presenting DRL and its properties in each setting (NMDP and MDP). In the NMDP setting, DRL amounts to the cross-fold version of the RL OPE doubly robust estimator, which was proposed in the experiments of Jiang and Li (2016, Section 6.1) but not analyzed. In the MDP setting, DRL is the first semiparametrically efficient and doubly robust estimator.

Throughout this section we assume that Assumptions 1 and 2 hold.

3.1. Double Reinforcement Learning for NMDPs

Given a learning algorithm to estimate the q-function $q(\mathcal{H}_{a_t})$ and cumulative density ratio function $\lambda_t(\mathcal{H}_{a_t})$, DRL with K-fold sample splitting ($K \geq 2$) for NMDPs proceeds as follows:

1. Randomly permute the data indices and let $\mathcal{D}_j = \{\lceil (j-1)n/K \rceil + 1, \ldots, \lceil jn/K \rceil\}$ for $j = 1, \ldots, K$. Let j_i be the fold containing observation i so that $i \in \mathcal{D}_{j_i}$ (namely, $j_i = 1 + \lfloor (i-1)K/n \rfloor$).

2. For $j = 1, \ldots, K$, construct estimators $\hat{\lambda}_t^{(j)}(\mathcal{H}_{a_t})$ and $\hat{q}_t^{(j)}(\mathcal{H}_{a_t})$ based on the training data given by all trajectories excluding those in $\mathcal{D}_j$, that is, $\{1, \ldots, n\} \backslash \mathcal{D}_j$.

3. Let

$$\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T} \Bigg(\hat{\lambda}_t^{(j_i)}(\mathcal{H}_{a_t}^{(i)}) \left(r_t^{(i)} - \hat{q}_t^{(j_i)}(\mathcal{H}_{a_t}^{(i)}) \right) + \hat{\lambda}_{t-1}^{(j_i)}(\mathcal{H}_{a_{t-1}}^{(i)}) \int_{a'_t} \hat{q}_t^{(j_i)}((\mathcal{H}_{s_t}^{(i)}, a'_t)) d\pi_t^e(a'_t \mid \mathcal{H}_{s_t}^{(i)}) \Bigg).$$

   Here $(\mathcal{H}_{s_t}^{(i)}, a'_t)$ represents the trajectory given by appending a'_t to $\mathcal{H}_{s_t}^{(i)}$; note this differs from $\mathcal{H}_{a_t}^{(i)}$. Note further that the integral becomes a simple sum when π^e has finite support over actions (*e.g.*, if π^{e} is deterministic or if there are finitely many actions).

In other words, we approximate the efficient influence function $\phi_{\mathrm{eff}}^{\mathcal{M}_1}(\mathcal{H}) + \rho^{\pi^e}$ from Theorem 1 by replacing the unknown q- and density ratio functions with estimates thereof and we take empirical averages of this approximation over the data, where for each data point we use q- and density ratio function estimates based only on the half-sample that does *not* contain the data point. While $K = 2$ is sufficient to achieve efficiency, larger K allows

nuisances to be fit on more data and may prove practically successful. Note also that we take only a single random K-fold split and that is enough to achieve our results below. In practice, repeating the above process over several splits of the data and taking the average of resulting DRL estimates can only reduce the variance without increasing bias.

This estimator has several desirable properties. To state them, we assume the following conditions for the estimators, reflecting Assumptions 1 and 2:

Assumption 3 (Bounded estimators)**.** $\limsup_{n\to\infty} \|\hat{\lambda}_t^{(j)}\|_\infty < \infty$, $\limsup_{n\to\infty} \|\hat{q}_t^{(j)}\|_\infty < \infty$ for each $0 \le t \le T$, $1 \le j \le K$.

Assumption 3 provides that the estimators are eventually almost surely bounded. This to be expected when their target estimands are bounded, and, in particular, will necessarily be the case for many non-parametric estimators such as random forests and kernel regression. And, per Assumptions 1 and 2, we have $\|\lambda_t^{(j)}\|_\infty \le C^{t+1}$ and $\|q_t^{(j)}\|_\infty < (T+1-t)R_{\max}$. Note, however, we need not assume the same bound applies to the estimates; any finite bound (independent of n) suffices. For the rest of this subsection we will assume Assumption 3 hold.

We first prove that DRL achieves the semiparametric efficiency bound, even if each nuisance estimator has a slow, nonparametric convergence rate (sub-$\sqrt{n}$).

Theorem 6 (Efficiency of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ under $\mathcal{M}_1$)**.** *Suppose* $\|\hat{\lambda}_t^{(j)} - \lambda_t\|_2 \|\hat{q}_t^{(j)} - q_t\|_2 = \mathrm{o}_p(n^{-1/2})$, $\|\hat{\lambda}_t^{(j)} - \lambda_t\|_2 = \mathrm{o}_p(1)$, $\|\hat{q}_t^{(j)} - q_t\|_2 = \mathrm{o}_p(1)$ *for* $0 \le t \le T, 1 \le j \le K$. *Then, the estimator* $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ *achieves the semiparametric efficiency bound under* $\mathcal{M}_1$.

Remark 9. Assumptions 1 to 3 posited L^∞ bounds on density ratios, rewards, and their estimates. These assumptions are standard in both reinforcement learning and causal inference. It is possible to relax these to L^p bounds at the cost of requiring stronger convergence on nuisance estimates above, requiring $L^{2/(1-1/p)}$ convergence instead of the L^2 convergence above. Since L^2 convergence in estimation is usually the standard convergence mode considered and standard results can be invoked to ensure such rates, and similarly L^∞ bounds on density ratios and rewards are also standard, we focus our analysis on this most common case of the assumptions in order to avoid a cumbersome presentation.

Remark 10. There are two important points to make about this result. First, we have not assumed a Donsker condition (van der Vaart, 1998) on the class of estimators $\hat{\lambda}_t$ and $\hat{q}_t$. This is why this type of sample splitting estimator is called a double machine learning: the only required condition is a convergence rate condition at a nonparametric rate, allowing the use of complex machine learning estimators, for which one cannot verify the Donsker condition (Chernozhukov et al., 2018). In fact, many adaptive or high-dimensional estimators fail to satisfy Donsker conditions (Díaz, 2019). Eschewing such conditions allows us to use such estimators as the highly adaptive LASSO (Benkeser and van der Laan, 2016), càdlàg function estimators in very high dimensions (Bibaut and van der Laan, 2019b), and random forests (Wager and Walther, 2016) as long as their convergence rates are ensured under certain conditions. Benkeser and van der Laan (2016); Bibaut and van der Laan (2019b) in particular establish $\mathrm{o}_p(n^{-1/4})$ rates, which are compatible with our assumptions. Second, relative to the efficient influence function, which is defined in terms of the true q-function and cumulative density ratio, there is no inflation in DRL's asymptotic variance due to plugging

in *estimated* nuisance functions. This is due to the doubly robust structure of efficient influence function so that the estimation errors multiply and drop out of the first-order variance terms. This is in contrast to inefficient importance sampling estimators, as we will see in Theorem 20.

In addition to efficiency, we can also establish finite-sample guarantees for DRL, where the leading term is controlled by the efficient variance.

Theorem 7. *Suppose that for some C_1, C_2, for every n, with probability at least $1-\delta$, we simultaneously have that $\|\hat{\lambda}_t^{(j)} - \lambda_t\|_2^2 \leq \kappa_1 = C_1(n^{-2\alpha_1} + \log(2KT/\delta)/n)$, $\|\hat{q}_t^{(j)} - q_t\|_2^2 \leq \kappa_2 = C_2(n^{-2\alpha_2} + \log(2KT/\delta)/n)$ $\forall t \leq T$, $\forall j \leq K$. Suppose moreover that $0 \leq \hat{\lambda}_t^{(j)} \leq C^t$, $0 \leq \hat{q}_t^{(j)} \leq (T+1-t)R_{\max}$ for $\forall t \leq T$. Then, for every n, with probability at least $1-7\delta$, we have*

$$\begin{aligned}\left|\hat{\rho}^{\pi_e}_{\mathrm{DRL}(\mathcal{M}_1)} - \rho^{\pi_e}\right| \leq & \sqrt{\frac{2\log(2/\delta)\mathrm{Effbd}(\mathcal{M}_1)}{n}} \\ & + Q_1\sqrt{\frac{\log(2/\delta)T^2(TR_{\max}C^{T+1}\sqrt{\kappa_1\kappa_2} + \kappa_1 T^2 R_{\max}^2 + \kappa_2 C^{2(T+1)})}{n}} \\ & + Q_2\frac{\log(2/\delta)TR_{\max}C^{T+1}}{n} + Q_3 T\sqrt{\kappa_1\kappa_2},\end{aligned}$$

where Q_1, Q_2, Q_3 are constants not depending on $\delta, T, R_{\max}, C, n, C_1, C_2$.

Notice that, if $\alpha_1 > 0, \alpha_2 > 0, \alpha_1 + \alpha_2 > 1/2$ as in Theorem 6, then the leading term in the above is exactly $\sqrt{\frac{2\log(2/\delta)\mathrm{Effbd}(\mathcal{M}_1)}{n}}$ (with no additional constant factor), while the other terms are of strictly smaller order in n.

The rate assumptions in Theorem 7 are standard finite-sample estimation guarantees. For example, if estimators for nuisances are obtained by empirical risk minimization methods based on L^2-loss, then the results of Bartlett et al. (2005) apply. In this cases, the rates α_1 and α_2 would be determined by the local Rademacher complexity of the posited function classes. The number $2KT$ in $\log(2KT/\delta)$ comes from the fact that there are $2KT$ nuisance estimators. The boundedness assumptions on the estimates reflect the bounds on the estimands, per Assumptions 1 and 2.

Often in RL, the behavior policy is known and need not be estimated. That is, we can let $\hat{\lambda}_t^{(j)} = \lambda$. In this case, as an immediate corollary of Theorem 6, we have a much weaker condition for semiparametric efficiency: just that we estimate the q-function consistently, *without a rate.*

Corollary 8 (Efficiency of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ under $\mathcal{M}_{1b}$)**.** *Suppose $\hat{\lambda}_t = \lambda_t$ and $\|\hat{q}_t^{(j)} - q_t\|_2 = \mathrm{o}_p(1)$ for $0 \leq t \leq T, 1 \leq j \leq K$. Then, the estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ achieves the semiparametric efficiency bound under $\mathcal{M}_{1b}$.*

Without sample splitting, we have to assume a Donsker condition for the class of estimators in order to control a stochastic equicontinutiy term (see, *e.g.*, van der Vaart, 1998, Lemma 19.24). Although this is more restrictive, for completeness, we also include a theorem establishing the semiparametric efficiency of the standard plug-in doubly robust estimator for

NMDPs (Jiang and Li, 2016) when assuming the Donsker condition for in-sample-estimated nuisance functions, since this result was never precisely established before. We do not recommend this estimator due to its restrictive requirements on nuisance estimators.

Theorem 9 (Efficiency without sample splitting)**.** *Let $\hat{\lambda}_t, \hat{q}_t$ be estimators based on $\mathcal{D}$ and let $\hat{\rho}^{\pi^e}_{\mathrm{DR}} = \mathrm{E}_n\left[\sum_{t=0}^{T}\left(\hat{\lambda}_t\left(r_t - \hat{q}_t\right) - \hat{\lambda}_{t-1}\mathrm{E}_{\pi^e}\left[\hat{q}_t \mid \mathcal{H}_{s_t}\right]\right)\right]$. Suppose $\|\hat{\lambda}_t - \lambda_t\|_2\|\hat{q}_t - q_t\|_2 = \mathrm{o}_p(n^{-1/2}), \|\hat{\lambda}_t - \lambda_t\|_2 = \mathrm{o}_p(1)$, $\|\hat{q}_t - q_t\|_2 = \mathrm{o}_p(1)$ for $0 \le t \le T$ and that $\hat{q}_t$, $\hat{\lambda}_t$ belong to a Donsker class. Then, the estimator $\hat{\rho}^{\pi^e}_{\mathrm{DR}}$ achieves the semiparametric efficiency bound under $\mathcal{M}_1$.*

Thus, in $\mathcal{M}_1$, in comparison to the standard doubly robust estimator, DRL enjoys efficiency under milder conditions. To our knowledge, Theorems 6 and 9 are the first results precisely showing semiparametric efficiency for any OPE estimator.

In addition to efficiency, DRL enjoys a double robustness guarantee (Rotnitzky and Vansteelandt, 2014; Rotnitzky et al., 2019). Specifically, if at least just one model is correctly specified, then the DRL is estimator is still $\sqrt{n}$-consistent.

Theorem 10 (Double robustness ($\sqrt{n}$-consistency))**.** *Suppose $\|\hat{\lambda}_t^{(j)} - \lambda_t^\dagger\|_2 = \mathrm{O}_p(n^{-\alpha_1})$ and $\|\hat{q}_t^{(j)} - q_t^\dagger\|_2 = \mathrm{O}_p(n^{-\alpha_2})$ for $0 \le t \le T, 1 \le j \le K$. If, for each $0 \le t \le T$, either $\lambda_t^\dagger = \lambda_t$ and $\alpha_1 \ge 1/2$, $\alpha_2 > 0$ or $q_t^\dagger = q_t$ and $\alpha_2 \ge 1/2$, $\alpha_1 > 0$, then the estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ is $\sqrt{n}$-consistent around ρ^{π^e}.*

In particular, if the behavior policy is known so that $\hat{\lambda}_t^{(j)} = \lambda_t$, we can always ensure the estimator is $\sqrt{n}$-consistent (an example is the IS estimator, which has $\hat{q}_t^{(j)} = q_t^\dagger = 0$).

For consistency without a rate, it is sufficient for one nuisance to be consistent without a rate.

Corollary 11 (Double robustness (Consistency))**.** *Suppose $\|\hat{\lambda}_t^{(j)} - \lambda_t^\dagger\|_2 = \mathrm{o}_p(1)$ and $\|\hat{q}_t^{(j)} - q_t^\dagger\|_2 = \mathrm{o}_p(1)$ for $0 \le t \le T, 1 \le j \le K$. If, for each $0 \le t \le T$, either $\lambda_t^\dagger = \lambda_t$ or $q_t^\dagger = q_t$, then the estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ is consistent around ρ^{π^e}.*

A remaining question is when can we get nonparametric estimators achieving the necessary rates for the q- and density ratio functions. We discuss estimating q-functions in Section 4. Regarding the density ratio, λ_k, if the behavior policy is known then the density ratio is known and if the behavior policy is unknown it must be estimated. *Any* estimator satisfying the slow rate conditions would suffice.

For example, we may let $\hat{\lambda}_k^{(j)} = \prod_{t=0}^{k} \pi_t^e / \hat{\pi}_t^{b,(j)}$, where $\hat{\pi}_t^{b,(j)}$ is some nonparametric regression estimator. Then $\hat{\lambda}_k^{(j)}$ would enjoy the same rates as $\hat{\pi}_t^{b,(j)}$:

Lemma 12. *Suppose $\hat{\pi}_t^{b,(j)}$ and π_t^b are uniformly bounded by some constant below and that $\|\hat{\pi}_t^{b,(j)} - \pi_t^b\|_2 = \mathrm{o}_p(n^{-\alpha})$. Then, $\|\hat{\lambda}_t^{(j)} - \lambda_t\|_2 = \mathrm{o}_p(n^{-\alpha})$.*

Rates for $\hat{\pi}^{b,(j)}$ can be obtained from standard results for nonparametric regression, such as for kernel and sieve estimators (Newey and Mcfadden, 1994; Stone, 1994) or nonparametric estimators suited for high dimensions and non-smooth models (Bibaut and van der Laan, 2019a; Imaizumi and Fukumizu, 2018; Khosravi et al., 2019).

Alternatively, parametric models can be used for q_t and (if behavior policy is unknown) λ_t. Then, under standard regularity conditions, using MLE and other parametric regression estimators for behavior policy would yield $\|\hat{\lambda}_t^{(j)} - \lambda_t^\dagger\|_2 = \mathrm{O}_p(n^{-1/2})$, where $\lambda_t^\dagger = \lambda_t$ if the model is well-specified. Similarly, in Section 4, we discuss how using parametric q-models yields $\|\hat{q}_t^{(j)} - q_t^\dagger\|_2 = \mathrm{O}_p(n^{-1/2})$. If both models are correctly specified then Theorem 6 immediately implies DRL achieves the efficiency bound. When using parametric models, this is sometimes termed *local efficiency* (*i.e.*, local to the specific parametric model). If only one model is correctly specified then Theorem 10 ensures the estimator is still $\sqrt{n}$-consistent.

3.2. Double Reinforcement Learning for MDPs

Leveraging our derivation of the efficient influence function for $\mathcal{M}_2$ in Theorem 3, we can similarly construct our DRL estimator for MDPs. Given a learning algorithm to estimate the q-function $q_t(s_t, a_t)$ and marginal density ratio function $\mu_t(s_t, a_t)$, DRL with K-fold sample splitting ($K \geq 2$) for MDPs proceeds as follows:

1. Randomly permute the data indices and let $\mathcal{D}_j = \{\lceil (j-1)n/K \rceil + 1, \ldots, \lceil jn/K \rceil\}$ for $j = 1, \ldots, K$. Let j_i be the fold containing observation i so that $i \in \mathcal{D}_{j_i}$ (namely, $j_i = 1 + \lfloor (i-1)K/n \rfloor$).

