Double Reinforcement Learning for Efficient and Robust Off-Policy Evaluation

Nathan Kallus [1] **Masatoshi Uehara** [2]

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.

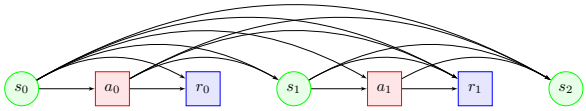

Figure 1. Non-Markov decision process (NMDP)

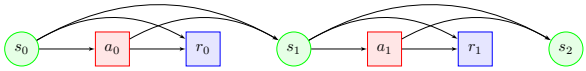

Figure 2. Markov decision process (MDP)

Table 1. Comparison of estimators under MDP. DR (doubly robust) means only one nuisance is needed for consistency. Efficient refers to whether the estimator achieves the efficiency bound under MDP.

Estimator	DR	Efficient	Nuisances
IS			ν
DM			q
DR (Jiang & Li, 2016)	✓		ν, q
MIS (Xie et al., 2019)			μ
DRL (Proposed)	✓	✓	μ, q

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) (Precup et al., 2000; Mahmood et al., 2014; Li et al., 2015; Thomas & Brunskill, 2016; Jiang & Li, 2016; Munos et al., 2016; Liu et al., 2018; Bibaut et al., 2019) 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 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 NMDP (Jiang & Li, 2016; Thomas & Brunskill, 2016; Dudik et al., 2014; Kallus & Uehara, 2019a), work on MDP has been restricted to the parametric, finite-state-finite-action case (Jiang & 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. 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.

We propose the Double Reinforcement Learning (DRL) estimator, which is given by cross-fold estimation and plug-in

[1]Cornell University, Ithaca, NY, USA [2]Harvard University, Massachusetts, Boston, USA. Correspondence to: Masatoshi Uehara <ueharamasatoshi136@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

of the q- and density ratio functions into the efficient influence function for each model, which we derive for the first time here. 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 machine learning method for the nuisance estimation in the spirit of Chernozhukov et al. (2018). Especially, the asymptotic MSE of DRL is polynomial in horizon T, i.e, $\mathcal{O}(T^2/n)$ under mild condition though the one of existing estimators is exponential in T. 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. Properties of DRL are summarized in Table 1 in comparison with other estimators.

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 $t = 1, \dots, T+1$, 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 $\mathcal{H} = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_T, a_T, r_T, s_{T+1})$, given by the product $P_{s_0}(s_0)\pi_0(a_0 \mid \mathcal{H}_{s_0})P_{r_0}(r_0 \mid \mathcal{H}_{a_0}) \cdots P_{r_T}(r_T \mid \mathcal{H}_{a_T})P_{s_{T+1}}(s_{T+1} \mid \mathcal{H}_{a_T})$. 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 Markov decision process (MDP) is an NMDP where transitions and emissions only depend only on the recent state and action, $P_{s_t}(s_t \mid \mathcal{H}_{a_{t-1}}) = P_{s_t}(s_t \mid s_{t-1}, a_{t-1})$ 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. To sum up, a model for the data generating process P_π of $\mathcal{D}$ in NMDP is given by the set of products

$$P_{s_0}(s_0) \prod_{t=0}^{T} \pi_t(a_t \mid \mathcal{H}_{s_t}) P_{r_t}(r_t \mid \mathcal{H}_{a_t}) P_{s_{t+1}}(s_{t+1} \mid \mathcal{H}_{a_t}), \tag{1}$$

over some possible values for each probability distribution in the product. In MDP, it becomes

$$P_{s_0}(s_0) \prod_{t=0}^{T} \pi_t(a_t \mid s_t) P_{r_t}(r_t \mid s_t, a_t) P_{s_{t+1}}(s_{t+1} \mid s_t, a_t). \tag{2}$$

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:

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

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. 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

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

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

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

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

We assume the following throughout this paper.

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

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

1.2. Summary of Semiparametric Inference

We give a concise introduction of semiparametric theory; then, apply it to our problem. For details, refer to van der Vaart (1998); Tsiatis (2006). For details, refer to Section 1.2.

Our target estimand is ρ^{π^e} so a natural question is what is the least-possible error we can achieve in estimating it. In parametric models, the Cramér-Rao bound lower bounds the variance of all unbiased estimators and, due to Hájek (1970), also the asymptotic MSE of *all* (regular) estimators. Our model, however, is nonparametric as it consists of *all* MDP distributions, *i.e.*, *any* choice for $P_{s_0}(s_0)$, $P_{s_t}(s_t \mid s_{t-1}, a_{t-1})$, $P_{r_t}(r_t \mid s_t, a_t)$, and $\pi_t(a_t \mid s_t)$ in Eq. (2), or of all NMDP distributions, *i.e.*, *any* choice for $P_{s_0}(s_0)$, $P_{s_t}(s_t \mid \mathcal{H}_{a_t})$, $P_{r_t}(r_t \mid \mathcal{H}_{a_t})$, and $\pi_t(a_t \mid \mathcal{H}_{s_t})$ in Eq. (1).

Semiparametric theory gives an answer to this question. We first informally state the key property of the *efficient influence functions* (EIF) from semiparametric theory in terms of our own model, which is all MDP distributions, and our estimand, which is ρ^{π^e}.

Theorem 1 (Informal description of van der Vaart (1998), Theorem 25.20)**.** *The EIF $\phi_{\text{eff}}^{\text{MDP}}(\mathcal{H})$ satisfies that*

$$\text{var}[\phi_{\text{eff}}^{\text{MDP}}(\mathcal{H})] = \inf_{\hat{\rho}^{\pi^e} \in \mathcal{R}} \text{AMSE}[\hat{\rho}^{\pi^e}],$$

where $\text{AMSE}[\rho^{\pi^e}] = \int zz^T dF(z)$ is the second moment of F the limiting distribution of $\sqrt{n}(\hat{\rho}^{\pi^e} - \rho^{\pi^e})$ and $\mathcal{R}$ is the class of regular estimators $\hat{\rho}^{\pi^e}$ for ρ^{π^e},

The same theorem holds for the EIF $\phi_{\text{eff}}^{\text{NMDP}}(\mathcal{H})$ in NMDP. Here, $\text{var}[\phi_{\text{eff}}^{\text{MDP}}]$ and $\text{var}[\phi_{\text{eff}}^{\text{NMDP}}]$ are called the *efficiency bounds* under NMDP and MDP. A regular estimator is any whose limiting distribution is insensitive to small changes of order $\mathcal{O}(1/\sqrt{n})$ to P_{π^b} that keep it an MDP distribution (see van der Vaart, 1998, Chapter 25). So the above provides a lower bound on the variance of all regular estimators, which is a very general class. It is so general that the bound also applies to *all* estimators at all in a local asymptotic minimax sense (see van der Vaart, 1998, Theorem 25.21).

The remaining question is how to derive efficient estimators achieving these efficiency bounds. It is achieved by taking the average of the EIFs. In practice, since the EIFs include some unknown functions, we have to plug them in. Under some mild conditions, we can typically prove that there is no inflation of the asymptotic MSE in the first order due to this estimation as we will see in Section 3.

2. Summary of Literature on OPE

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. More specifically, once we have an estimate $\hat{q}_0$ of the first q-function, the DM estimate is simply

$$\hat{\rho}_{\text{DM}}^{\pi^e} = \mathbb{E}_n \left[\mathbb{E}_{\pi^e} \left[\hat{q}_0(s_0, a_0) | s_0 \right] \right],$$

where the inner expectation is simply over $a_0 \sim \pi^e(\cdot \mid s_0)$ (Ernst et al., 2005; Le et al., 2019). However, DM is weak against model–misspecification of q–functions.

The second approach is *importance sampling* (IS), which averages the data weighted by the density ratio of the evaluation and behavior policies (Precup et al., 2000). Given estimates $\hat{\nu}_t$ of the cumulative density ratios (or, letting $\hat{\nu}_t = \nu_t$ if the behavior policy is known), the IS estimate is simply

$$\hat{\rho}_{\text{IS}}^{\pi^e} = \mathbb{E}_n \left[\sum_{t=0}^T \hat{\nu}_t r_t \right].$$

When behavior policy is known, IS is unbiased, but its variance tends to be large due to extreme weights.