2. For $j = 1, \ldots, K$, construct estimators $\hat{\mu}_t^{(j)}(s_t, a_t)$ and $\hat{q}_t^{(j)}(s_t, a_t)$ based on the training data given by all trajectories excluding those in $\mathcal{D}_j$, that is, $\{1, \ldots, n\} \backslash \mathcal{D}_j$.

3. Let

$$\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_2)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T} \Bigg(\hat{\mu}_t^{(j_i)}(s_t{}^{(i)}, a_t{}^{(i)}) \left(r_t^{(i)} - \hat{q}_t^{(j_i)}(s_t{}^{(i)}, a_t{}^{(i)}) \right) + \hat{\mu}_{t-1}^{(j_i)}(s_{t-1}{}^{(i)}, a_{t-1}{}^{(i)}) \int_{a'_t} \hat{q}_t^{(j_i)}(s_t{}^{(i)}, a'_t) d\pi_t^e(a'_t \mid s_t{}^{(i)}) \Bigg).$$

   Again, note the difference between the dummy a'_t and the data $a_t^{(i)}$, and that the integral becomes a simple sum when π^e has finite support over actions (*e.g.*, if π^{e} is deterministic or if there are finitely many actions).

Again, what we have done is approximate the efficient influence function $\phi_{\mathrm{eff}}^{\mathcal{M}_2}(\mathcal{H}) + \rho^{\pi^e}$ from Theorem 3 and taken its empirical average over the data, where for each data point we use q- and marginal density ratio function estimates based only on the half-sample that does *not* contain the data point. Again, taking one split suffices for our results, but one can repeat the above over many splits and take averages without deterioration.

Since both estimators are approximating their respective influence function as we derive in Theorems 1 and 3, the differences between $\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_1)}$ and $\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_2)}$, as noted in Remark 6, is (a) λ_t is replaced with μ_t and (b) q- and v-functions only depend on recent state and action rather than full past trajectory. Again notice that in $\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_1)}$, $\hat{q}_t^{(j)}$ refers to $\hat{q}_t^{(j)}(\mathcal{H}_{a_t})$, while in $\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_2)}$, $\hat{q}_t^{(j)}$ refers to the much simpler $\hat{q}_t^{(j)}(s_t, a_t)$.

Again, to establish the properties of DRL for MDPs, we assume the following conditions reflecting Assumptions 1 and 2:

Assumption 4 (Bounded estimators)**.** $\limsup_{n\to\infty} \|\hat{\mu}_t^{(j)}\|_\infty < \infty$, $\limsup_{n\to\infty} \|\hat{q}_t^{(j)}\|_\infty < \infty$ for each $0 \le t \le T$, $1 \le j \le K$.

Like Assumption 3, we expect Assumption 4 to hold under Assumptions 1 and 2 as the latter provides that $\|\mu_t^{(j)}\|_\infty \le C'$, $\|q_t^{(j)}\| \le (T+1-t)R_{\max}$. And, for many non-parametric estimators, Assumption 4 will necessarily hold. Again, we need not have that the same bounds hold for the estimators as hold for the estimands; any finite bound will do. For the rest of this subsection we will assume Assumption 4 hold.

The following result establishes that DRL is the first efficient OPE estimator for MDPs. In fact, it is efficient even if each nuisance estimator has a slow, nonparametric convergence rate (sub-$\sqrt{n}$). Moreover, as before, we make no restrictive Donsker assumption; the only required condition is the convergence rate condition.. This result leverages our novel derivation of the efficient influence function in Theorem 3 and the structure of the influence function, which ensures no variance inflation due to estimating the nuisance functions.

Theorem 13 (Efficiency of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$ under $\mathcal{M}_2$)**.** *Suppose* $\|\hat{\mu}_t^{(j)} - \mu_t\|_2 \|\hat{q}_t^{(j)} - q_t\|_2 = \mathrm{o}_p(n^{-1/2})$, $\|\hat{\mu}_t^{(j)} - \mu_t\|_2 = \mathrm{o}_p(1)$, $\|\hat{q}_t^{(j)} - q_t\|_2 = \mathrm{o}_p(1)$ *for* $0 \le t \le T, 1 \le j \le K$. *Then, the estimator* $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$ *achieves the semiparametric efficiency bound under* $\mathcal{M}_2$.

Remark 11 (Example cases)**.** We note a few specific cases of Theorem 13.

- *Tabular case*: Suppose the state and action spaces are finite: $|\mathcal{S}_t|, |\mathcal{A}_t| < \infty$. Then both μ_t and q_t are *parametric* functions with parameters given by their values at each (s_t, a_t) pair. They can therefore be easily estimated at $\mathrm{O}_p(n^{-1/2})$ rates, ensuring the above rate conditions are easily satisfied. For example, we can use simple frequency estimators that simply take sample averages within each (s_t, a_t) bin (Li and Racine, 2007, Chapter 3). Other examples and additional detail are given in Sections 3.3 and 4.

- *Finite state space, known behavior policy*: Suppose now only $|\mathcal{S}_t| < \infty$ while $\mathcal{A}_t$ can be continuous and that η_t is known. Then $\mu_t(s_t, a_t) = \eta_t(s_t, a_t) w_t(s_t)$ and $w_t(s_t) = \mathrm{E}[\lambda_{t-1} \mid s_t]$ is a parametric function easily estimated at $\mathrm{O}_p(n^{-1/2})$ rates using a frequency estimator or using a recursive estimator as in Xie et al. (2019) (more detail in Section 3.3). It therefore suffices for q-function estimators to have errors that are $\mathrm{o}_p(1)$, *i.e.*, only consistency is needed without a rate. Since $\sum_{k=t}^{T} r_k$ has finite variance (it is bounded) and q_t is its regression on (s_t, a_t), Theorems 4.2, 5.1, 6.1, 10.3, and 16.1 of Györfi et al. (2006) establish that this can be done with any of histogram estimates, kernel regression, k-nearest neighbor, sieve regression, or neural networks of growing width, respectively. (These provide L^2 convergence of L^2 errors, which is stronger than the in-probability convergence of L^2 errors we require.)

- *Nonparametric case*: In the fully nonparametric case, our nuisance estimators may converge more slowly than $\mathrm{O}_p(n^{-1/2})$. Our result nonetheless accommodates such lower rates and, crucially, does not impose strong metric entropy conditions that would exclude flexible machine learning estimators. We discuss in greater detail how one might estimate the nuisance functions in Sections 3.3 and 4.

As before, we can also obtain a finite-sample guarantee for DRL in $\mathcal{M}_2$ with a leading constant controlled by the asymptotic variance, which in this case is efficient. If we can say

C_1, C_2 do not depend on C^{T+1}, this gives a finite sample result with a sample complexity, which depends only C' not on C^{T+1}, noting $\text{Effbd}(\mathcal{M}_2)$ is bounded by $C' R_{\max}^2 T^2$.

Theorem 14. *Suppose that for some C_1, C_2, for every n, with probability at least $1-\delta$, we simultaneously have that $\|\hat{\mu}_t^{(j)} - \mu_t\|_2^2 \leq \kappa_1 = C_1(n^{-2\alpha_1} + \log(2KT/\delta)/n)$, $\|\hat{q}_t^{(j)} - \lambda_t\|_2^2 \leq \kappa_2 = C_2(n^{-2\alpha_2} + \log(2KT/\delta)/n)$ $\forall t \leq T$, $1 \leq j \leq K$. Suppose moreover that $0 \leq \hat{\mu}_t \leq C'$, $0 \leq \hat{q}_t \leq (T+1-t)R_{\max}$ for $t \leq T$. Then, for every n, with probability at least $1-7\delta$, we have*

$$\begin{aligned}\left|\hat{\rho}^{\pi_e}_{\text{DRL}(\mathcal{M}_2)} - \rho^{\pi_e}\right| \leq & \sqrt{\frac{2\log(2/\delta)\text{Effbd}(\mathcal{M}_2)}{n}} \\ & + Q_1\sqrt{\frac{\log(2/\delta)T^2(TR_{\max}C'\sqrt{\kappa_1\kappa_2} + \kappa_1 T^2 R_{\max}^2 + \kappa_2\{C'\}^2)}{n}} \\ & + Q_2\frac{\log(2/\delta)TR_{\max}C'}{n} + Q_3 T\sqrt{\kappa_1\kappa_2},\end{aligned}$$

where Q_1, Q_2, Q_3 are constants not depending on $\delta, T, R_{\max}, C', n, C_1, C_2, C$.

As before, if $\alpha_1 > 0, \alpha_2 > 0, \alpha_1 + \alpha_2 > 1/2$, then the leading term in n in the above bound is exactly $\sqrt{\frac{2\log(2/\delta)\text{Effbd}(\mathcal{M}_2)}{n}}$ (with no constant factor). Note that, while C_1 (embedded inside κ_1) does not appear in the leading n term, ensuring the conditions of Theorem 14 may require C_1 to depend on T, depending on which estimate $\hat{\mu}_t^{(j)}$ for μ_t is used; see Remark 17.

The DRL estimator in $\mathcal{M}_2$ can also achieve efficiency without sample splitting (*i.e.*, with adaptive in-sample estimation of nuisances) if we impose an Donsker condition on the estimated nuisances.

Theorem 15 (Efficiency without sample splitting)**.** *Let $\hat{\mu}_t, \hat{q}_t$ be estimators based on $\mathcal{D}$ and let $\hat{\rho}^{\pi^e}_{\text{DRL}(\mathcal{M}_2),\,\text{adaptive}} = \mathrm{E}_n\left[\sum_{t=0}^{T}\left(\hat{\mu}_t\left(r_t - \hat{q}_t\right) - \hat{\mu}_{t-1}\mathrm{E}_{\pi^e}\left[\hat{q}_t \mid \mathcal{H}_{s_t}\right]\right)\right]$. Suppose $\|\hat{\mu}_t - \mu_t\|_2\|\hat{q}_t - q_t\|_2 = \mathrm{o}_p(n^{-1/2})$, $\|\hat{\mu}_t - \mu_t\|_2 = \mathrm{o}_p(1)$, $\|\hat{q}_t - q_t\|_2 = \mathrm{o}_p(1)$ for $0 \leq t \leq T$ and that $\hat{q}_t$, $\hat{\mu}_t$ belong to a Donsker class. Then, the estimator $\hat{\rho}^{\pi^e}_{\text{DRL}(\mathcal{M}_2),\,\text{adaptive}}$ achieves the semiparametric efficiency bound under $\mathcal{M}_2$.*

In addition to efficiency, DRL enjoys a double robustness guarantee in $\mathcal{M}_2$.

Theorem 16 (Double robustness ($\sqrt{n}$-consistency))**.** *Suppose $\|\hat{\mu}_t^{(j)} - \mu_t^\dagger\|_2 = \mathrm{O}_p(n^{-\alpha_1})$ and $\|\hat{q}_t^{(j)} - q_t^\dagger\|_2 = \mathrm{O}_p(n^{-\alpha_2})$ for $0 \leq t \leq T, 1 \leq j \leq K$. If, for each $0 \leq t \leq T$, either $\mu_t^\dagger = \mu_t$ and $\alpha_1 \geq 1/2$, $\alpha_2 > 0$ or $q_t^\dagger = q_t$ and $\alpha_2 \geq 1/2$, $\alpha_1 > 0$, then the estimator $\hat{\rho}_{\text{DRL}(\mathcal{M}_2)}$ is $\sqrt{n}$-consistent around ρ^{π^e}.*

Again, we obtain consistency without a rate even if just one nuisance is consistent without a rate.

Corollary 17 (Double robustness (consistency))**.** *Suppose $\|\hat{\mu}_t^{(j)} - \mu_t^\dagger\|_2 = \mathrm{o}_p(1)$ and $\|\hat{q}_t^{(j)} - q_t^\dagger\|_2 = \mathrm{o}_p(1)$ for $0 \leq t \leq T, 1 \leq j \leq K$. If, for each $0 \leq t \leq T$, either $\mu_t^\dagger = \mu_t$ or $q_t^\dagger = q_t$, then the estimator $\hat{\rho}_{\text{DRL}(\mathcal{M}_2)}$ is consistent around ρ^{π^e}.*

Remark 12. When the behavior policy is known, the estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ is still $\sqrt{n}$-consistent under $\mathcal{M}_2$ even without smoothness conditions on μ_t because $\mathcal{M}_2$ is included in $\mathcal{M}_1$ so that Theorem 10 applies. On the other hand, in the nonparametric setting, the estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$ requires some smoothness conditions even if the behavior policy is known because μ_t must still be estimated. In this sense, when the behavior policy is known, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ is more robust than $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$ under $\mathcal{M}_2$ but its asymptotic variance is bigger, and generally strictly so.

A reaming question is how to estimate the nuisances at the necessary rates. We discuss q-function estimation in Section 4. For estimating μ_k, one can leverage the following relationship to reduce it to a regression problem:

$$\mu_t(s_t, a_t) = \eta_t(s_t, a_t) w_t(s_t), \quad \text{where } w_t(s_t) = \mathrm{E}[\lambda_{t-1} \mid s_t]. \tag{5}$$

Thus, for example, when the behavior policy is known, we need only estimate w_t, which amounts to regressing λ_{t-1} on s_t. So, in particular, if $w_t(s_t)$ belongs to a Hölder class with smoothness α and s_t has dimension d_s, estimating w_t with a sieve-type estimator $\hat{w}_t$ based on the loss function $(\lambda_{t-1} - w_t(s_t))^2$ and letting $\hat{\mu}_t^{(j)}(s_t, a_t) = \eta_t(s_t, a_t)\hat{w}_t^{(j)}(s_t)$ will give a convergence rate $\|\hat{\mu}_t^{(j)}(s_t, a_t) - \mu_t(s_t, a_t)\|_2 = \mathrm{O}_p(n^{-\alpha/(\alpha+d_{s_t})})$ (Chen, 2007). When the behavior policy is unknown, it can be first estimated to construct $\hat{\lambda}_t$ and we can repeat the above replacing λ_t with $\hat{\lambda}_t$. In particular, there will be no deterioration in rate if π_t^b also belongs to a Hölder class with smoothness α and if we further split each $\mathcal{D}_j$, estimate π_b^t as in Theorem 12 on one half, and plug it in to estimate w_t on the other half. Further strategies for estimating μ_t are discussed in Section 3.3 below.

In the special case where we use parametric models for μ_t and q_t, under some regularity conditions, parametric estimators will generally satisfy $\|\hat{\mu}_t - \mu_t^\dagger\|_2 = \mathrm{O}_p(n^{-1/2})$ and $\|\hat{q}_t - q_t^\dagger\|_2 = \mathrm{O}_p(n^{-1/2})$, where $q_t^\dagger = q_t$ and $\mu_t^\dagger = \mu_t$ if the models are well-specified. (See Section 4 regarding estimating the q-function). Thus, if both models are correctly specified, then Theorem 13 yields local efficiency. If only one model is correctly specified, Theorem 16 yields double robustness.

3.3. Estimating Marginalized Density Ratios and the Inefficiency of Marginalized Importance Sampling

In this section we discuss strategies for estimating μ_t and also show that doing OPE estimation using only marginalized density ratios, as recently proposed, leads to inefficient evaluation in $\mathcal{M}_2$.

Given an estimate $\hat{\mu}_t$ of μ_t, the Marginalized Importance Sampling (MIS) is given by

$$\hat{\rho}_{\mathrm{MIS}}^{\pi^e} = \mathrm{E}_n\left[\sum_{t=0}^{T} \hat{\mu}_t r_t\right], \tag{6}$$

which resembles the IS estimator but where we replace $\hat{\lambda}_t$ with $\hat{\mu}_t$. Note that $\mu_t = \mathrm{E}[\lambda_t \mid s_t, a_t]$, *i.e.*, the marginalization of λ_t over $\mathcal{H}_{a_{t-1}}$; hence the name MIS. Equation (6) can also be seen as DRL without sample splitting when we let $\hat{q}_t = 0$.

MIS is a generic meta-estimator that depends on a particular estimate $\hat{\mu}_t$. As we will see in this section, the asymptotic variance of MIS depends on *how* we estimate $\hat{\mu}_t$; this is

unlike DRL, whose asymptotic variance is insensitive to this choice as long as it satisfied lax rate conditions. To study MIS, we therefore study different instantiations of Eq. (6) for different μ_t estimators. Xie et al. (2019) provides one possible estimate in the special finite-state case. We will next study general-case estimators $\hat{\mu}_t$ based on regression. See also Remark 17 regarding different μ_t estimators. Note that since $\mu_t = \eta_t w_t$, when behavior policy is known we can estimate μ_t as $\hat{\mu}_t = \eta_t \hat{w}_t$. We focus on two cases: when $\hat{w}$ is estimated using a histogram by averaging λ_{t-1} by state over a finite state space and a nonparametric extension.