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 (Scharfstein et al., 1999; Dudik et al., 2014; Jiang & Li, 2016). The DR estimate has the form

$$\hat{\rho}_{\text{DR}}^{\pi^e} = \mathbb{E}_n \left[\sum_{t=0}^T \left(\hat{\nu}_t (r_t - \hat{q}_t) + \hat{\nu}_{t-1} \mathbb{E}_{\pi^e} \left[\hat{q}_t | s_t \right] \right) \right].$$

Many variations of DR have been proposed (Thomas & Brunskill, 2016; Farajtabar et al., 2018; Kallus & Uehara, 2019a). However, all of the aforementioned IS and DR estimators do not exploit MDP structure and, in particular, will *fail* to be efficient under MDP. Recently, in the same finite-state-and-action-space setting studied by Jiang & Li (2016), Xie et al. (2019) studied an IS-type estimator that exploits MDP structure by replacing density ratios with marginalized density ratios, estimated by within-state-action averages since state and action spaces are assumed finite. More spefically, the estimator has the form

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

given estimates $\hat{\mu}_t$. However, this estimator is also not efficient, even in the finite setting though near optimal. In fact, the asymptotic MSE is $\mathcal{O}(T^3/n)$, not $\mathcal{O}(T^2/n)$.

Finally, note that our focus is a finite horizon problem. This setting is different from the recent work of infinite horizon OPE utilizing the existence of stationary distributions (Liu et al., 2018; Nachum et al., 2019; Kallus & Uehara, 2019b). Their methods *cannot* be directly applied to a finite horizon problem since the stationary distribution does not exist.

3. Semiparametric Inference for Off-Policy Evaluation

In this section, we derive the efficiency bounds and efficient influence functions for ρ^{π^e} under MDP and NMDP.

3.1. Semiparametric Efficiency in Non-Markov Decision Processes

First, we consider the efficiency bound under NMDP. 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, Jiang & Li, 2016; Farajtabar et al., 2018; Kallus & Uehara, 2019a; Thomas & Brunskill, 2016), we are aware of no result showing rigorously that it in fact corresponds to the efficient influence function in the NMDP model or deriving the semiparametric efficiency bound. (Note that, in contrast, the influence function we derive for the MDP model in the next section appears to be novel.)

Theorem 2 (Efficiency bound under NMDP)**.** *The efficient influence function of ρ^{π^e} under NMDP is*

$$\phi_{\text{eff}}^{\text{NMDP}}(\mathcal{H}) = -\rho^{\pi^e} + \sum_{t=0}^{T} \left(\nu_t \left(r_t - q_t\right) + \nu_{t-1} v_t\right), \tag{3}$$

where $v_{T+1} = 0, \nu_{-1} = 0$. The semiparametric efficiency bound under NMDP is

$$\text{EffBd}(\text{NMDP}) = \sum_{t=-1}^{T} \text{E}\left[\nu_t^2 \text{var}\left(r_t + v_{t+1} \mid \mathcal{H}_{a_t}\right)\right], \tag{4}$$

Remark 1. When the action and state spaces are discrete, NMDP 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 & 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. There is a significant gap to deriving the semiparametric bound, which generalizes these results to more general action and state spaces and nonparametric models. Our result is 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 2. The efficient influence function $\phi_{\text{eff}}^{\text{NMDP}}$ has the oft-noted doubly robust structure. Specifically, $\rho^{\pi^e} +$ $\text{E}\left[\phi_{\text{eff}}^{\text{NMDP}}(\mathcal{H})\right]$ is equal to

$$= \underbrace{\text{E}\left[\sum_{t=0}^{T} \nu_t r_t\right]}_{=\rho^{\pi^e}} + \underbrace{\text{E}\left[\sum_{t=0}^{T} \left(-\nu_t q_t + \nu_{t-1} v_t\right)\right]}_{=0}$$

$$= \underbrace{\text{E}\left[v_0\right]}_{=\rho^{\pi^e}} + \underbrace{\text{E}\left[\sum_{t=0}^{T} \nu_t (r_t - q_t + v_{t+1})\right]}_{=0}.$$

The first term in each line corresponds to IPW and direct method (DM) estimators, 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 9 to achieve double robustness for DRL.

3.2. Semiparametric Efficiency in Markov Decision Processes

Next, we derive the efficiency bound and efficient influence function for ρ^{π^e} under MDP. 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 MDP, we have $q_t = q_t(s_t, a_t)$ and $v_t = v_t(s_t)$.