Theorem 18 (Asymptotic variance of $\hat{\rho}^{\pi^e}_{\text{MIS}}$ with finite state space)**.** *Suppose $|\mathcal{S}_t| < \infty$ for $0 \leq t \leq T$. Let*

$$\hat{w}_t(s_t) = \frac{\sum_{i=1}^n \mathbb{I}[s_t^{(i)} = s_t]\lambda_{t-1}}{\sum_{i=1}^n \mathbb{I}[s_t^{(i)} = s_t]}. \tag{7}$$

Then $\hat{\rho}^{\pi^e}_{\text{MIS}}$ is consistent and asymptotically normal (CAN) around ρ^{π^e} and its asymptotic MSE is

$$\text{var}\left[\sum_{t=0}^T \mu_t r_t + (\lambda_{t-1} - w_t)\text{E}_{\pi_e}[r_t \mid s_t]\right]. \tag{8}$$

For the proof of Theorem 18, we use an argument based on the theory of U-statistics (van der Vaart, 1998, Ch. 12) in order to rephrase the MIS estimator with $\hat{w}$ as in Eq. (7) in an asymptotically linear form: $\hat{\rho}^{\pi^e}_{\text{MIS}} = \text{E}_n[\sum_{t=0}^T \mu_t r_t + (\lambda_{t-1} - w_t)\text{E}_{\pi_e}[r_t|s_t]] + \text{o}_p(n^{-1/2})$.

This influence function is different from the efficient influence function; therefore, $\hat{\rho}^{\pi^e}_{\text{MIS}}$ with histogram nuisance estimators is not efficient (the efficient influence function is unique). In fact, we can confirm this fact by calculating and comparing the variances.

Theorem 19. *If $P_{\pi^b} \in \mathcal{M}_2$ (i.e., the underlying distribution is an MDP), Eq.* (8) *is greater than or equal to* $\text{EffBd}(\mathcal{M}_2)$*. The difference is*

$$\text{var}[v_0] + \sum_{t=0}^{T-1} \text{E}\left[(w_t - \lambda_{t-1})^2 \text{var}\left(\eta_t v_t(s_{t+1})|s_t\right)\right].$$

We now turn to the nonparametric case, where we first consider a sieve-type extension of the $\hat{w}$ estimator.

Theorem 20 (Asymptotic variance of $\hat{\rho}^{\pi^e}_{\text{MIS}}$ with nonparametric w_t estimate)**.** *Suppose $\text{E}[(\lambda_t - \mu_t)^q] < \infty$ for some $q > 1$. Let*

$$\hat{w}_t(s_t) = \underset{w_t(s_t) \in \Lambda^\alpha_{d_{s_t}}}{\arg\min}\ \text{E}_n[(w_t(s_t) - \lambda_{t-1})^2], \tag{9}$$

where $\Lambda^\alpha_{d_{s_t}}$ is the space of Hölder functions with smoothness α and the dimension d_{s_t}. Assume $w_t \in \Lambda^\alpha_{d_{s_t}}$, $\alpha/(2\alpha + d_{s_t}) > 1/4$. Then the estimator $\hat{\rho}^{\pi^e}_{\text{MIS}}$ is CAN around ρ^{π^e} and its asymptotic MSE is equal to Eq. (8)*.*

Remark 13. The estimator Eq. (9) is over an infinite-dimensional function space. It can be replaced with a finite-dimensional approximation Λ_n^α such that $\Lambda_n^\alpha \to \Lambda_{ds}^\alpha$. Following Example 1(b) in Section 8 (Shen, 1997), it can be shown that this will lead to the same asymptotic MSE as in Eq. (8) and not change the conclusion of Theorem 20.

Remark 14. When the action and sample space is continuous, the histogram estimator in Eq. (7) can also easily be extended to a kernel estimator:

$$\hat{w}_t(s_t) = \frac{\sum_{i=1}^n \mathrm{K}_h(s_t^{(i)} - s_t)\lambda_{t-1}}{\sum_{i=1}^n \mathrm{K}_h(s_t^{(i)} - s_t)}, \tag{10}$$

where K_h is a kernel with a bandwidth h.

The smoothness condition in Theorem 20 ensures we can estimate w_t at fourth-root rates using Eq. (9). Following Newey and Mcfadden (1994) and utilizing a high-order kernel, we can obtain similar fourth-root rates for Eq. (10) and a similar variance result for MIS. Unlike Eq. (9), we cannot invoke a Donsker condition to prove a stochastic equicontinuity condition. However, it is still possible to show this directly based on a V-statistics theory (see Chapter 8 of Newey and Mcfadden, 1994).

Finally, we also consider estimating μ_t directly and nonparametrically using the relation

$$\mu_t(s_t, a_t) = \mathrm{E}[\lambda_t \mid s_t, a_t]. \tag{11}$$

A sieve-type regression estimator for μ_t is then constructed as

$$\hat{\mu}_t(s_t, a_t) = \operatorname*{arg\,min}_{\mu_t(s_t,a_t)\in\Lambda^\alpha_{d_{s_t}+d_{a_t}}} \mathrm{E}_n[(\mu_t(s_t, a_t) - \lambda_t)^2]. \tag{12}$$

Theorem 21 (Asymptotic variance of $\hat{\rho}^{\pi^e}_{\mathrm{MIS}}$ with nonparametric μ_t estimate)**.** *Suppose* $\mathrm{E}[(\lambda_t - \mu_t)^q] < \infty$ *for some* $q > 1$. *Let* $\hat{\mu}_t$ *be as in Eq.* (12). *Assume* $\mu_t \in \Lambda^\alpha_{d_{s_t}+d_{a_t}}$, $\alpha/(2\alpha + d_{s_t} + d_{a_t}) > 1/4$. *Then the estimator* $\mathrm{E}_n\left[\sum_{t=0}^T \hat{\mu}_t r_t\right]$ *is CAN around* ρ^{π^e} *and its asymptotic MSE is equal to*

$$\mathrm{var}\left[\sum_{t=0}^T \mu_t r_t + (\lambda_t - \mu_t)\mathrm{E}[r_t \mid s_t, a_t]\right]. \tag{13}$$

Remark 15. While both estimators for μ_t in Theorems 20 and 21 achieve fourth-root rates under the respective conditions, the resulting asymptotic variances in Eqs. (8) and (13) are different and generally incomparable. Both are inefficient, but which is larger is problem-dependent. Note that, in contrast, the asymptotic variance of DRL (Theorem 13) is the same (and is efficient) regardless of which way is used to estimate μ_t as long as we have the necessary rate. When the behavior policy is known, using Eq. (9) may be better than Eq. (12) when estimating μ_t nonparametrically because the smoothness condition is weaker and the convergence rate is faster (since $d_{s_t} < d_{a_t} + d_{s_t}$). However, when using parametric models, the rates are the same (under correct specification) and sometimes it is easier to model $\mu_t(s_t, a_t)$ rather than $w_t(s_t)$, as we do in Section 5.1.

Remark 16 (The Inefficiency of MIS)**.** Theorems 18, 20 and 21 each study the MIS estimator $\hat{\rho}^{\pi^e}_{\text{MIS}}$ in Eq. (6) but with different estimator for the nuisance $\mu_t = \eta_t w_t$. As noted above, unlike DRL, the variance of the MIS estimator actually depends on the way this nuisance is estimated. And, in each case, the MIS estimator was inefficient. In the finite-state-space setting with known behavior policy, Xie et al. (2019) propose another, different MIS estimator based on estimating the MDP transition kernel; but per their Remark 4 it is also inefficient. (In contrast, Remark 11 shows that DRL is efficient in the finite-state-space setting *without* requiring any smoothness conditions.) This does not immediately imply the MIS estimator is always inefficient, as it may depend on how μ_t is estimated, but semiparametric theory strongly suggests there is reason to believe that MIS would in general be inefficient.

One natural question that sheds light on this is how would a hypothetical MIS estimator perform with oracle values for μ_t. In fact, the variance of $\sum_{t=0}^T \mu_t r_t$ is in general *incomparable* to $\text{EffBd}(\mathcal{M}_2)$, that is, it may be smaller *or* larger depending on the particular instance. This may surprising but is not contradictory since one can in fact prove that no regular estimator (let alone an efficient one) in either $\mathcal{M}_2$ or $\mathcal{M}_{2b}$ could ever have the form $\text{E}_n[\sum_{t=0}^T \mu_t r_t] + \text{o}_p(n^{-1/2})$, that is, asymptotically linear with influence function $\sum_{t=0}^T \mu_t r_t$. This is because $\sum_{t=0}^T \mu_t r_t$ is *not* a gradient of ρ^{π^e} under either $\mathcal{M}_2$ or $\mathcal{M}_{2b}$ (see Theorem 24). This is in stark contrast to IS: $\sum_{t=0}^T \lambda_t r_t$ is always a valid influence function under either $\mathcal{M}_{1b}$ or $\mathcal{M}_{2b}$ since we know its empirical average always gives an unbiased linear estimator (not just asymptotically). Indeed, we similarly have that the variance of $\sum_{t=0}^T \mu_t r_t$ is *also* incomparable to $\sum_{t=0}^T \lambda_t r_t$. The function $\sum_{t=0}^T \mu_t r_t$ is an influence function (a gradient) under the MDP model with *known* transition kernel, but that is a very restrictive and unrealistic model.

One interesting specific case is the fully tabular setting (finite state *and* action spaces). Since our paper was posted, the more recent Yin and Wang (2020) considered a "modified" version of the estimator of Xie et al. (2019) in order to obtain efficiency under the tabular case of $\mathcal{M}_{2b}$. By simple algebra, the estimator of Yin and Wang (2020), which is defined as

$$\begin{aligned}
&\tfrac{1}{n}\textstyle\sum_{i=1}^n \sum_{t=0}^T \hat{w}_t(s_t^{(i)}) \int r_t \hat{P}_{r_t}(r_t|s_t^{(i)}, a_t)\pi^\text{e}(a_t|s_t^{(i)})\text{d}(r_t, a_t),\\
&\text{where } \hat{w}_t(s_t) = \tfrac{1}{\hat{p}_{\pi_t^\text{b}}(s_t)} \textstyle\int \hat{P}_{s_t}(s_t|s_{t-1}, a_{t-1}) \prod_{k=0}^{t-1} \left(\pi_k^\text{e}(a_k|s_k)\hat{P}_{s_k}(s_k|s_{k-1}, a_{k-1})\right) \text{d}(\mathcal{H}_{a_{t-1}}),
\end{aligned}$$

and where $\hat{P}_{s_t}, \hat{P}_{r_t}, \hat{p}_{\pi_t^\text{b}}$ are each an empirical frequency (histogram) estimator, can in fact be rewritten simply as

$$\sum_{t=0}^T \int r_t \hat{P}_{r_t}(r_t|s_t, a_t) \prod_{k=0}^t \left(\pi_k^\text{e}(a_k|s_k)\hat{P}_{s_k}(s_k|s_{k-1}, a_{k-1})\right) \text{d}(\mathcal{H}_{a_t}, r_t).$$

This is essentially a model-based OPE estimator, where we first fit *all* MDP parameters and then explicitly integrate with respect to the resulting estimated trajectory density function in order to compute the expectation ρ^{π^e}. This is also often called the G-formula in the causal inference literature (Hernan and Robins, 2019; Robins, 1986). In the tabular setting, the efficiency of this estimator is immediate since it is exactly the (parametric) MLE estimate for this setting, which is well known to achieve the Cramér-Rao lower bound, and the Cramér-Rao lower bound is the efficiency bound in the tabular setting since the

model is parametric. In the case of continuous state and/or action spaces, the simple extension of replacing $\hat{P}_{s_{k+1}}(s_{k+1}|s_k,a_k), \hat{P}_{r_t}(r_t|s_t,a_t)$ with some nonparametric conditional density estimators would have poor performance since the nonparametric density estimation is unstable and would significantly inflate the variance. Alternatively, if we extend the estimator by instead estimating μ_t or w_t nonparametrically, the above already argues why we expect this would generally be inefficient.

Similarly to the model-based approach, in the tabular case, our results in the next section have shown that also simple DM estimates based on q-function estimation are also efficient, since in the tabular case q-functions are parametric. Moreover, DRL was already shown to be efficient in the the tabular MDP setting with *any* parametric μ_t estimator as they all have $\mathrm{O}_p(1/\sqrt{n})$ convergence in this setting, and the particular choice of estimator does not affect this (see also Remark 11). Hence, DRL was the *first* efficient OPE estimator, both in general and in the tabular MDP setting in particular.

Remark 17 (Other estimators for μ_t)**.** Xie et al. (2019); Yin and Wang (2020) may in fact both offer alternative estimators for w_t and hences μ_t (in their respective settings) that may be used in DRL and either will ensure efficiency for DRL (see Remark 11). In particular, since λ_t may have variance growing exponentially in T, this may affect the variance of estimates of μ_t based on the regression of it on s_t or on s_t, a_t, as studied above. Although this will not appear in the leading term of the variance of DRL and will not affect efficiency, it may still be a concern. Developing and analyzing alternative estimators for w_t and/or μ_t may be fruitful future work. For example, still other possible estimation approaches for w_t include a fitted w-iteration: start with $\hat{w}_0 \equiv 1$, regress $\hat{w}_{t-1}\eta_t$ on s_t using any supervised regression method to obtain $\hat{w}_t$, and repeat.

4. Estimating the q-function and Efficiency Under $\mathcal{M}_{1q}$, $\mathcal{M}_{2q}$

In this section, we discuss the estimation of q-functions in an off-policy manner, parametrically or nonparametrically, which can be plugged into our estimators, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}, \hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$. On the way, we also derive the semiparametric efficiency bound when we impose parametric restrictions on q-functions, *i.e.*, the models $\mathcal{M}_{1q}$, $\mathcal{M}_{2q}$.

To do this, we will leverage a recursive definition of the q-functions (Bertsekas, 2012). Under $\mathcal{M}_1$, the following recursion equation holds:

$$q_t = \mathrm{E}\left[r_t + \mathrm{E}_{\pi^e}\left[q_{t+1} \mid \mathcal{H}_{s_{t+1}}\right] \mid \mathcal{H}_{a_t}\right]. \tag{14}$$

Under $\mathcal{M}_2$, we can further replace $\mathcal{H}_{s_{t+1}}$ with s_{t+1} and $\mathcal{H}_{a_t}$ with (s_t, a_t) in the above.

The recursion in Eq. (14) can equivalently be written as a set of conditional moment equations satisfied by the q-functions:

$$\begin{aligned} &m_t(\mathcal{H}_{a_t}; \{q_1, \dots, q_T\}) = 0 \quad \forall t \le T, \\ &\text{where} \quad m_t(\mathcal{H}_{a_t}; \{q'_1, \dots, q'_T\}) = \mathrm{E}\left[r_t + \mathrm{E}_{\pi^e}\left[q'_{t+1}(\mathcal{H}_{a_t+1}) \mid \mathcal{H}_{s_{t+1}}\right] - q'_t(\mathcal{H}_{a_t}) \mid \mathcal{H}_{a_t}\right]. \end{aligned} \tag{15}$$

This formulation of the q-function in terms of conditional moment equations, along with the observation that $\rho^{\pi^e} = \mathrm{E}\left[\mathrm{E}_{\pi^e}\left[q_0(s_0,a_0) \mid s_0\right]\right]$ is determined by the q-function, allows us both to estimate the q-function efficiently, either parametrically and nonparametrically, and to characterize the efficiency bounds under $\mathcal{M}_{1q}$ and $\mathcal{M}_{2q}$. We start with the latter.

4.1. Efficiency Bounds Under $\mathcal{M}_{1q}$, $\mathcal{M}_{2q}$

In this section we consider the models where we restrict q-functions parametrically:

$$\mathcal{M}_{1q} = \{P_{\pi^b} \in \mathcal{M}_1 : \exists \beta_t^* \in \Theta_{\beta_t},\ q_t(\mathcal{H}_{a_t}) = q_t(\mathcal{H}_{a_t}; \beta_t^*)\ \ \forall t \leq T\},$$
$$\mathcal{M}_{2q} = \{P_{\pi^b} \in \mathcal{M}_2 : \exists \beta_t^* \in \Theta_{\beta_t},\ q_t(s_t, a_t) = q_t(s_t, a_t; \beta_t^*)\ \ \forall t \leq T\},$$

where $q_t(\mathcal{H}_{a_t}; \beta_t)$ or $q_t(s_t, a_t; \beta_t)$ is some parametric model for the q-function at time t that is continuously differentiable with respect to the parameter β_t, Θ_{β_t} is some compact parameter space, and β_t^* is the true parameter, which is assumed to lie in the interior of Θ_{β_t}. For brevity we define $v_t(\mathcal{H}_{s_t}; \beta_t) = \mathrm{E}_{\pi^e}\left[q_t(\mathcal{H}_{a_t}; \beta_t) \mid \mathcal{H}_{s_t}\right]$ and similarly $v_t(s_t; \beta_t)$.

Under $\mathcal{M}_{1q}$, $\mathcal{M}_{2q}$, Eq. (14) can be rephrased as as a set of conditional moment restrictions on the parameter β defined by $\beta = (\beta_1^\top, \cdots \beta_T^\top)^\top$. In particular, overloading notation and letting $m_t(\mathcal{H}_{a_t}; \beta) = m_t(\mathcal{H}_{a_t}; \{q_1(\cdot, \beta_1), \ldots, q_T(\cdot, \beta_T)\})$, we have that β is defined by the set of conditional moment equations $m_t(\mathcal{H}_{a_t}; \beta) = 0\ \forall t \leq T$. This observation is key in establishing the following result.

Theorem 22 (Efficiency bound under $\mathcal{M}_{1q}$, $\mathcal{M}_{2q}$)**.** *Define* $e_{q,t} = r_t + v_{t+1} - q_t$

$$\begin{aligned}
A_t &= C_t^{-1} + C_t^{-1} B_t A_{t+1} B_t^\top C_t^{-1},\\
B_t &= \mathrm{E}\left[\nabla_{\beta_t} q_t(\mathcal{H}_{a_t}; \beta_t^*) \mathrm{var}(e_{q,t} \mid \mathcal{H}_{a_t})^{-1} \nabla_{\beta_{t+1}}^\top v_{t+1}(\mathcal{H}_{s_{t+1}}; \beta_{t+1}^*)\right],\\
C_t &= \mathrm{E}\left[\nabla_{\beta_t} q_t(\mathcal{H}_{a_t}; \beta_t^*) \mathrm{var}(e_{q,t} \mid \mathcal{H}_{a_t})^{-1} \nabla_{\beta_t}^\top q_t(\mathcal{H}_{a_t}; \beta_t^*)\right],\\
A_T &= \mathrm{E}\left[\nabla_{\beta_T} q_T(\mathcal{H}_{a_T}; \beta_T^*) \mathrm{var}(e_{q,T} \mid \mathcal{H}_{a_T})^{-1} \nabla_{\beta_T}^\top q_T(\mathcal{H}_{a_T}; \beta_T^*)\right]^{-1},\\
B_{-1} &= \mathrm{E}[\eta_0(s_0, a_0) \nabla_{\beta_0}^\top q_0(s_0, a_0; \beta_0^*)].
\end{aligned}$$

Then

$$\mathrm{EffBd}(\mathcal{M}_{1q}) = \mathrm{var}\left(v_0\right) + B_{-1} A_0 B_{-1}^\top.$$

Moreover, the efficiency bound for estimating β_t is A_t.

Finally, the corresponding efficiency bounds under $\mathcal{M}_{2q}$ are given by replacing $\mathcal{H}_{s_{t+1}}$ with s_{t+1} and $\mathcal{H}_{a_t}$ with (s_t, a_t) everywhere in the above.

Remark 18. When $T = 1$, $\mathrm{EffBd}(\mathcal{M}_{1q})$ above is equal to

$$\begin{aligned}
&\mathrm{var}[v_0(s_0)] + B_{-1} A_0 B_{-1}^\top,\\
\text{where}\quad &A_0 = \mathrm{E}[\nabla_{\beta_0} q_0(s_0, a_0; \beta_0^*) \mathrm{var}[r_0 \mid s_0, a_0]^{-1} \nabla_\beta^\top q_0(s_0, a_0; \beta_0^*)]^{-1},
\end{aligned}$$

The Matrix Cauchy-Schwarz Inequality (Tripathi, 1999) immediately shows that this is upper bounded by $\mathrm{EffBd}(\mathcal{M}_1)$, as is also implied by $\mathcal{M}_{1q} \subset \mathcal{M}_1$ albeit less directly.

4.2. Parametric Estimation of q-functions

Next, we consider an estimation method for β_t and ρ^{π^e}. Given the above observations, a natural way to estimate β is by solving the following set of conditional moment equations given by $m_t(\mathcal{H}_{a_t}; \beta) = 0\ \forall t \leq T$. For example, one approach when the q-model is a linear model specified as $\beta_t^\top \phi_t(\mathcal{H}_{a_t})$ for some d_{ϕ_t}-dimensional feature expansion ϕ_t is to choose $\hat{\beta}$

to minimize $\sum_{t=0}^{T}\sum_{i=1}^{d_{\phi_t}}\left(\mathrm{E}_n\left[\left(r_t+v_{t+1}(\mathcal{H}_{s_{t+1}};\beta_{t+1})-q_t(\mathcal{H}_{a_t};\beta_t)\right)\phi_{ti}(\mathcal{H}_{a_t})\right]\right)^2$, which corresponds exactly to backward-recursive ordinary least squares. That is, first r_T is regressed on $\phi_t(\mathcal{H}_{a_T})$ to obtain $\hat{\beta}_T$, then $q_T(\mathcal{H}_{a_T};\hat{\beta}_T)$ is averaged over $\pi_T^e(a_T \mid \mathcal{H}_{s_T})$ to obtain $\hat{v}_T$, then $r_{T-1}+\hat{v}_T$ is regressed on $\phi_t(\mathcal{H}_{a_{T-1}})$ to obtain $\hat{\beta}_{T-1}$, and so on.

Although such an estimator can achieve the rate $\mathrm{O}_p(n^{-1/2})$ under correct specification and standard conditions for M-estimators, it might not yield an efficient estimator for β or for ρ^{π^e}. When the q-model is linear as above, this can be easily solved by instead applying any efficient variant of the generalized method of moments (GMM), such as two-step GMM (Hansen, 1982; Hansen et al., 1996), to the set of moment equations given by $m_t(\mathcal{H}_{a_t};\beta)\phi_{ti}(\mathcal{H}_{a_t})=0\ \forall t\le T,\ i\le d_{\phi_t}$. This is almost the same as the above backward-recursive ordinary least squares but with an optimal weighting of the different moment conditions in the sum above.

When the q-model may be nonlinear, we can obtain an efficient estimator by instead applying the method of Hahn (1997) to our set of conditional moment equations. Specifically, we can consider the set of Tm_n moment equations $\mathrm{E}\left[m_t(\mathcal{H}_{a_t};\beta)\phi_{ti}(\mathcal{H}_{a_t})\right]=0\ \forall t\le T,\ i\le m_n$, where $\phi_{t1}(\mathcal{H}_{a_t}),\phi_{t2}(\mathcal{H}_{a_t}),\dots$ is a basis expansion of the L^2-space and $m_n\to\infty$ as $n\to\infty$. Then, applying any efficient variant of GMM to this set of moment conditions will yield an efficient estimator $\hat{\beta}$ of β.

In all of the above, replacing $\mathcal{H}_{s_{t+1}}$ with s_{t+1} and $\mathcal{H}_{a_t}$ with (s_t,a_t), the same techniques can be applied in $\mathcal{M}_2$. In either case, once we have an efficient estimate $\hat{\beta}$ of β, an efficient estimate for ρ^{π^e}, achieving the semiparametric efficiency bound in the appropriate model, is given by $\hat{\rho}_{DM}=\mathrm{E}_n\left[v_0(s_0;\hat{\beta}_0)\right]$.

Remark 19 (Tabular setting)**.** Consider a tabular case. Then, by treating $q_t(s_t,a_t)=\beta^\top\phi(a_t,s_t)$, where $\phi(a,s)=(\mathrm{I}(a=a_1^\dagger,s=s_1^\dagger),\cdots,\mathrm{I}(a=a_{|A_t|}^\dagger,s=s_{|S_t|}^\dagger))$ and $a_i^\dagger$ and $s_i^\dagger$ are the elements of the finite $\mathcal{A}_t$ and $\mathcal{S}_t$, we can observe that $\mathrm{Effbd}(\mathcal{M}_{2q})=\mathrm{Effbd}(\mathcal{M}_2)$ by some algebra. This result is natural since $\mathcal{M}_{2q}=\mathcal{M}_2$ in the tabular setting.

4.3. Nonparametric Estimation of q-functions

The above observation in Eq. (14) that q-functions satisfy a set of conditional moment equations also lends itself to nonparametric estimation of the q-functions. In this section we briefly review how one approach to this, following the application of the method of Ai and Chen (2012) to this set of conditional moment equation, can obtain the necessary fourth-root rates for use in DRL.

The estimator $\{\hat{q}_t\}_{t=0}^T$ is constructed as the following sieve minimum distance estimator:

$$\{\hat{q}_t\}_{t=0}^T\in\operatorname*{arg\,min}_{q_t\in\Lambda_{t,n}\,\forall t\le T}\sum_{t=0}^{T}\mathrm{E}_n\left[\hat{m}_t(\mathcal{H}_{a_t};q_t)\hat{\Sigma}_t^{-1}\hat{m}_t(\mathcal{H}_{a_t};q_t)\right],$$

where $\hat{m}_t(\mathcal{H}_{a_t};q_t)$ is a nonparameric estimator for $m_t(\mathcal{H}_{a_t};q_t)$, $\hat{\Sigma}_t$ is a nonparametric estimator for $\mathrm{var}\left(e_{q,t}\mid\mathcal{H}_{a_t}\right)$, and $\Lambda_{k,n}$ is a sequence of approximation space whose union $\cup_{n=1}^{\infty}\Lambda_{t,n}$ is dense in some infinite dimensional space Λ_t. Alternatively, in $\mathcal{M}_2$, we replace $\mathcal{H}_{a_t}$ with (s_t,a_t) in the above.

Table 1: Experiment from Section 5.1: RMSE (and standard errors).

Setting	n	$\hat{\rho}_{\text{IS}}$	$\hat{\rho}_{\text{DRL}(\mathcal{M}_1)}$	$\hat{\rho}_{\text{DM}}$	$\hat{\rho}_{\text{MIS}}$	$\hat{\rho}_{\text{DRL}(\mathcal{M}_2)}$
(1)	1500	42.4 (12.4)	36.1 (16.8)	0.70 (0.002)	40.8 (12.5)	0.70 (0.002)
	3000	20.4 (3.1)	7.8 (0.8)	0.50 (0.001)	20.8 (2.8)	0.50 (0.001)
	4500	20.2 (3.1)	6.6 (0.75)	0.43 (0.001)	21.5 (3.5)	0.43 (0.001)
(2)	1500	42.4 (12.4)	77.6 (29.1)	10.8 (0.002)	40.8 (12.5)	10.3 (3.5)
	3000	20.4 (2.5)	36.6 (6.9)	10.8 (0.001)	20.8 (2.8)	6.0 (0.6)
	4500	20.2 (3.1)	34.4 (9.6)	10.8 (0.001)	21.5 (3.5)	5.5 (2.0)
(3)	1500	42.4 (12.4)	36.1 (16.8)	0.70 (0.002)	87.7 (25.5)	0.73 (0.03)
	3000	20.4 (3.1)	7.8 (0.8)	0.50 (0.001)	37.3 (3.2)	0.51 (0.002)
	4500	20.2 (3.1)	6.6 (0.75)	0.43 (0.001)	53.5 (15.1)	0.44 (0.005)

Ai and Chen (2003) prove that applying the above with appropriate nonparametric estimators, under some smoothness conditions, we can obtain $\|\hat{q}_t - q_t\|_{F,t} = \mathrm{o}_p(n^{-1/4})$, where $\|\cdot\|_{F,t}$ is the Fisher metric, which in our setting of Eq. (15) is defined as

$$\|g(\mathcal{H}_{a_t})\|_{F,t}^2 = \mathrm{E}[\mathrm{var}(e_{q,t} \mid \mathcal{H}_{a_t})g^2 + \mathrm{var}(e_{q,t-1} \mid \mathcal{H}_{a_{t-1}})\mathrm{E}_{\pi^e}\left[g(\mathcal{H}_{a_t}) \mid \mathcal{H}_{s_t}\right]^2].$$

We omit the details and refer the interested reader to Ai and Chen (2003). We only prove that this norm is in fact equivalent to the L^2-norm under mild conditions.

Lemma 23. *Suppose* $\mathrm{var}[e_{q,t} \mid \mathcal{H}_{a_t}]$ *and* $\mathrm{var}[e_{q,t-1} \mid \mathcal{H}_{a_{t-1}}]$ *are bounded away from zero. Then,* $\|\cdot\|_{F,k}$ *and* $\|\cdot\|_2$ *are equivalent norms.*

This means that, under the appropriate conditions, the estimator $\hat{q}$ obtains the rate $\mathrm{o}_p(n^{-1/4})$ in terms of L^2-norm, as necessary for Theorems 6, 10, 13 and 16.

5. Experiments

We now turn to an empirical study of OPE and DRL. First, we construct a simulation to investigate the effect of using memorylessness on estimation variance as well as the effect of double robustness on model specification sensitivity. Then, we study comparative performance of different OPE estimators in two standard OpenAI Gym tasks.

Replication code for all experiments is available at `http://github.com/CausalML/DoubleReinforcementLearningMDP`.

5.1. The Effects of Leveraging Memorylessness and of Double Robustness

In this section we consider an MDP with a horizon of $T = 30$, binary actions, univariate continuous state, initial state distribution $p(s_0) \sim \mathcal{N}(0.5, 0.2)$, transition probabilities $P_t(s_{t+1} \mid s_t, a_t) \sim \mathcal{N}(s + 0.3a - 0.15, 0.2)$. The target and behavior policies we consider are, respectively,

$$\pi^e(a \mid s) \sim \text{Bernoulli}(p_e),\ p_e = 0.2/(1 + \exp(-0.1s)) + 0.2U,\ U \sim \text{Uniform}[0, 1]$$
$$\pi^b(a \mid s) \sim \text{Bernoulli}(p_b),\ p_b = 0.9/(1 + \exp(-0.1s)) + 0.1U,\ U \sim \text{Uniform}[0, 1].$$

We assume the behavior policy is known. Note that this setting is an MDP and belongs to $\mathcal{M}_2$.

We compare five estimators: $\hat{\rho}_{\mathrm{IS}}$, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$, $\hat{\rho}_{\mathrm{DM}}$, $\hat{\rho}_{\mathrm{MIS}}$, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$ when nuisance functions $q_t(s,a)$ and $\mu_t(s)$ are estimated parametrically. We consider three settings:

(1) Both models correct: $q_t(s_t,a_t) = \beta_{1t}s_t + \beta_{2t}s_ta_t + \beta_{3t}$, $\mu_t(s_t,a_t) = \beta_{4t}s_t + \beta_{5t}s_ta_t + \beta_{6t}$.

(2) Only μ-model correct: $q_t(s_t,a_t) = \beta_{1t}s_t^2 + \beta_{2t}s_t^2a_t + \beta_{3t}$, $\mu_t(s_t,a_t) = \beta_{4t}s_t + \beta_{5t}s_ta_t + \beta_{6t}$.

(3) Only q-model correct: $q_t(s_t,a_t) = \beta_{1t}s_t + \beta_{2t}s_ta_t + \beta_{3t}$, $\mu_t(s_t,a_t) = \beta_{4t}s_t^2 + \beta_{5t}s_t^2a_t + \beta_{6t}$.

Note that in the above, the "correct" models are in fact not exactly correct because $\mathrm{E}_{\pi^e}[a_t \mid s_t]$ is actually nonlinear in s_t, but it is very nearly linear in the space of observed s_t values (for example, best linear fit for $\mathrm{E}_{\pi^e}[a_t \mid s_t]$ has an L^2 distance 3×10^{-5} on $[0,1]$, which spans ± 2.5 standard deviations for s_0). We therefore treat them as correctly specified.

In all cases, to estimate q-models we use backward-recursive ordinary least squares as in Section 4.2. To estimate μ-models we use ordinary least squares regression on λ_t (which is assumed known) as in Eq. (11).

For each $n = 1500, 3000, 4500$, we consider 50000 Monte Carlo replications. In each replication, we estimate the q- and μ-models as above and compute, for each setting, each of $\hat{\rho}_{\mathrm{IS}}$, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$, $\hat{\rho}_{\mathrm{DM}}$, $\hat{\rho}_{\mathrm{MIS}}$, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$. We report the RMSE of each estimator in each setting (and the standard error) in Table 1.

Our first immediate observation is that $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$ nearly dominates all other estimators, achieving similar or better performance in every setting and sample size. In particular, in settings (1) and (3), where the q-model is correct, it has performance similar to $\hat{\rho}_{\mathrm{DM}}$. Note that in settings (1) and (3), $\hat{\rho}_{\mathrm{DM}}$ is efficient for $\mathcal{M}_{2q}$ per Section 4.2 (or almost so; it would be efficient if we used efficient GMM instead of one-step GMM). In setting (1), $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$ is locally efficient, while in setting (3), it is only doubly robust and performs almost imperceptibly worse than the efficient $\hat{\rho}_{\mathrm{DM}}$.

In setting (2), where the q-model is incorrect, $\hat{\rho}_{\mathrm{DM}}$ is inconsistent and $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$ handily outperforms it. In the same setting (2), the consistent $\hat{\rho}_{\mathrm{IS}}$ and $\hat{\rho}_{\mathrm{MIS}}$ also outperform the inconsistent $\hat{\rho}_{\mathrm{DM}}$ but not by as much as $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$. While $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ is doubly robust in setting (2) guaranteeing consistency, unlike the case of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$, the combination of large (unmarginalized) cumulative density ratios and a misspecified q-model leads to still worse performance in the sample sizes tested.

Generally, $\hat{\rho}_{\mathrm{IS}}$, $\hat{\rho}_{\mathrm{MIS}}$, and $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ all have high RMSE due to the significant mismatch between the behavior and target policies so that cumulative density ratios are very large and only marginalizing them without also using a q-model helps only a little. In settings (1) and (2), where the μ-model is correct, $\hat{\rho}_{\mathrm{MIS}}$ improves on $\hat{\rho}_{\mathrm{IS}}$ only slightly, while in setting (3), where μ-model is incorrect, it performs significantly worse. This highlights the potential danger of misspecifying μ-models compared to the robustness of importance sampling with known behavior policy (see also Remark 12).

While both $\hat{\rho}_{\mathrm{IS}}$ and $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ remain consistent throughout all settings, they are outperformed by the also-consistent $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$, which leverages the MDP structure of $\mathcal{M}_2$ and exhibits local efficiency in setting (1) and doubly robustness in settings (2) and (3).