Theorem 3 (Efficiency bound under MDP)**.** *The efficient influence function of ρ^{π^e} under MDP is*

$$\phi_{\text{eff}}^{\text{MDP}}(\mathcal{H}) = -\rho^{\pi^e} + \sum_{t=0}^{T} \left(\mu_t \left(r_t - q_t\right) + \mu_{t-1} v_t\right), \tag{5}$$

where $v_{T+1} = 0$, $\mu_{-1} = 0$. The semiparametric efficiency bound under MDP is

$$\text{EffBd}(\text{MDP}) = \sum_{t=-1}^{T} \text{E}\left[\mu_t^2 \text{var}\left(r_t + v_{t+1} \mid s_t, a_t\right)\right], \tag{6}$$

Remark 3. Again, when the action and state spaces are discrete, MDP 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 & 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 MDP.

Remark 4. The difference between the efficient influence functions in the NMDP and MDP models, $\phi_{\text{eff}}^{\text{NMDP}}$ and $\phi_{\text{eff}}^{\text{MDP}}$, is that (a) the cumulative density ratio ν_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 $\phi_{\text{eff}}^{\text{NMDP}}$, q_t refers to $q_t(\mathcal{H}_{a_t})$, while in $\phi_{\text{eff}}^{\text{MDP}}$, 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 2. 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 MDP, the function $\phi_{\text{eff}}^{\text{NMDP}}(\mathcal{H})$ from Theorem 2 can be shown to still be an influence function of some regular asymptotic linear estimator in MDP. Projecting it onto the tangent space in MDP, 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 $\phi_{\text{eff}}^{\text{MDP}}(\mathcal{H})$.

Remark 5. The efficient influence function $\phi_{\text{eff}}^{\text{MDP}}(\mathcal{H})$ also has a doubly robust structure. Specifically, $\rho^{\pi^e} + \mathrm{E}\left[\phi_{\text{eff}}^{\text{MDP}}(\mathcal{H})\right]$ is equal to

$$\underbrace{\mathrm{E}\left[\sum_{t=0}^{T} \mu_t r_t\right]}_{=\rho^{\pi^e}} + \underbrace{\mathrm{E}\left[\sum_{t=0}^{T} \left(-\mu_t q_t + \mu_{t-1} v_t\right)\right]}_{=0}$$

$$= \underbrace{\mathrm{E}\left[v_0\right]}_{=\rho^{\pi^e}} + \underbrace{\mathrm{E}\left[\sum_{t=0}^{T} \mu_t (r_t - q_t + v_{t+1})\right]}_{=0}.$$

The first term on the first line corresponds to the marginalized IPW estimator of 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 12 to achieve double robustness for DRL.

By comparing the efficiency bounds of Theorem 2 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 $P_{\pi^b} \in$ MDP (i.e., the underlying distribution is an MDP), then*

$$\mathrm{EffBd(MDP)} \leq \mathrm{EffBd(NMDP)}.$$

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

Finally, we can see that $\mathrm{EffBd(MDP)}$ is polynomial in T, more specifically, $\mathcal{O}(T^2/n)$ if C' does not depend on T. For example, if the MDP is stationary, this is often satisfied since the marginal density ratio μ_t does not depend on t. On the other hand, $\mathrm{EffBd(NMDP)}$ is always exponential in T, more specifically, $\mathcal{O}(C^T T^2)$.

Theorem 5. $\mathrm{EffBd(MDP)} \leq C' R_{\max}^2 T^2$ *and* $\mathrm{EffBd(NMDP)} \leq C^T R_{\max}^2 T^2$.

Note that if $\mathrm{E}[\log(\eta_t)] \geq c > 0$ for some constant c, we can also show that $\mathrm{EffBd(NMDP)}$ is bounded below by C^T up to some constant. This suggests $\mathrm{EffBd(NMDP)} = \Theta(C^T)$ under mild conditions. This implies that the curse of horizon is inevitable in NMDPs; however, it can be avoidable in MDPs.

4. Efficient Estimation

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). We proceed by presenting DRL and its properties. Especially,in the MDP setting, DRL is the first semiparametrically efficient and doubly robust estimator.

4.1. Double Reinforcement Learning for NMDPs

Given a learning algorithm to estimate the q-function $q(\mathcal{H}_{a_t})$ and cumulative density ratio function $\nu_t(\mathcal{H}_{a_t})$, DRL for NMDPs proceeds as follows:

1. Split the data randomly into two halves, $\mathcal{D}_0$ and $\mathcal{D}_1$. Let $J^{(i)} \in \{0, 1\}$ be i's half so that $i \in \mathcal{D}_{J^{(i)}}$.

2. For $j = 0, 1$, construct estimators $\hat{\nu}_t^{(j)}(\mathcal{H}_{a_t})$ and $\hat{q}_t^{(j)}(\mathcal{H}_{a_t})$ based on the training data $\mathcal{D}_j$ alone.

3. Let $\hat{\rho}_{\text{DRL(NMDP)}}^{\pi^e}$ be

$$\mathrm{E}_n[\textstyle\sum_{t=0}^{T} \left(\hat{\nu}_t^{(1-J)}(r_t - \hat{q}_t^{(1-J)}) + \hat{\nu}_{t-1}^{(1-J)} \hat{v}_t^{(1-J)}\right)],$$

where $\hat{v}_t^{(1-J)} = \mathrm{E}_{\pi^e}[\hat{q}_t^{(1-J)} \mid \mathcal{H}_{s_t}]$, which is computable as a sum or integral over the known E_{π^e}. Here, in i^{th} term of the empirical expectation, J refers to the data $J^{(i)}$.

In other words, we approximate the efficient influence function $\phi_{\text{eff}}^{\text{NMDP}}(\mathcal{H}) + \rho^{\pi^e}$ from Theorem 2 by replacing the unknown q- and density ratio functions with estimates thereof and we take empirical averages of this approximation, 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.

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

Assumption 3. $0 \leq \hat{\nu}_t \leq C^t, 0 \leq \hat{q}_t \leq (T+1-t)R_{\max}$ for $0 \leq t \leq T$.

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 $\hat{\rho}_{\text{DRL(NMDP)}}$ under NMDP)**.** *Suppose* $\|\hat{\nu}_t^{(j)} - \nu_t\|_2 \|\hat{q}_t^{(j)} - q_t\|_2 = o_p(n^{-1/2})$, $\|\hat{\nu}_t^{(j)} - \nu_t\|_2 = o_p(1)$, $\|\hat{q}_t^{(j)} - q_t\|_2 = o_p(1)$ *for* $0 \leq t \leq T, j = 0, 1$. *Then, the estimator* $\hat{\rho}^{\pi^e}_{\text{DRL(NMDP)}}$ *achieves the semiparametric efficiency bound under* NMDP.

Remark 6. 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 $\hat{\nu}_t$ 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). 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 MIS estimator.

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

Corollary 7 (Efficiency of $\hat{\rho}_{\text{DRL(NMDP)}}$ when the behavior policy is known)**.** *Suppose* $\hat{\nu}_t = \nu_t$ *and* $\|\hat{q}_t^{(j)} - q_t\|_2 = o_p(1)$ *for* $0 \leq t \leq T, j = 0, 1$. *Then, the estimator* $\hat{\rho}^{\pi^e}_{\text{DRL(NMDP)}}$ *achieves the semiparametric efficiency bound under NMDP.*

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 & Li, 2016) when assuming the Donsker condition for in-sample-estimated nuisance functions, since this result was never precisely established before.

Theorem 8 (Efficiency without sample splitting)**.** *Let* $\hat{\nu}_t, \hat{q}_t$ *be estimators based on* $\mathcal{D}$ *and let*

$$\hat{\rho}^{\pi^e}_{\text{DR}} = \mathrm{E}_n \left[\textstyle\sum_{t=0}^{T} \left(\hat{\nu}_t \left(r_t - \hat{q}_t \right) + \hat{\nu}_{t-1} \mathrm{E}_{\pi^e} \left[\hat{q}_t \mid \mathcal{H}_{s_t} \right] \right) \right].$$

Suppose $\|\hat{\nu}_t - \nu_t\|_2 = o_p(n^{-\alpha_{1,t}})$, $\|\hat{q}_t - q_t\|_2 = o_p(n^{-\alpha_{2,t}})$, $\alpha_{1,t} + \alpha_{2,t} \geq 1/2, \alpha_{1,t}, \alpha_{2,t} > 0$ *for* $0 \leq t \leq T$ *and that* $\hat{q}_t$, $\hat{\nu}_t$ *belong to a Donsker class. Then, the estimator* $\hat{\rho}^{\pi^e}_{\text{DR}}$ *achieves the semiparametric efficiency bound under* NMDP.