Table 2: Cliff Walking: RMSE (and standard errors)

Size	$\hat{\rho}_{\text{IS}}$	$\hat{\rho}_{\text{DRL}(\mathcal{M}_1)}$	$\hat{\rho}_{\text{DM}}$	$\hat{\rho}_{\text{MIS}}$	$\hat{\rho}_{\text{DRL}(\mathcal{M}_2)}$
500	18.8 (7.67)	3.78(1.14)	2.63 (0.01)	12.8 (4.96)	1.44 (0.29)
1000	7.99 (0.89)	0.28 (0.026)	1.27 (0.002)	5.92 (0.78)	0.22 (0.34)
1500	7.64 (1.63)	0.098 (0.013)	1.01 (0.001)	5.55 (1.10)	0.075 (0.008)

Table 3: Mountain Car: RMSE (and standard errors)

n	$\hat{\rho}_{\text{IS}}$	$\hat{\rho}_{\text{DRL}(\mathcal{M}_1)}$	$\hat{\rho}_{\text{DM}}$	$\hat{\rho}_{\text{MIS}}$	$\hat{\rho}_{\text{DRL}(\mathcal{M}_2)}$
500	6.85 (0.13)	3.72 (0.08)	4.30 (0.05)	6.82 (0.12)	3.53 (0.12)
1000	4.73 (0.07)	2.12 (0.04)	3.40 (0.008)	4.83 (0.06)	2.07 (0.04)
1500	3.41 (0.04)	1.82 (0.02)	3.30 (0.008)	3.40 (0.05)	1.69 (0.03)

5.2. Investigating Performance in RL Tasks: Cliff Walking and Mountain Car

We next compare the same OPE estimators using nonparametric nuisance estimation in two standard RL settings included in OpenAI Gym (Brockman et al., 2016): Cliff Walking and Mountain Car. For further detail on each setting, see Appendix C.

First, we used q-learning to learn an optimal policy for the MDP and define it as π^d. Then we generate the dataset from the behavior policy $\pi^b = (1-\alpha)\pi^d + \alpha\pi^u$ where π^u is a uniform random policy and $\alpha = 0.8$. We define the target policy similarly but with $\alpha = 0.9$. Again, we assume the behavior policy is known. Note that this π_d is fixed in each setting.

We estimate all μ-functions by first estimating w-functions and using Eq. (5). For Cliff Walking, we use a histogram estimator for w as in Eq. (7). For Mountain Car, we use a kernel estimator for w as in Eq. (10). We use the Epanechnikov kernel and choose an optimal bandwidth based on an L^2-risk criterion for $t = 1$; we then use this bandwidth for all other t values as well for simplicity. For q-functions, we use backward-recursive regression as in Section 4.2. For Cliff-Walking, we use a histogram model, $q(s,a;\beta) = \sum_{s_j,a_k \in \mathcal{S},\mathcal{A}} \beta_{jk} \mathbb{I}[s_j = s, a_k = a]$. For Mountain-Car, we use the mode $q(s,a;\beta) = \beta^\top \phi(s,a)$ where $\phi(s,a)$ is a 400-dimensional feature vector based on a radial basis function, generated using the `RBFSampler` method of `scikit-learn` based on Rahimi and Recht (2008).

We again compare $\hat{\rho}_{\text{IS}}$, $\hat{\rho}_{\text{DRL}(\mathcal{M}_1)}$, $\hat{\rho}_{\text{DM}}$, $\hat{\rho}_{\text{MIS}}$, $\hat{\rho}_{\text{DRL}(\mathcal{M}_2)}$. In each setting we consider varying evaluation dataset sizes and for each consider 1000 replications. We report the RMSE of each estimator in each setting (and the standard error) in Tables 2 and 3.

We again find that the performance of $\hat{\rho}_{\text{DRL}(\mathcal{M}_2)}$ is superior to all other estimators in either setting. This is especially true in Cliff Walking. The estimator $\hat{\rho}_{\text{DRL}(\mathcal{M}_2)}$ also improves upon $\hat{\rho}_{\text{IS}}$ and $\hat{\rho}_{\text{DM}}$ but not as much as $\hat{\rho}_{\text{DRL}(\mathcal{M}_2)}$. The estimator $\hat{\rho}_{\text{MIS}}$ offers a slight improvement over $\hat{\rho}_{\text{IS}}$, but is still outperformed by $\hat{\rho}_{\text{DRL}(\mathcal{M}_2)}$, $\hat{\rho}_{\text{DRL}(\mathcal{M}_1)}$, and $\hat{\rho}_{\text{DM}}$. That the improvement of $\hat{\rho}_{\text{MIS}}$ over $\hat{\rho}_{\text{IS}}$ and the overall improvements of $\hat{\rho}_{\text{DRL}(\mathcal{M}_2)}$ is starker in Cliff Walking than in Mountain Car may be attributable to the difficulty of learning w_t nonparametrically in a continuous state space.

6. Conclusions

We established the semiparametric efficiency bounds and efficient influence functions for OPE under either NMDP or MDP model, which quantify how fast one could hope to estimate policy value. While in the NMDP case, the influence function we derived has appeared frequently in OPE estimators, in the MDP case, the influence function is novel and has not appeared in existing estimators. Our results also suggested how one could construct efficient estimators. We used this to develop DRL, which used our newly derived efficient influence function, with nuisances estimated in a cross-fold manner. This ensured efficiency under very weak and mostly agnostic conditions on the nuisance estimation method used. Notably, DRL is the *first efficient OPE estimator for MDPs*. In addition, DRL enjoyed double robustness properties. This efficiency and robustness translated to better performance in experiments.

Acknowledgments

The authors thank Nan Jiang and Yu-Xiang Wang for helpful discussions. This material is based upon work supported by the National Science Foundation under Grant No. 1846210.

References

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71:1795–1843, 2003.

Chunrong Ai and Xiaohong Chen. Semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170:442–457, 2012.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33:1497–1537, 2005.

David Benkeser and Mark van der Laan. The highly adaptive lasso estimator. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, volume 2016, pages 689–696. IEEE, 2016.

Dimitri P Bertsekas. *Dynamic programming and optimal control*. Athena Scientific optimization and computation series. Athena Scientific, Belmont, Mass, 4th ed. edition, 2012.

Aurélien Bibaut and Mark van der Laan. Fast rates for empirical risk minimization over cdlg functions with bounded sectional variation norm. *arXiv preprint arXiv:1907.09244*, 2019a.

Aurelien Bibaut, Ivana Malenica, Nikos Vlassis, and Mark van der Laan. More efficient off-policy evaluation through regularized targeted learning. In *Proceedings of the 36th*

International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 654–663, 2019.

Aurélien F Bibaut and Mark J van der Laan. Fast rates for empirical risk minimization over c\adl\ag functions with bounded sectional variation norm. *arXiv preprint arXiv:1907.09244*, 2019b.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

Clive G. Bowsher and Peter S. Swain. Identifying sources of variation and the flow of information in biochemical networks. *Proceedings of the National Academy of Sciences*, 109, 2012.

G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, , and W Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Bibhas Chakraborty and EE Moodie. *Statistical methods for dynamic treatment regimes*. Springer, 2013.

Gary Chamberlain. Comment: Sequential moment restrictions in panel data. *Journal of Business & Economic Statistics*, 10:20–26, 1992.

Xiaohong Chen. Chapter 76 large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632, 2007.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68, 2018.

Iván Díaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 2019.

Miroslav Dudik, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29:485–511, 2014.

Ashkan Ertefaie and Robert L Strawderman. Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, 105:963–977, 2018.

M. Farajtabar, Y. Chow, and M. Ghavamzadeh. More robust doubly robust off-policy evaluation. *In Proceedings of the 35th International Conference on Machine Learning*, pages 1447–1456, 2018.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nat Med*, 25:16–18, 2019.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

J Hahn. Efficient estimation of panel data models with sequential moment restrictions. *Journal of Econometrics*, 79:1–21, 1997.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331, 1998.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14:262–280, 1996.

M.A. Hernan and J.M. Robins. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, 2019.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71:1161–1189, 2003.

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. *arXiv preprint arXiv:1802.04474*, 2018.

N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. *In Proceedings of the 33rd International Conference on International Conference on Machine Learning*, pages 652–661, 2016.

Nathan Kallus and Masatoshi Uehara. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. *In Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient and robust off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Khashayar Khosravi, Greg Lewis, and Vasilis Syrgkanis. Non-parametric inference adaptive to intrinsic dimension. *arXiv preprint arXiv:1901.03719*, 2019.

Chris A. J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, 15:1548–1562, 1987.

Michael R Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer New York, New York, NY, 2008.

Michail Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2004.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712, 2019.

L. Li, R. Munos, and C. Szepesvari. Toward minimax off-policy value estimation. *In Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 608–616, 2015.

Qi Li and Jeffrey Scott Racine. *Nonparametric econometrics : theory and practice*. Princeton University Press, Princeton, N.J., 2007.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31*, pages 5356–5366. 2018.

Daniel J. Luckett, Eric B. Laber, Anna R. Kahkoska, David M. Maahs, Elizabeth Mayer-Davis, and Michael R. Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, pages 1–34, 2018.

A. Rupam Mahmood, Hado P van Hasselt, and Richard S Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems 27*, pages 3014–3022. 2014.

T. Mandel, Y. Liu, S. Levine, E. Brunskill, and Z Popovic. Off-policy evaluation across representations with applications to educational games. *In Proceedings of the 13th International Conference on Autonomous Agentsand Multi-agent Systems*, pages 1077–1084, 2014.

Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

Remi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems 29*, pages 1054–1062. 2016.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65:331–355, 2003.

S A Murphy, M J van der Laan, and J M Robins. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96:1410–1423, 2001.

W. K. Newey and D. L Mcfadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, IV:2113–2245, 1994.

D. Precup, R. Sutton, and S Singh. Eligibility traces for off-policy policy evaluation. *In Proceedings of the 17th International Conference on Machine Learning*, pages 759–766, 2000.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. 2008.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

James M Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000.

James M Robins, Miguel Angel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:551, 2000.

Andrea Rotnitzky and Stijn Vansteelandt. Double-robust methods. In *Handbook of missing data methodology. In Handbooks of Modern Statistical Methods*, pages 185–212. Chapman and Hall/CRC, 2014.

Andrea Rotnitzky, Ezequiel Smucler, and James Robins. Characterization of parameters with a mixed bias property. *arXiv preprint arXiv:1509.02556*, 2019.

D. Scharfstein, A. Rotnizky, and J. M. Robins. Adjusting for nonignorable dropout using semi-parametric models. *Journal of the American Statistical Association*, 94:1096–1146, 1999.

Xiaotong Shen. On methods of sieves and penalization. *The Annals of Statistics*, 25: 2555–2591, 1997.

Charles J. Stone. Rejoinder: The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 22:179–184, 1994.

Richard S Sutton. *Reinforcement learning : an introduction*. MIT Press, Cambridge, Mass., 2018.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems 28*, pages 3231–3239. 2015.

P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. *In Proceedings of the 33rd International Conference on Machine Learning*, pages 2139–2148, 2016.

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

Anastasios A Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, New York, NY, 2006.

Mark J. van der Laan and James M Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics,. Springer New York, New York, NY, 2003.

A. W van der Vaart. On differentiable functionals. *Ann. Statist.*, 19:178–204, 03 1991.

A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK, 1998.

A. W. van der Vaart. *Semiparametric Statistics*. Lecture Notes in Mathematics ; 1781. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

Karel Vermeulen. Semiparametric efficiency. *Gent, Faculteit WetenschappenVakgroep Toegepaste Wiskunde en Informatica*, 2010.

Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2016.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3589–3597, 2017.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems 32*, pages 9665–9675. 2019.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. *In Proceedings of the 23nd International Workshop on Artificial Intelligence and Statistics (To appear)*, 2020.

Baqun Zhang, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100: 681–694, 2013.

Wenjing Zheng and Mark J van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, pages 459–474. Springer New York, New York, NY, 2011.

Appendix A. Notation

We first summarize the notation we use in Table 4 and the abbreviations we use in Table 5. Notice in particular that, following empirical process theory literature, in the proofs we also use $\mathbb{P}$ to denote expectations (interchangeably with E).

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 $\mathcal{M}_1, \mathcal{M}_2$ respectively
$q_t = q_t(\mathcal{H}_{a_t})$, $q_t(s_t, a_t)$	q-function at t, in $\mathcal{M}_1, \mathcal{M}_2$ respectively
λ_t	Cumulative density ratio $\prod_{k=0}^t \pi_t^e/\pi_t^b$
μ_t	Marginal density ratio $\mathrm{E}[\lambda_t \mid s_t, a_t]$
η_t	Instantaneous density ratio π_t^e/π_t^b
Λ	Tangent space
$\mathcal{M}$	A model for the data generating distribution
$\mathcal{M}_1, \mathcal{M}_{1b}, \mathcal{M}_{1q}$	NMDP model with unknown behavior policy, known behavior policy, and parametric q-function, respectively
$\mathcal{M}_2, \mathcal{M}_{2b}, \mathcal{M}_{2q}$	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 $\sqrt{n}(\mathbb{P}_n - \mathbb{P})$
Asmse[·], var[·]	Asymptotic variance, variance
$\mathcal{N}(a, b)$	Normal distribution with mean a and variance b
$\mathrm{Uni}[a, b]$	Uniform distribution on $[a, b]$
$A_n = \mathrm{o}_p(a_n)$	The term A_n/a_n converges to zero in probability
$A_n = \mathrm{O}_p(a_n)$	The term A_n/a_n is bounded in probability
Λ_d^α	Hölder space with smoothness α with a dimension d

Table 5: Abbreviations

NMDP	Non-Markov Decision Process
MDP	Markov Decision Process
RL	Reinforcement Learning
CB	Contextual Bandit
OPE	Off policy Evaluation
MLE	Maximum Likelihood Estimation
RAL	Regular and Asymptotic Linear
CAN	Consistent and Asymptotically Normal
MSE	Mean Squared Error

Appendix B. Proofs

Before going into details of the proof, we summarize definitions and proofs to derive a semiparametric lower bound. As we mentioned in Section 1.2, for a complete and rigorous treatment, refer to Bickel et al. (1998); van der Laan and Robins (2003); van der Vaart (2002). Additional accessible treatments are also given in (Bibaut and van der Laan, 2019a; Tsiatis, 2006; Vermeulen, 2010)

B.1. Semiparametric theory

We overload notation on Section 1.2. 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 24 (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 25 (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 26 (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 27 (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 28. *Under the same assumptions of Theorem 27,*

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

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 29 (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.

B.2. Proof

Proof of Theorem 1.

Efficient influence function under $\mathcal{M}_1$. The entire regular (regular model as defined in Chapter 7 van der Vaart, 1998) parametric submodel under $\mathcal{M}_1$ is

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

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

The score function of the model $\mathcal{M}_1$ is decomposed as

$$\begin{aligned} g(\mathcal{J}_{s_T}) &= \sum_{k=0}^{T} \nabla \log p_\theta(s_k|\mathcal{H}_{a_{k-1}}) + \sum_{k=0}^{T} \nabla \log p_\theta(a_k|\mathcal{H}_{s_k}) + \sum_{k=0}^{T} \nabla \log p_\theta(r_k|\mathcal{H}_{a_k}) \\ &= \sum_{k=0}^{T} g_{s_k|\mathcal{H}_{a_{k-1}}} + \sum_{k=0}^{T} g_{a_k|\mathcal{H}_{s_k}} + \sum_{k=0}^{T} g_{r_k|\mathcal{H}_{a_k}}. \end{aligned}$$

We first calculate a derivative for the target functional w.r.t the model $\mathcal{M}_1$. Note that this derivative is not unique. We have

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

This concludes that the following function is a derivative:

$$
\rho_{\mathrm{eff}}^{\mathcal{M}_1} = -\rho^{\pi^e} + \sum_{c=0}^{T}\left\{\lambda_c r_c - \lambda_c \sum_{t=c}^{T} \mathrm{E}_{\pi_e}(r_t|\mathcal{H}_{a_c}) - \lambda_{c-1} \sum_{t=c}^{T} \mathrm{E}_{\pi_e}(r_t|\mathcal{H}_{s_c})\right\}. \tag{16}
$$

Next, we show that this derivative is the efficient influence function. In order to show this, we calculate the tangent space of model $\mathcal{M}_1$. The tangent space of the model $\mathcal{M}_1$ 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

$$
\bigoplus_{0\le t\le T} (A'_t \bigoplus B'_t \bigoplus C'_t) \tag{17}
$$

such that

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

More specifically, we have the following lemma.

Lemma 30. *The orthogonal tangent space is represented as*

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

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. (16) is orthogonal to the orthogonal tangent space Eq. (17), we can see that the above derivative is actually the efficient influence function under the model $\mathcal{M}_1$. This fact is shown as follows.

Lemma 31. *The derivative Eq.* (16) *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 $t(\mathcal{J}_{s_k}) \in A'_k$

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

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

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

$$= \mathrm{E}\left[\left\{\left\{\lambda_k \sum_{t=k}^{T} \mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{a_k}]\right\} - \left\{\lambda_k \sum_{t=k}^{T} \mathrm{E}_{\pi^e}[r_t|\mathcal{H}_{a_k}]\right\}\right\} t(\mathcal{J}_{a_k})\right] = 0.$$

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

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

This concludes the proof for $\mathcal{M}_1$.