Thus, in NMDP, in comparison to the standard doubly robust estimator, DRL enjoys efficiency under milder conditions. To our knowledge, Theorems 6 and 8 are the first results precisely showing semiparametric efficiency for any OPE estimator.

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

Theorem 9 (Double robustness)**.** *Suppose* $\|\hat{\nu}_t^{(j)} - \nu_t^\dagger\|_2 = \mathcal{O}_p(n^{-\alpha_{t,1}})$ *and* $\|\hat{q}_t^{(j)} - q_t^\dagger\|_2 = \mathcal{O}_p(n^{-\alpha_{t,2}})$ *for* $0 \leq t \leq T, j = 0, 1$. *If, for each* $0 \leq t \leq T$, *either* $\nu_t^\dagger = \nu_t$ *and* $\alpha_{t,1} \geq 1/2$, $\alpha_{t,2} > 0$ *or* $q_t^\dagger = q_t$ *and* $\alpha_{t,1} > 0$, $\alpha_{t,2} \geq 1/2$, *then the estimator* $\hat{\rho}^{\pi^e}_{\text{DRL(NMDP)}}$ *is* $\sqrt{n}$*-consistent.*

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

4.2. Estimation of Nuisance Functions for $\hat{\rho}^{\pi^e}_{\text{DRL(NMDP)}}$

A remaining question is when can we get nonparametric estimators achieving the necessary rates for the density ratio and q- functions.

Cumulative density ratio: When the behavior policy is unknown, ν_k can be estimated by estimating and plugging in π^b, which can in turn be estimated by nonparametric regression. Specifically, we let $\hat{\nu}_k^{(j)} = \prod_{0=t}^{k} \pi_t^e / \hat{\pi}_t^{b,j}$, where $\hat{\pi}_t^{b,j}$ is a standard kernel regression estimator or sieve regression estimator (Newey & Mcfadden, 1994; Stone, 1994). When $\pi_t^b(a_t \mid \mathcal{H}_{s_t})$ belongs to the Hölder class with smoothness parameter α and the dimension of the space $\mathcal{H}_{s_t}$ is $d_{\mathcal{H}_{s_t}}$ and the dimension of the action space is 1, it can be shown (ibid.) that $\|\hat{\pi}_t^{b,j} - \pi_t^b\|_2 = \mathcal{O}_p(n^{-\alpha/(2\alpha + d_{\mathcal{H}_{s_t}})})$. We therefore have the following result.

Lemma 10. *Assume $\hat{\pi}_t^{b,j}$ and π_t^b are uniformly bounded by some constant below and that $\pi_t^b(a_t \mid \mathcal{H}_{s_t})$ is Hölder with parameter α. Then, $\|\hat{\nu}_t^{(j)} - \nu_t\|_2 = \mathcal{O}_p(n^{-\alpha/(2\alpha+d_{\mathcal{H}_{s_t}})})$.*

Q-function: The q-function estimation is discussed in many literature. For example, recursive type fitted Q–iteration based on Bellman–equation is one of the common estimators (Le et al., 2019). The rate is obtained in some literature (Antos et al., 2008). The other possible estimator is a nonparametric regression estimator based on the relation: $\mathbb{E}[q_t(\mathcal{H}_{a_t})] = \mathrm{E}_\pi[\sum_{k=t}^T r_k \mid \mathcal{H}_{a_t}]$. In this case, the standard nonparametric regression results can be applied (Chen, 2007).

Remark 7. Alternatively, parametric models can be used for q_t and (if behavior policy is unknown) ν_t. Then, under standard regularity conditions, using MLE and other parametric regression estimators for behavior policy would yield $\|\hat{\nu}_t^{(j)} - \nu_t^\dagger\|_2 = \mathcal{O}_p(n^{-1/2})$, where $\nu_t^\dagger = \nu_t$ if the model is well-specified. Similarly, we have $\|\hat{q}_t^{(j)} - q_t^\dagger\|_2 = \mathcal{O}_p(n^{-1/2})$, where $q_t^\dagger = q_t$ if the model of Q-function is well-specified. 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 9 ensures the estimator is still $\sqrt{n}$-consistent.