Efficient influence function under $\mathcal{M}_{1b}$. 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

$$\bigoplus_{0 \le t \le T} (A_t' \bigoplus B_t'' \bigoplus C_t') \tag{18}$$

This space Eq. (18) is orthogonal to the obtained efficient influence function under $\mathcal{M}_1$. Therefore, the efficient influence function under $\mathcal{M}_{1b}$ is the same as the one under $\mathcal{M}_b$.

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

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

$$= \sum_{t=0}^{T+1} \mathrm{E}\left[\lambda_{t-1}^2 \mathrm{var}\left(r_{t-1} + v_t(\mathcal{H}_{s_t})|\mathcal{H}_{a_{t-1}}\right)\right].$$

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

Proof of Theorem 2. The proof is analogous to the proof of Theorem 1. □

Proof of Theorem 3.

Efficient influence function under $\mathcal{M}_2$. 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

$$\begin{aligned} g(\mathcal{J}_{s_T}) &= \sum_{k=0}^{T} \nabla_\theta \log p_\theta(s_k \mid s_{k-1}, a_{k-1}) + \nabla_\theta \log p_\theta(a_{k+1} \mid s_k) + \nabla_\theta \log p_\theta(r_k \mid s_k, a_k) \\ &= \sum_{k=0}^{T} g_{s_k|s_{k-1},a_{k-1}} + \sum_{k=0}^{T} g_{a_{k+1}|s_k} + \sum_{k=0}^{T} g_{r_k|s_k,a_k}. \end{aligned}$$

We first calculate a derivative of the target functional w.r.t the model $\mathcal{M}_2$. Note that this derivative is not only derivative. We have

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

Therefore, the following function is a derivative:

$$-\rho^{\pi_c^e} + \sum_{c=0}^{T} \frac{p_{\pi_c^e}(s_c, a_c)}{p_{\pi_c^b}(s_c, a_c)} r_c - \left\{ \frac{p_{\pi_c^e}(s_c, a_c)}{p_{\pi_c^b}(s_c, a_c)} \sum_{t=c}^{T} \mathrm{E}_{\pi_e}[r_t | s_c, a_c] - \frac{p_{\pi_c^e}(s_{c-1}, a_{c-1})}{p_{\pi_c^b}(s_{c-1}, a_{c-1})} \sum_{t=c}^{T} \mathrm{E}_{\pi_e}[r_t | s_c] \right\}. \tag{19}$$

We will show this derivative is the efficient influence function.

In order to show this, we calculate the tangent space of model $\mathcal{M}_2$. The tangent space of the model $\mathcal{M}_2$ 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

$$\bigoplus_{0 \le t \le T} (A'_t \bigoplus B'_t \bigoplus C'_t), \tag{20}$$

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, 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 derivative Eq. (19) is orthogonal to the orthogonal tangent space Eq. (20), we can see that the above derivative is actually the efficient influence function under the model $\mathcal{M}_2$. This fact is shown as follows.

Lemma 32. *The derivative Eq.* (19) *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. (19) is orthogonal to A'_k; for $t(\mathcal{J}_{s_k}) \in A'_k$

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

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

Second, the influence function Eq. (19) 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. (19) is orthogonal to C'_k; for $t(\mathcal{J}_{r_k}) \in C'_k$

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

□

Efficient influence function under $\mathcal{M}_{2b}$. In Theorem 32, we check that the $\phi_{\mathrm{eff}}^{\mathcal{M}2}$ is orthogonal to $B"_t$. This concludes the proof noting that the orthogonal tangent space of $\mathcal{M}_{2b}$ 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 and Swain, 2012). Recall that we can also easily derive this variance form using another equivalent form of efficient influence function.

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

$$
\begin{aligned}
&= \sum_{t=0}^{T+1} \mathrm{E}\left[\mathrm{var}\left[\mathrm{E}\left[\mu_{t-1}(s_{t-1}, a_{t-1})(r_{t-1} + v_t - q_{t-1})|\mathcal{J}_{a_t}\right]|\mathcal{J}_{a_{t-1}}\right]\right] \\
&= \sum_{t=0}^{T+1} \mathrm{E}\left[\mathrm{var}\left[\mu_{t-1}(s_{t-1}, a_{t-1})(r_{t-1} + v_t - q_{t-1})|\mathcal{J}_{a_{t-1}}\right]\right] \\
&= \sum_{t=0}^{T+1} \mathrm{E}\left[\mu_{t-1}^2(s_{t-1}, a_{t-1})\mathrm{var}\left[(r_{t-1} + v_t(s_t))|\mathcal{J}_{a_{t-1}}\right]\right].
\end{aligned}
$$

□

Proof of Theorem 4. From Jensen's inequality,

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

When λ_{t-1}^2 is not constant given s_{t-1}, a_{t-1} and $\mathrm{var}\left\{r_t + v_t(s_t)|s_{t-1}, a_{t-1}\right\} \neq 0$, 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

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

The last equality follows by the law of total variance.

Similarly, the efficiency bound under MDP is

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

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

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. Without loss of generality, we consider the case $K = 2$. Define $\phi(\{\hat{\lambda}_k\}, \{\hat{q}_k\})$ as:

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

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

$$
\frac{n_1}{n} \mathbb{P}_{n_1} \phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) + \frac{n_2}{n} \mathbb{P}_{n_2} \phi(\{\hat{\lambda}_k^{(2)}\}, \{\hat{q}_k^{(2)}\}),
$$

where $\mathbb{P}_{n_1}$ is an empirical approximation based on a set of samples such that $i \in \mathcal{D}_1$, $\mathbb{P}_{n_2}$ is an empirical approximation based on a set of samples such that $i \in \mathcal{D}_2$. Then, we have

$$
\begin{aligned}
\sqrt{n}(\mathbb{P}_{n_1} \phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) &= \sqrt{n/n_1} \mathbb{G}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k\}, \{q_k\})] && (21) \\
&+ \sqrt{n/n_1} \mathbb{G}_{n_1}[\phi(\{\lambda_k^{(1)}\}, \{q_k^{(1)}\})] && (22) \\
&+ \sqrt{n}(\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) | \{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] - \rho^{\pi^e}). && (23)
\end{aligned}
$$

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

$$
\phi(\{\hat{\lambda}_k\}, \{\hat{q}_k\}) - \phi(\{\lambda_k\}, \{q_k\}) = D_1 + D_2 + D_3, \quad \text{where}
$$

$$D_1 = \sum_{k=0}^{T} (\hat{\lambda}_k - \lambda_k)(-\hat{q}_k + q_k) + (\hat{\lambda}_{k-1} - \lambda_{k-1})(\hat{v}_k - v_k),$$
$$D_2 = \sum_{k=0}^{T} \lambda_k(-\hat{q}_k + q_k) + \lambda_{k-1}(\hat{v}_k - v_k),$$
$$D_3 = \sum_{k=0}^{T} (\hat{\lambda}_k - \lambda_k)(r_k - q_k + v_{k+1}).$$

First, we show the term Eq. (21) is $\mathrm{o}_p(1)$.

Lemma 33. *The term Eq.* (21) *is* $\mathrm{o}_p(1)$.

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

$$\lim_{n\to\infty} \sqrt{n_1} P[\mathbb{P}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k\}, \{q_k\})] \\ - \mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k\}, \{q_k\})|\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon|\mathcal{D}_2] = 0, \tag{24}$$

Then, by bounded convergence theorem, we would have

$$\lim_{n\to\infty} \sqrt{n_1} P[\mathbb{P}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k\}, \{q_k\})] \\ - \mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k\}, \{q_k\})|\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon] = 0,$$

yielding the statement.

To show Eq. (24), we show that the conditional mean is 0 and conditional variance is $\mathrm{o}_p(1)$. The conditional mean is

$$\mathrm{E}[\mathbb{P}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k\}, \{q_k\})|\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] \\ - \mathbb{P}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k\}, \{q_k\})]|\mathcal{D}_2] = 0.$$

Here, we leveraged the sample splitting construction, that is, $\hat{\lambda}_k^{(1)}$ and $\hat{q}_k^{(1)}$ only depend on $\mathcal{D}_2$. The conditional variance is

$$\begin{aligned}
&\mathrm{var}[\sqrt{n_1}\mathbb{P}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k\}, \{q_k\})]|\mathcal{D}_2] \\
&\quad = \mathrm{E}[\mathrm{E}[D_1^2 + D_2^2 + D_3^2 + 2D_1D_2 + 2D_2D_3 + 2D_2D_3|\{\hat{q}_k^{(1)}\}, \{\lambda_k^{(1)}\}]|\mathcal{D}_2] \\
&\quad = \mathrm{o}_p(1).
\end{aligned}$$

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

$$\begin{aligned}
&\sqrt{n_1} P[\mathbb{P}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k\}, \{q_k\})] - \mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) \\
&\qquad - \phi(\{\lambda_k\}, \{q_k\})|\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon|\mathcal{D}_2] \\
&\le \frac{1}{\epsilon^2}\mathrm{var}[\sqrt{n_1}\mathbb{P}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k\}, \{q_k\})]|\mathcal{D}_2] = \mathrm{o}_p(1).
\end{aligned}$$

□

Lemma 34. *The term Eq.* (23) *is* $\mathrm{o}_p(1)$.

Proof.

$$\begin{aligned}
&\sqrt{n}\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \mathrm{E}[\phi(\{\lambda_k\},\{q_k\})]|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\lambda}_k^{(1)} - \lambda_k)(-\hat{q}_k^{(1)} + q_k) + (\hat{\lambda}_{k-1}^{(1)} - \lambda_{k-1})(-\hat{v}_k^{(1)} + v_k)|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}\lambda_k(-\hat{q}_k^{(1)} + q_k) + \lambda_{k-1}(\hat{v}_k^{(1)} - v_k)|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\lambda}_k^{(1)} - \lambda_k)(r_k - q_k + v_{k+1})|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\lambda}_k^{(1)} - \lambda_k)(-\hat{q}_k^{(1)} + q_k) + (\hat{\lambda}_{k-1}^{(1)} - \lambda_{k-1})(-\hat{v}_k + v_k)|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\sum_{k=0}^{T}\mathrm{O}(\|\hat{\lambda}_k^{(1)} - \lambda_k\|_2\|\hat{q}_k^{(1)} - q_k\|_2) = \sqrt{n}\sum_{k=0}^{T}\mathrm{o}_p(n^{-1/2}) = \mathrm{o}_p(1). \qquad \square
\end{aligned}$$

Finally, we get

$$\sqrt{n}(\mathbb{P}_{n_1}\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) = \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\lambda_k\},\{q_k\})] + \mathrm{o}_p(1).$$

Therefore,

$$\begin{aligned}
&\sqrt{n}(\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_1)} - \rho^{\pi^e}) \\
&= n_1/n\sqrt{n}\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) + n_2/n\sqrt{n}(\mathbb{P}_{n_2}\phi(\{\hat{\lambda}_k^{(2)}\},\{\hat{q}_k^{(2)}\}) - \rho^{\pi^e}) \\
&= \sqrt{n_1/n}\mathbb{G}_{n_1}[\phi(\{\lambda_k\},\{q_k\})] + \sqrt{n_2/n}\mathbb{G}_{n_2}[\phi(\{\lambda_k\},\{q_k\})] + \mathrm{o}_p(1) \\
&= \mathbb{G}_n[\phi(\{\lambda_k\},\{q_k\})] + \mathrm{o}_p(1),
\end{aligned}$$

concluding the proof by showing the influence function of $\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_1)}$ is the efficient one. $\square$

Proof of Theorem 7. Here, $a \lessapprox b$ means there exists constant Q not depending on $T, R_{\max}, C, n, C_1, C_2$ such that $a < Qb$.

Define $\phi(\{\hat{\lambda}_k\},\{\hat{q}_k\})$ as:

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

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

$$\frac{n_1}{n}\mathbb{P}_{n_1}\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) + \frac{n_2}{n}\mathbb{P}_{n_2}\phi(\{\hat{\lambda}_k^{(2)}\},\{\hat{q}_k^{(2)}\}),$$

where $\mathbb{P}_{n_1}$ is an empirical approximation based on a set of samples such that $i \in \mathcal{D}_1$, $\mathbb{P}_{n_2}$ is an empirical approximation based on a set of samples such that $i \in \mathcal{D}_2$.

Then, we have

$$(\mathbb{P}_{n_1}\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\rho^{\pi^e}) = \sqrt{1/n_1}\mathbb{G}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\phi(\{\lambda_k\},\{q_k\})] \tag{25}$$
$$+\sqrt{1/n_1}\mathbb{G}_{n_1}[\phi(\{\lambda_k\},\{q_k\})] \tag{26}$$
$$+(\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}]-\rho^{\pi^e}). \tag{27}$$

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

$$\phi(\{\hat{\lambda}_k\},\{\hat{q}_k\})-\phi(\{\lambda_k\},\{q_k\}) = D_1+D_2+D_3, \quad \text{where}$$
$$D_1 = \sum_{k=0}^{T}(\hat{\lambda}_k-\lambda_k)(-\hat{q}_k+q_k)+(\hat{\lambda}_{k-1}-\lambda_{k-1})(\hat{v}_k-v_k),$$
$$D_2 = \sum_{k=0}^{T}\lambda_k(-\hat{q}_k+q_k)+\lambda_{k-1}(\hat{v}_k-v_k),$$
$$D_3 = \sum_{k=0}^{T}(\hat{\lambda}_k-\lambda_k)(r_k-q_k+v_{k+1}).$$

Lemma 35. *With probability $1-2\delta$, the absolute value of the term Eq. (25) is bounded by*

$$\sqrt{\frac{\log(2/\delta)T^2(TR_{\max}C^{T+1}\sqrt{\kappa_1\kappa_2}+\kappa_1T^2R_{\max}^2+\kappa_2C^{2(T+1)})}{n}}+\frac{\log(2/\delta)TR_{\max}C^{T+1}}{n},$$

up to some constant independent of $n, C, R_{\max}, T$.

Proof. From Bernstein inequality, with probability $1-\delta$,

$$P[|\mathbb{P}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\phi(\{\lambda_k\},\{q_k\})]$$
$$-\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\phi(\{\lambda_k\},\{q_k\})|\mathcal{D}_1]|>\epsilon]|\mathcal{D}_2]$$
$$\le 2\exp\left(-\frac{0.5n\epsilon^2}{\mathrm{E}[\{\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\phi(\{\lambda_k\},\{q_k\})\}^2|\mathcal{D}_2]+Q_1TR_{\max}C^{T+1}\epsilon}\right),$$

noting the conditional mean is 0;

$$\mathrm{E}[\mathbb{P}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\phi(\{\lambda_k\},\{q_k\})|\mathcal{D}_1]$$
$$-\mathbb{P}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\phi(\{\lambda_k\},\{q_k\})]|\mathcal{D}_2]=0.$$

Here, Q_1 is some constant independent of $n, C, R_{\max}, T, \delta$.

With probability $1-\delta$, the conditional variance is

$$\mathrm{E}[\{\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\phi(\{\lambda_k\},\{q_k\})\}^2|\mathcal{D}_2]$$
$$=\mathrm{E}[D_1^2+D_2^2+D_3^2+2D_1D_2+2D_2D_3+2D_2D_3|\mathcal{D}_2]$$
$$\lessapprox T^2(TR_{\max}C^{T+1}\sqrt{\kappa_1\kappa_2}+\kappa_1(TR_{\max})^2+\kappa_2C^{2(T+1)}).$$

Therefore, with probability $1-2\delta$,

$$P[|\mathbb{P}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\phi(\{\lambda_k\},\{q_k\})] - \mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\phi(\{\lambda_k\},\{q_k\})|\mathcal{D}_1]|>\epsilon]|\mathcal{D}_2] \leq 2\exp\left(-\frac{0.5n_1\epsilon^2}{Q_2T^2(TR_{\max}C^{T+1}\sqrt{\kappa_1\kappa_2}+\kappa_1T^2R_{max}^2+\kappa_2C^{2(T+1)})+Q_1TR_{\max}C^{T+1}\epsilon}\right).$$

Here, Q_2 is some constant independent of $n, C, R_{\max}, T, \delta$. Then, by the law of total probability and some algebra, the statement is concluded. □

Lemma 36. *With probability $1-\delta$, the absolute value of the term Eq. (27) is bounded by*

$$T\sqrt{\kappa_1\kappa_2}.$$

up to some constant independent of $n, C, R_{\max}, T, \delta$.