4.3. Double Reinforcement Learning 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 for MDPs proceeds as follows:

1. Split the data randomly into two halves, $\mathcal{D}_0$ and $\mathcal{D}_1$. Let $J^{(i)} \in \{0, 1\}$ be i's half so that $i \in \mathcal{D}_{J^{(i)}}$.

2. For $j = 0, 1$, construct estimators $\hat{\mu}_t^{(j)}(s_t, a_t)$ and $\hat{q}_t^{(j)}(s_t, a_t)$ based on the training data $\mathcal{D}_j$.

3. Let $\hat{\rho}^{\pi^e}_{\mathrm{DRL(MDP)}}$ be

$$\mathrm{E}_n\left[\sum_{t=0}^T \left(\hat{\mu}_t^{(1-J)}\left(r_t - \hat{q}_t^{(1-J)}\right) + \hat{\mu}_{t-1}^{(1-J)}\hat{v}_t^{(1-J)}\right)\right],$$

   where $\hat{v}_t^{(1-J)} = \mathrm{E}_{\pi^e}[\hat{q}_t^{(1-J)} \mid s_t]$, which is computable as a sum or integral over the known E_{π^e}.

Again, what we have done is approximating the efficient influence function $\phi^{\mathrm{MDP}}_{\mathrm{eff}}(\mathcal{H}) + \rho^{\pi^e}$ from Theorem 3 and taken its empirical average, 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, to establish the properties of DRL for MDPs, we assume the following conditions for the estimators, reflecting Assumptions 1 and 2:

Assumption 4. $0 \le \hat{\mu}_t \le C^t$, $0 \le \hat{q}_t \le (T + 1 - t)R_{\max}$ for $0 \le t \le T$.

The following result establishes that DRL is the first efficient OPE estimator for MDPs. In addition, from Theorem 5, this implies that the asymptotic MSE is $\mathcal{O}(C'T^2/n)$. 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 11 and the structure of the influence function, which ensures no variance inflation due to estimating the nuisance functions.

Theorem 11 (Efficiency of $\hat{\rho}^{\pi^e}_{\mathrm{DRL(MDP)}}$ under MDP)**.** *Suppose $\|\hat{\mu}_t^{(j)} - \mu_t\|_2 \|\hat{q}_t^{(j)} - q_t\|_2 = o_p(n^{-1/2})$, $\|\hat{\mu}_t^{(j)} - \mu_t\|_2 = o_p(1)$, $\|\hat{q}_t^{(j)} - q_t\|_2 = o_p(1)$ for $0 \le t \le T, j = 0, 1$. Then, the estimator $\hat{\rho}^{\pi^e}_{\mathrm{DRL(MDP)}}$ achieves the semiparametric efficiency bound under* MDP.

In addition to efficiency, DRL again enjoys a double robustness guarantee in MDP, as in NMDP.

Theorem 12 (Double robustness)**.** *Suppose $\|\hat{\mu}_t^{(j)} - \mu_t^\dagger\|_2 = \mathcal{O}_p(n^{-\alpha_{t,1}})$ and $\|\hat{q}_t^{(j)} - q_t^\dagger\|_2 = \mathcal{O}_p(n^{-\alpha_{t,2}})$ for $0 \le t \le T, j = 0, 1$. If, for each $0 \le t \le T$, either $\mu_t^\dagger = \mu_t$ and $\alpha_{t,1} \ge 1/2$, $\alpha_{t,2} > 0$ or $q_t^\dagger = q_t$ and $\alpha_{t,2} \ge 1/2$, $\alpha_{t,1} > 0$, then the estimator $\hat{\rho}^{\pi^e}_{\mathrm{DRL(MDP)}}$ is $\sqrt{n}$-consistent.*

4.4. Estimation of Nuisance Functions for $\hat{\rho}^{\pi^e}_{\mathrm{DRL(MDP)}}$

A remaining question is how to estimate the nuisances at the necessary rates. Regarding q-functions, see Section 4.2. Here, we discuss how to estimate the marginal ratio μ_k.