Proof. Here, we have

$$\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\mathrm{E}[\phi(\{\lambda_k\},\{q_k\})]|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \leq \sum_{k=0}^{T} 2\|\hat{\lambda}_k^{(1)}-\lambda_k\|_2\|\hat{q}_k^{(1)}-q_k\|_2.$$

Then, with probability $1-\delta$, this is bounded by $T\sqrt{\kappa_1\kappa_2}$. □

Combining all results so far, with probability $1-6\delta$, we have

$$\left|\frac{n_1}{n}\mathbb{P}_{n_1}\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})+\frac{n_2}{n}\mathbb{P}_{n_2}\phi(\{\hat{\lambda}_k^{(2)}\},\{\hat{q}_k^{(2)}\})-\rho^{\pi_e}\right|$$
$$< Q_3T\sqrt{\kappa_1\kappa_2}+Q_2\sqrt{\frac{\log(2/\delta)T^2(TR_{\max}C^{T+1}\sqrt{\kappa_1\kappa_2}+\kappa_1T^2R_{\max}^2+\kappa_2C^{2(T+1)})}{n}}+\frac{Q_1\log(2/\delta)TR_{\max}C^{T+1}}{n}$$
$$+\mathbb{P}_{n_1}[\phi(\{\lambda_k\},\{q_k\})]+\mathbb{P}_{n_2}[\phi(\{\lambda_k\},\{q_k\})]-\rho^{\pi_e}.$$

Here, Q_3 is some constant independent of $n, C, R_{\max}, T$. Noting $|\mathbb{P}_n[\phi(\{\lambda_k\},\{q_k\})]-\rho^{\pi_e}|$ is bounded by

$$\sqrt{\frac{2\log(2/\delta)\mathrm{Effbd}(\mathcal{M}_1)}{n}}+\frac{Q_1\log(2/\delta)TR_{\max}C^{T+1}}{n}$$

with probability $1-\delta$, the statement is concluded. □

Proof of Theorem 9. We define $\phi(\{\hat{\lambda}_k\},\{\hat{q}_k\})$ as:

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

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

$$\sqrt{n}(\mathbb{P}_n\phi(\{\hat{\lambda}_k\},\{\hat{q}_k\})-\rho^{\pi^e})=\mathbb{G}_n[\phi(\{\hat{\lambda}_k\},\{\hat{q}_k\})-\phi(\{\lambda_k\},\{q_k\})] \tag{28}$$

$$+ \mathbb{G}_n[\phi(\{\lambda_k\},\{q_k\})] \tag{29}$$
$$+ \sqrt{n}(\mathrm{E}[\phi(\{\hat{\lambda}_k\},\{\hat{q}_k\})|\{\hat{\lambda}_k\},\{\hat{q}_k\}] - \rho^{\pi^e}). \tag{30}$$

If we can prove that the term Eq. (28) is $\mathrm{o}_p(1)$, the statement is concluded as in the proof of Theorem 6. We proceed to prove this.

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

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

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

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

as in Lemma 33. Therefore, from Lemma 19.24 in van der Vaart (1998), the term Eq. (28) is $\mathrm{o}_p(1)$, concluding the proof. □

Proof of Theorem 10. Without loss of generality, we consider the case $K = 2$.

We use the following doubly robust structure

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

Then, as in the proof of Theorem 6,

$$\begin{aligned}
&\sqrt{n}(\mathbb{P}_{n_1}\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) \\
&= \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})] + \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})] \\
&+ \sqrt{n/n_1}(\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] - \mathrm{E}[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})]) + \sqrt{n}(\mathrm{E}[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})] - \rho^{\pi^e}) \\
&= \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})] + \sqrt{n}(\mathrm{E}[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})] - \rho^{\pi^e}) + \mathrm{o}_p(1).
\end{aligned}$$

Here, we used

$$\sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})] = \mathrm{o}_p(1)$$

from Theorem 33 and

$$\sqrt{n/n_1}(\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] - \mathrm{E}[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})]) = \mathrm{o}_p(1),$$

which we prove below as in Theorem 34.

Lemma 37.

$$\sqrt{n/n_1}(\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] - \mathrm{E}[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})]) = \mathrm{o}_p(1).$$

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

$$
\begin{aligned}
&\sqrt{n}\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \mathrm{E}[\phi(\{\lambda_k\},\{q_k^\dagger\})]|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\lambda}_k^{(1)} - \lambda_k)(-\hat{q}_k^{(1)} + q_k^\dagger) + (\hat{\lambda}_{k-1}^{(1)} - \lambda_{k-1})(-\hat{v}_k + v_k^\dagger)|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}\lambda_k(\hat{q}_k^{(1)} - q_k^\dagger) + \lambda_{k-1}(\hat{v}_k - v_k^\dagger)|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\lambda}_k^{(1)} - \lambda_k)(r_k - q_k^\dagger + v_{k+1}^\dagger)|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\lambda}_k^{(1)} - \lambda_k)(-\hat{q}_k^{(1)} + q_k^\dagger) + (\hat{\lambda}_{k-1}^{(1)} - \lambda_{k-1})(-\hat{v}_k + v_k^\dagger)|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\lambda}_k^{(1)} - \lambda_k)(r_k - q_k^\dagger + v_{k+1}^\dagger)|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\sum_{k=0}^{T}\mathrm{O}(\|\hat{\lambda}_k^{(1)} - \lambda_k\|_2\|\hat{q}_k^{(1)} - q_k^\dagger\|_2 + \|\hat{\lambda}_k^{(1)} - \lambda_k\|_2) \\
&= \sqrt{n}\sum_{k=0}^{T}\left\{\mathrm{O}_p(n^{-\alpha_1})\mathrm{O}_p(n^{-\alpha_2}) + \mathrm{O}_p(n^{-\alpha_1})\right\} = \mathrm{O}_p(1).
\end{aligned}
$$

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

$$
\begin{aligned}
&\sqrt{n}\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \mathrm{E}[\phi(\{\lambda_k^\dagger\},\{q_k\})]|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}(\hat{\lambda}_k^{(1)} - \lambda_k^\dagger)(-\hat{q}_k^{(1)} + q_k) + (\hat{\lambda}_{k-1}^{(1)} - \lambda_{k-1}^\dagger)(-\hat{v}_k + v_k)|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&+ \sqrt{n}\mathrm{E}[\sum_{k=0}^{T}\lambda_k^\dagger(\hat{q}_k^{(1)} - q_k) + \lambda_{k-1}^\dagger(\hat{v}_k - v_k)|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&= \sqrt{n}\sum_{k=0}^{T}\mathrm{O}(\|\hat{\lambda}_k^{(1)} - \lambda_k^\dagger\|_2\|\hat{q}_k^{(1)} - q_k\|_2 + \|\hat{q}_k^{(1)} - q_k\|_2) \\
&= \sqrt{n}\sum_{k=0}^{T}\left\{\mathrm{O}_p(n^{-\alpha_1})\mathrm{O}_p(n^{-\alpha_2}) + \mathrm{O}_p(n^{-\alpha_2})\right\} = \mathrm{O}_p(1).
\end{aligned}
$$

□

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

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

$$
\mathrm{E}[\phi(\{\lambda_k\},\{q_k^\dagger\})] = \mathrm{E}[\sum_{k=0}^{T}[\lambda_k r_k - \{\lambda_k q_k^\dagger(\mathcal{H}_{a_k}) - \lambda_{k-1}\mathrm{E}_{\pi^e}[q_k^\dagger(\mathcal{H}_{a_k})|s_k]]
$$

$$= \mathrm{E}[\sum_{k=0}^{T} \lambda_k r_k] = \rho^{\pi^e}.$$

Then,

$$\sqrt{n}(\mathbb{P}_{n_1}\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) = \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\lambda_k\}, \{q_k^\dagger\})] + \mathrm{O}_p(1).$$

Therefore,

$$\begin{aligned}\sqrt{n}(\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_1)} - \rho^{\pi^e}) &= \sqrt{n_1/n}\mathbb{G}_{n_1}[\phi(\{\lambda_k\}, \{q_k^\dagger\})] + \sqrt{n_2/n}\mathbb{G}_{n_2}[\phi(\{\lambda_k\}, \{q_k^\dagger\})] + \mathrm{O}_p(1) \\ &= \mathbb{G}_n[\phi(\{\lambda_k\}, \{q_k^\dagger\})] + \mathrm{O}_p(1) = \mathrm{O}_p(1),\end{aligned}$$

which shows $\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_1)}$ is $\sqrt{n}$-consistent around ρ^{π^e} when the model for the behavior policy is well-specified.

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

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

Then,

$$\sqrt{n}(\mathbb{P}_{n_1}\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) = \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\lambda_k^\dagger\}, \{q_k\})] + \mathrm{O}_p(1).$$

Therefore,

$$\begin{aligned}\sqrt{n}(\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_1)} - \rho^{\pi^e}) &= \sqrt{n_1/n}\mathbb{G}_{n_1}[\phi(\{\lambda_k^\dagger\}, \{q_k\})] + \sqrt{n_2/n}\mathbb{G}_{n_2}[\phi(\{\lambda_k^\dagger\}, \{q_k\})] + \mathrm{O}_p(1) \\ &= \mathbb{G}_n[\phi(\{\lambda_k^\dagger\}, \{q_k\})] + \mathrm{O}_p(1) = \mathrm{O}_p(1).\end{aligned}$$

which shows $\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_1)}$ is $\sqrt{n}$-consistent around ρ^{π^e} when the model for the q-function is well-specified. □

Proof of Theorem 11. Without loss of generality, we consider the case $K = 2$.

Then, as in the proof of Theorem 6,

$$\begin{aligned}&\mathbb{P}_{n_1}\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \rho^{\pi^e} \\ &= \sqrt{1/n_1}\mathbb{G}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k^\dagger\}, \{q_k^\dagger\})] + \sqrt{1/n_1}\mathbb{G}_{n_1}[\phi(\{\lambda_k^\dagger\}, \{q_k^\dagger\})] \\ &+ \sqrt{1/n_1}(\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\})|\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] - \mathrm{E}[\phi(\{\lambda_k^\dagger\}, \{q_k^\dagger\})]) + (\mathrm{E}[\phi(\{\lambda_k^\dagger\}, \{q_k^\dagger\})] - \rho^{\pi^e}) \\ &= \sqrt{1/n_1}\mathbb{G}_{n_1}[\phi(\{\lambda_k^\dagger\}, \{q_k^\dagger\})] + (\mathrm{E}[\phi(\{\lambda_k^\dagger\}, \{q_k^\dagger\})] - \rho^{\pi^e}) + \mathrm{o}_p(1) \\ &= (\mathbb{P}_{n_1} - \mathbb{P})[\phi(\{\lambda_k^\dagger\}, \{q_k^\dagger\})] + \mathrm{o}_p(1).\end{aligned}$$

Here, under the assumption, we use the following equations:

$$\sqrt{1/n_1}\mathbb{G}_{n_1}[\phi(\{\hat{\lambda}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\lambda_k^\dagger\}, \{q_k^\dagger\})] = \mathrm{o}_p(1)$$

$$(\mathrm{E}[\phi(\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\})|\{\hat{\lambda}_k^{(1)}\},\{\hat{q}_k^{(1)}\}]-\mathrm{E}[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})])=\mathrm{o}_p(1)$$
$$(\mathrm{E}[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})]-\rho^{\pi^e})=0.$$

These equations are proved as in the proof of Theorem 10. Then,

$$\begin{aligned}(\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_1)}-\rho^{\pi^e}) &= (\mathbb{P}_{n_1}-\mathbb{P})[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})]+(\mathbb{P}_{n_2}-\mathbb{P})[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})]+\mathrm{o}_p(1)\\ &= (\mathbb{P}_n-\mathbb{P})[\phi(\{\lambda_k^\dagger\},\{q_k^\dagger\})]+\mathrm{o}_p(1).\end{aligned}$$

From the law of large numbers, the statement is concluded.

□

Proof of Theorem 12. Let any $\hat{\pi}_t^b$ be given. Due to boundedness away from zero, we have

$$\begin{aligned}&\left\|\prod_{0=t}^{k}\frac{\pi_t^e}{\hat{\pi}_t^b}-\lambda_k\right\|_2 \le \left\|\sum_{i=0}^{T}\left(\prod_{t=0}^{i}\frac{\pi_t^e}{\hat{\pi}_t^b}\prod_{t=i}^{k}\frac{\pi_t^e}{\pi_t^b}-\prod_{t=0}^{i-1}\frac{\pi_t^e}{\hat{\pi}_t^b}\prod_{t=i}^{k}\frac{\pi_t^e}{\pi_t^b}\right)\right\|_2\\ &\le \sum_{t=0}^{k}\mathrm{o}_p(\|1/\hat{\pi}_t^b-1/\pi_t^b\|_2)\\ &\le \sum_{t=0}^{k}\mathrm{o}_p(\|\hat{\pi}_t^b-\pi_t^b\|_2)\\ &\le \mathrm{o}_p(n^{-\alpha}).\end{aligned}$$

□

Proof of Theorem 13. Without loss of generality, we consider the case $K=2$.

Define $\phi(\{\hat{\mu}_k\},\{\hat{q}_k\})$ as:

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

The estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$ is given by

$$\frac{n_1}{n}\mathbb{P}_{n_1}\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\})+\frac{n_2}{n}\mathbb{P}_{n_2}\phi(\{\hat{\mu}_k^{(2)}\},\{\hat{q}_k^{(2)}\}).$$

Then, we have

$$\begin{aligned}\sqrt{n}(\mathbb{P}_{n_1}\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\rho^{\pi^e}) &= \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\phi(\{\mu_k\},\{q_k\})] &(31)\\ &+\sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\mu_k\},\{q_k\})] &(32)\\ &+\sqrt{n}(\mathrm{E}[\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\})|\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}]-\rho^{\pi^e}). &(33)\end{aligned}$$

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

$$\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\phi(\{\mu_k\},\{q_k\})=D_1+D_2+D_3,\quad\text{where}$$

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

First, we show the term Eq. (31) is $\mathrm{o}_p(1)$.

Lemma 38. *The term Eq.* (31) *is* $\mathrm{o}_p(1)$.

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

$$\lim_{n\to\infty} \sqrt{n_1} P[\mathbb{P}_{n_1}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k^{(1)}\}, \{q_k^{(1)}\})] \\ - \mathrm{E}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon|\mathcal{D}_2] = 0. \tag{34}$$

Then, by bounded convergence theorem, we have

$$\lim_{n\to\infty} \sqrt{n_1} P[\mathbb{P}_{n_1}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})] \\ - \mathrm{E}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon] = 0,$$

yielding the statement.

To show Eq. (34), we show that the conditional mean is 0 and conditional variance is $\mathrm{o}_p(1)$. The conditional mean is

$$\mathrm{E}[\mathbb{P}_{n_1}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] - \\ \mathbb{P}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})]|\mathcal{D}_2] = 0.$$

Here, we used a sample splitting construction, that is, $\hat{\mu}_k^{(1)}$ and $\hat{q}_k^{(1)}$ only depend on $\mathcal{D}_2$. The conditional variance is

$$\begin{aligned} &\mathrm{var}[\sqrt{n_1}\mathbb{P}_{n_1}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})]|\mathcal{D}_2] \\ &= \mathrm{E}[\mathrm{E}[D_1^2 + D_2^2 + D_3^2 + 2D_1D_2 + 2D_2D_3 + 2D_2D_3|\{\hat{q}_k^{(1)}\}, \{\mu_k^{(1)}\}]|\mathcal{D}_2] \\ &= \mathrm{o}_p(1). \end{aligned}$$

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

$$\sqrt{n_1} P[\mathbb{P}_{n_1}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})] - \mathrm{E}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \\ \phi(\{\mu_k\}, \{q_k\})|\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}] > \epsilon|\mathcal{D}_2] \\ \le \frac{1}{\epsilon^2}\mathrm{var}[\sqrt{n_1}\mathbb{P}_{n_1}[\phi(\{\hat{\mu}_k^{(1)}\}, \{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k\}, \{q_k\})]|\mathcal{D}_2] = \mathrm{o}_p(1). \qquad \square$$

Lemma 39. *The term Eq.* (33) *is* $\mathrm{o}_p(1)$.

Proof.

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

□

Finally, we get

$$\sqrt{n}(\mathbb{P}_{n_1}\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) = \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\mu_k\},\{q_k\})] + \mathrm{o}_p(1).$$

Therefore,

$$\begin{aligned}
&\sqrt{n}(\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_2)} - \rho^{\pi^e}) \\
&= n_1/n\sqrt{n}\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) + n_2/n\sqrt{n}(\mathbb{P}_{n_2}\phi(\{\hat{\mu}_k^{(2)}\},\{\hat{q}_k^{(2)}\}) - \rho^{\pi^e}) \\
&= \sqrt{n_1/n}\mathbb{G}_{n_1}[\phi(\{\mu_k\},\{q_k\})] + \sqrt{n_2/n}\mathbb{G}_{n_2}[\phi(\{\mu_k\},\{q_k\})] + \mathrm{o}_p(1) \\
&= \mathbb{G}_n[\phi(\{\mu_k\},\{q_k\})] + \mathrm{o}_p(1),
\end{aligned}$$

concluding the proof by showing the influence function of $\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_2)}$ is the efficient one. □

Proof of Theorem 14. Almost same as the proof of Theorem 7. The differences are replacing λ_t with μ_t, and C^{T+1} with C' □

Proof of Theorem 15. We define $\phi(\{\hat{\mu}_k\},\{\hat{q}_k\})$ as:

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

The estimator $\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_2),\,\mathrm{adaptive}}$ is given by $\mathbb{P}_n\phi(\{\hat{\mu}_k\},\{\hat{q}_k\})$. Then, we have

$$\begin{aligned}
\sqrt{n}(\mathbb{P}_n\phi(\{\hat{\mu}_k\},\{\hat{q}_k\}) - \rho^{\pi^e}) &= \mathbb{G}_n[\phi(\{\hat{\mu}_k\},\{\hat{q}_k\}) - \phi(\{\mu_k\},\{q_k\})] &(35)\\
&+ \mathbb{G}_n[\phi(\{\mu_k\},\{q_k\})] &(36)
\end{aligned}$$

$$+ \sqrt{n}(\mathrm{E}[\phi(\{\hat{\mu}_k\},\{\hat{q}_k\})|\{\hat{\mu}_k\},\{\hat{q}_k\}] - \rho^{\pi^e}). \tag{37}$$

If we can prove that the term Eq. (35) is $\mathrm{o}_p(1)$, the statement is concluded as in the proof of Theorem 13. We proceed to prove this.

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

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

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

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

as in Lemma 33. Therefore, from Lemma 19.24 in van der Vaart (1998), the term Eq. (35) is $\mathrm{o}_p(1)$, concluding the proof. □

Proof of Theorem 16. Without loss of generality, we consider the case $K = 2$.

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,

$$\begin{aligned}
&\sqrt{n}(\mathbb{P}_{n_1}\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \rho^{\pi^e}) \\
&= \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}) - \phi(\{\mu_k^\dagger\},\{q_k^\dagger\})] + \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\mu_k^\dagger\},\{q_k^\dagger\})] \\
&\quad + \sqrt{n/n_1}(\mathrm{E}[\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\});\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\}] \\
&\quad - \mathrm{E}[\phi(\{\mu_k^\dagger\},\{q_k^\dagger\})]) + \sqrt{n}(\mathrm{E}[\phi(\{\mu_k^\dagger\},\{q_k^\dagger\})] - \rho^{\pi^e}) \\
&= \sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\mu_k^\dagger\},\{q_k^\dagger\})] + \sqrt{n}(\mathrm{E}[\phi(\{\mu_k^\dagger\},\{q_k^\dagger\})] - \rho^{\pi^e}) + \mathrm{O}_p(1).
\end{aligned}$$

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,

$$\sqrt{n}(\mathbb{P}_{n_1}\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\rho^{\pi^e})=\sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\mu_k\},\{q_k^\dagger\})]+\mathrm{O}_p(1).$$

Therefore,

$$\begin{aligned}\sqrt{n}(\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_2)}-\rho^{\pi^e})&=\sqrt{n_1/n}\mathbb{G}_{n_1}[\phi(\{\mu_k\},\{q_k^\dagger\})]+\sqrt{n_2/n}\mathbb{G}_{n_2}[\phi(\{\mu_k\},\{q_k^\dagger\})]+\mathrm{O}_p(1)\\&=\mathbb{G}_n[\phi(\{\mu_k\},\{q_k^\dagger\})]+\mathrm{O}_p(1)=\mathrm{O}_p(1),\end{aligned}$$

which shows $\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_2)}$ is $\sqrt{n}$-consistent around ρ^{π^e} 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

$$\sqrt{n}(\mathbb{P}_{n_1}\phi(\{\hat{\mu}_k^{(1)}\},\{\hat{q}_k^{(1)}\})-\rho^{\pi^e})=\sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\mu_k^\dagger\},\{q_k\})]+\mathrm{O}_p(1).$$

Therefore,

$$\begin{aligned}\sqrt{n}(\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_2)}-\rho^{\pi^e})&=\sqrt{n/n_1}\mathbb{G}_{n_1}[\phi(\{\mu_k^\dagger\},\{q_k\})]+\sqrt{n/n_2}\mathbb{G}_{n_2}[\phi(\{\mu_k^\dagger\},\{q_k\})]+\mathrm{O}_p(1)\\&=\mathbb{G}_n[\phi(\{\mu_k^\dagger\},\{q_k\})]+\mathrm{O}_p(1)=\mathrm{O}_p(1),\end{aligned}$$

which shows $\hat{\rho}^{\pi^e}_{\mathrm{DRL}(\mathcal{M}_2)}$ is $\sqrt{n}$-consistent around ρ^{π^e} when the model for the q-function is well-specified. □

Proof of Theorem 17. Almost the same as the proof of Theorem 11 □

Proof of Theorem 18. We first prove

$$\mathbb{P}_n[\hat{w}_t(s_t)\eta_t r_t]=\mathbb{P}_n[w_t(s_t)\eta_t r_t]+\mathbb{P}_n[(\lambda_{t-1}-w_t(s_t))\mathrm{E}_{\pi_e}[r_t|s_t]]+\mathrm{o}_p(n^{-1/2}). \tag{38}$$

Noting

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[s_t^{(i)}=s_t\right]\lambda_{t-1}-p_{\pi_t^e}(s_t)\right\|_\infty&=\mathrm{o}_p(n^{-1/4}),\\\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[s_t^{(i)}=s_t\right]-p_{\pi_t^b}(s_t)\right\|_\infty&=\mathrm{o}_p(n^{-1/4}),\\\hat{a}/\hat{b}=b^{-1}\{1-\hat{b}^{-1}(\hat{b}-b)\}\{(\hat{a}-a)-a/b(\hat{b}-b)\},&\end{aligned}$$

we have

$$\hat{w}_t(s_t)-w_t(s_t)+\mathrm{o}_p(n^{-1/2})$$

$$= \frac{1}{p_{\pi_t^b}(s_t)} \left[\left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[s_t^{(i)} = s_t\right] \lambda_{t-1} - p_{\pi_t^e}(s_t) \right\} - w_t(s_t) \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[s_t^{(i)} = s_t\right] - p_{\pi_t^b}(s_t) \right\} \right].$$

Then,

$$\begin{aligned}
\mathbb{P}_n[\hat{w}_t(s_t)\eta_t r_t] =& \frac{1}{n} \sum_{i=1}^{n} w_t(s_t^{(i)}) \eta_t^{(i)} r_t^{(i)} + \\
&+ \frac{1}{n^2} \sum_{i=1}^{n} \frac{\eta_t^{(i)} r_t^{(i)}}{p_{\pi_t^b}(s_t^{(i)})} \left\{ \sum_{j=1}^{n} \mathbb{I}\left[s_t^{(j)} = s_t^{(i)}\right] \lambda_{t-1}^{(i)} - w_t(s_t^{(i)}) \sum_{j=1}^{n} \mathbb{I}\left[s_t^{(j)} = s_t^{(i)}\right] \right\} \\
=& \frac{2}{n(n-1)} \sum_{i<j} 0.5(a_{ij} + a_{ji}) + \mathrm{o}_p(n^{-1/2}),
\end{aligned}$$

where

$$a_{ij} = \frac{\eta_t^{(i)} r_t^{(i)} \mathbb{I}\left[s_t^{(j)} = s_t^{(i)}\right]}{p_{\pi_t^b}(s_t^{(i)})} \left(\lambda_{t-1}^{(j)} - w_t(s_t^{(i)})\right).$$

From U-statistics theory, by defining $b_{ij}(\mathcal{H}^{(i)}, \mathcal{H}^{(j)}) = 0.5(a_{ij} + a_{ji})$, we have

$$\frac{2}{n(n-1)} \sum_{i<j} b_{ij}(\mathcal{H}^{(i)}, \mathcal{H}^{(j)}) = \frac{2}{n} \sum_{i=1}^{n} \mathrm{E}[b_{ij}(\mathcal{H}^{(i)}, \mathcal{H}^{(j)}) | \mathcal{H}^{(i)}] + \mathrm{o}_p(n^{-1/2}).$$

In addition,

$$\begin{aligned}
\mathrm{E}[a_{i,j}|\mathcal{H}^{(i)}] &= \eta_t^{(i)} r_t^{(i)} \{w_t(s_t^{(i)}) - w_t(s_t^{(i)})\} = 0, \\
\mathrm{E}[a_{j,i}|\mathcal{H}^{(i)}] &= \mathrm{E}\left[\frac{\eta_t^{(j)} r_t^{(j)} \mathbb{I}\left[s_t^{(j)} = s_t^{(i)}\right]}{p_{\pi_t^b}(s_t^{(j)})} \left(\lambda_{t-1}^{(i)} - w_t(s_t^{(j)})\right) | \mathcal{H}^{(i)} \right] \\
&= (\lambda_{t-1}^{(i)} - w_t^{(i)}) \mathrm{E}[\eta_t^{(i)} r_t^{(i)} | s_t^{(i)}]
\end{aligned}$$

Therefore, we have shown Eq. (38). Summing over t yields

$$\mathbb{P}_n\left[\sum_{t=0}^{T} \hat{w}_t(s_t)\eta_t r_t\right] = \mathbb{P}_n\left[\sum_{t=0}^{T} \{w_t(s_t)\eta_t r_t + \lambda_{t-1}\mathrm{E}_{\pi_e}[r_t|s_t] - w_t \mathrm{E}_{\pi_e}[r_t|s_t]\}\right] + \mathrm{o}_p(n^{-1/2}),$$

which concludes the proof by establishing the influence function for $\hat{\rho}_{\mathrm{MIS}}$. □

Proof of Theorem 19. The difference of the influence functions belongs to the orthogonal tangent space. Therefore, the difference of variances is the variance of the difference of the influence functions. This is equal to

$$\mathrm{var}\left[v_0 + \sum_{t=0}^{T} -\mu_t q_t + \mu_t v_{t+1} - \{\lambda_{t-1} - w_t(s_t)\}\mathrm{E}_{\pi_e}[r_t|s_t] \right]$$

$$
\begin{aligned}
&= \mathrm{var}[v_0] + \sum_{t=1}^{T+1} \mathrm{E}\left[\mathrm{var}\left(\mathrm{E}\left[\sum_{k=0}^{T} -\mu_k q_k + \mu_k v_{k+1} - \{\lambda_{k-1} - w_k(s_k)\}\mathrm{E}_{\pi_e}[r_k|s_k]|\mathcal{J}_{s_t}\right]|\mathcal{J}_{s_{t-1}}\right)\right] \\
&= \mathrm{var}[v_0] + \sum_{t=1}^{T+1} \mathrm{E}\left[\mathrm{var}\left(\mathrm{E}\left[\mu_{t-1} v_t + \sum_{k=t}^{T} -\mu_k r_k - \{\lambda_{k-1} - w_k(s_k)\}\mathrm{E}_{\pi_e}[r_k|s_k]|\mathcal{J}_{s_t}\right]|\mathcal{J}_{s_{t-1}}\right)\right] \\
&= \mathrm{var}[v_0] + \sum_{t=1}^{T+1} \mathrm{E}\left[\mathrm{var}\left(\mathrm{E}\left[\mu_{t-1} v_t - \sum_{k=t}^{T} \lambda_{k-1}\mathrm{E}_{\pi_e}[r_k|s_k]|\mathcal{J}_{s_t}\right]|\mathcal{J}_{s_{t-1}}\right)\right] \\
&= \mathrm{var}[v_0] + \sum_{t=1}^{T+1} \mathrm{E}\left[\mathrm{var}\left(\mathrm{E}\left[\mu_{t-1} v_t - \lambda_{t-1} v_t(s_t)|\mathcal{J}_{s_t}\right]|\mathcal{J}_{s_{t-1}}\right)\right] \\
&= \mathrm{var}[v_0] + \sum_{t=1}^{T+1} \mathrm{E}\left[\mathrm{var}\left(\{\mu_{t-1} - \lambda_{t-1}\} v_t(s_t)|\mathcal{J}_{s_{t-1}}\right)\right] \\
&= \mathrm{var}[v_0] + \sum_{t=1}^{T} \mathrm{E}\left[\{w_{t-1} - \lambda_{t-1}\}^2 \mathrm{var}\left(\eta_{t-1} v_t(s_t)|s_{t-1}\right)\right]. \qquad \square
\end{aligned}
$$

Proof of Theorem 20. We have

$$
\begin{aligned}
&\sqrt{n}\{\mathbb{P}[\sum_{c=0}^{T} \hat{w}_c \eta_c(s_c, a_c) r_c] - \rho^{\pi^e}\} \\
&= \mathbb{G}_n[\sum_{c=0}^{T} \hat{w}_c \eta_c r_c - \sum_{c=0}^{T} w_c \eta_c r_c] + \mathbb{G}_n[\sum_{c=0}^{T} w_c \eta_c r_c] \\
&+ \sqrt{n}\left\{\mathrm{E}\{\sum_{c=0}^{T} \hat{w}_c \eta_c r_c|\{\hat{\mu}_c\}\} - \rho^{\pi^e}\right\} \\
&= \mathbb{G}_n[\sum_{c=0}^{T} w_c \eta_c r_c] + \sqrt{n}\left\{\mathrm{E}\{\sum_{c=0}^{T} \hat{w}_c \eta_c r_c|\{\hat{\mu}_c\}\} - \rho^{\pi^e}\right\} + \mathrm{o}_p(1) \\
&= \mathrm{o}_p(1) + \mathbb{G}_n\{\sum_{c=0}^{T} w_c \eta_c r_c\} + \mathbb{G}_n[\sum_{c=0}^{T}\{\lambda_{c-1} - w_c(s_c)\}\mathrm{E}_{\pi_e}[r_c|s_c]].
\end{aligned}
$$

From the second line to the third line, we used a fact that the stochastic equicontinuity term is $\mathrm{o}_p(1)$ as in the proof of Theorem 9 because $\{\hat{w}_c \eta_c r_c\}$ belongs to a Donsker class and the convergence rate condition holds. This fact is confirmed by the fact that a Hölder class with $\alpha > d_{\mathcal{H}_{s_c}}/2$ is a Donsker class (Example 19.9 in van der Vaart, 1998).

From the third line to the fourth line, we have used a result of Example 1(a) in Section 8 of Shen (1997). More specifically, the functional derivative of the loss function with respect to $\hat{w}_c$ is

$$g(s_c) \to \{w_c(s_c) - \lambda_{c-1}\} g(s_c),$$

and the induced metric from the loss function is L^2-metric with respect to $p_{\pi^b}(s_c)$. The functional derivative of the target function with respect to $\hat{\mu}_c$ is

$$g(s_c) \to \mathrm{E}[g(s_c)\eta_c r_c] = \mathrm{E}[g(s_c)\mathrm{E}_{\pi_e}[r_c|s_c]],$$

and Riesz representation of the Hilbert space with the induced L^2-metric with respect to $p_{\pi^b}(s_c)$ is $\mathrm{E}_{\pi_e}[r_c|s_c]$. Therefore, from Theorem 1 in Shen (1997),

$$\mathrm{E}[\sum_{c=0}^{T} \hat{w}_c \eta_c r_c | \{\hat{\mu}_c\}] - \rho^{\pi^e} = (\mathbb{P}_n - \mathbb{P}) \sum_{c=0}^{T} (\lambda_{c-1} - w_c(s_c)) \mathrm{E}_{\pi_e}[r_c|s_c] + \mathrm{o}_p(n^{-1/2}). \qquad \square$$

Proof of Theorem 21. The proof is done as in the proof of Theorem 20. We study the following drift term:

$$\mathrm{E}\left[\sum_{c=0}^{T} \hat{\mu}_c r_c \mid \hat{\mu}_c\right].$$

Here, the functional derivative of the loss function with respect to $\hat{\mu}_c(s_c, a_c)$ is

$$g(s_c, a_c) \to \mathrm{E}[g(s_c, a_c)(\mu_c - \lambda_c)].$$

and Riesz representation of the Hilbert space with the induced L^2-metric with respect to $p_{\pi_b}(s_c, a_c)$ is $\mathrm{E}[r_c|s_c, a_c]$. On the other hand, the functional derivative of the target function with respect to $\hat{\mu}_c$ is

$$g(s_c, a_c) \to \mathrm{E}[g(s_c, a_c) r_c] = \mathrm{E}[g(s_c, a_c) \mathrm{E}[r_c|s_c, a_c]],$$

From Theorem 1 in Shen (1997),

$$\mathrm{E}\left[\sum_{c=0}^{T} \hat{\mu}_c r_c \mid \hat{\mu}_c\right] = \mathbb{P}_n \left[\sum_{c=0}^{T} (\lambda_c - \mu_c) \mathrm{E}[r_c|s_c, a_c]\right] + \mathrm{o}_p(n^{-1/2}). \qquad \square$$

Proof of Theorem 22. We use the general framework developed in Chamberlain (1992) for establishing the efficiency bounds. For the current problem, noting that the orthogonal moment condition

$$\mathrm{E}[e_{q,k+i} e_{q,k}] = \mathrm{E}[\mathrm{E}[e_{q,k+i}|\mathcal{H}_{a_{k+i}}] e_{q,k}]] = 0, \ (0 \le k < k+i \le T)$$

holds, the efficiency bound for β is represented as

$$\left\{\sum_{k=0}^{T} \nabla_\beta m_k(\mathcal{H}_{a_k}; \beta^*) \Sigma_k^{-1}(\mathcal{H}_{a_k}) \nabla_\beta^\top m_k(\mathcal{H}_{a_k}; \beta^*)\right\}^{-1},$$

where $\Sigma_k(\mathcal{H}_{a_k}) = \mathrm{var}[e_{q,k}|\mathcal{H}_{a_k}]$. The statement of the theorem for the efficiency bound for β is arrived at by algebraic simplification of the above. The efficiency bound of ρ^{π^e} is calculated similarly. $\square$

Proof of Theorem 23. Note $\mathrm{var}[e_{q,k}|\mathcal{H}_{a_k}]$ and $\mathrm{var}[e_{q,k-1}|\mathcal{H}_{a_{k-1}}]$ are upper and lower bounded by some constants by assumption. From Jensen's inequality, we also know

$$\mathrm{E}[g^2] = \mathrm{E}[\mathrm{E}[g^2|\mathcal{H}_{s_k}]] \ge \mathrm{E}[\mathrm{E}[g|\mathcal{H}_{s_k}]^2].$$

This conludes that there exists some constant C_1, C_2 such that

$$C_1 \|g\|_2 \le \|g\|_{F,k} \le C_2 \|g\|_2. \qquad \square$$

Appendix C. Additional Details from Section 5.2

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.