For estimating μ_k, one can leverage the following relationship to reduce it to a regression problem:

$$\mu_t(s_t, a_t) = \eta_t(s_t, a_t)w_t(s_t), \quad \text{where } w_t(s_t) = \mathrm{E}[\nu_{t-1} \mid s_t]. \tag{7}$$

Thus, when the behavior policy is known, we need only estimate w_t, which amounts to regressing ν_{t-1} on s_t. So, for example, if $w_t(s_t)$ belongs to Hölder class with smoothness α and s_t has dimension d_s, we can estimate w_t with a sieve-type estimator $\hat{w}_t$ based on the loss function $(\nu_{t-1} - w_t(s_t))^2$:

$$\hat{w}_t(s_t) = \arg\min_{w_t(s_t) \in \Lambda^\alpha_{d_{s_t}}} \mathrm{E}_n[(w_t(s_t) - \nu_{t-1})^2], \tag{8}$$

where $\Lambda^\alpha_{d_{s_t}}$ is the space of Hölder functions with smoothness α and the dimension d_{s_t}. By letting $\hat{\mu}_t^{(j)}(s_t, a_t) =$

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

Size	$\hat{\rho}_{\text{IS}}$	$\hat{\rho}_{\text{DRL(NMDP)}}$	$\hat{\rho}_{\text{DM}}$	$\hat{\rho}_{\text{MIS}}$	$\hat{\rho}_{\text{DRL(MDP)}}$
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	$\hat{\rho}_{\text{IS}}$	$\hat{\rho}_{\text{DRL(NMDP)}}$	$\hat{\rho}_{\text{DM}}$	$\hat{\rho}_{\text{MIS}}$	$\hat{\rho}_{\text{DRL(MDP)}}$
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)

$\eta_t(s_t, a_t)\hat{w}_t^{(j)}(s_t)$, the convergence rate is $\|\hat{\mu}_t^{(j)}(s_t, a_t) - \mu_t(s_t, a_t)\|_2 = \mathcal{O}_p(n^{-\alpha/(\alpha+d_{s_t})})$ (Chen, 2007). If the behavior policy is unknown, we can first estimate η_t as $\pi_t^e/\hat{\pi}_t^b$, and then plug it into the regression:

$$\hat{w}_t(s_t) = \arg\min_{w_t(s_t)\in\Lambda^{\alpha}_{d_{s_t}}} \mathrm{E}_n[(w_t(s_t) - \hat{\nu}_{t-1})^2].$$

As long as the convergence rate for η is the same as the rate of the second-stage regression, we will obtain the same final rate for $\hat{w}$.

In the finite-state-and-action-space setting in (Xie et al., 2019), the other example is a histogram estimator:

$$\hat{w}_t(s_t) = \frac{\sum_{i=1}^n \mathbb{I}[s_t^{(i)}=s_t]\nu_{t-1}}{\sum_{i=1}^n \mathbb{I}[s_t^{(i)}=s_t]}. \tag{9}$$

In continuous space cases, the histogram estimator in Eq. (9) can also easily be extended to a kernel estimator:

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

where K_h is a kernel with a bandwidth h.

Remark 8 (Parametric model case)**.** In the special case where we use parametric models for μ_t and q_t, under some regularity conditions, parametric estimators will generally satisfy $\|\hat{\mu}_t - \mu_t^\dagger\|_2 = \mathcal{O}_p(n^{-1/2})$ and $\|\hat{q}_t - q_t^\dagger\|_2 = \mathcal{O}_p(n^{-1/2})$, where $q_t^\dagger = q_t$ and $\mu_t^\dagger = \mu_t$ if the models are well-specified. Thus, if both models are correctly specified, then Theorem 11 yields local efficiency. If only one model is correctly specified, Theorem 12 yields double robustness.

5. Experiments

We now turn to an empirical study of OPE and DRL. We study comparative performance of different OPE estimators $\hat{\rho}_{\text{IS}}^{\pi^e}$, $\hat{\rho}_{\text{DRL(NMDP)}}^{\pi^e}$, $\hat{\rho}_{\text{DM}}^{\pi^e}$, $\hat{\rho}_{\text{MIS}}^{\pi^e}$ and $\hat{\rho}_{\text{DRL(MDP)}}^{\pi^e}$ in two standard OpenAI Gym tasks: (Brockman et al., 2016): Cliff Walking and Mountain Car. Here, the IS estimate and DM estimate are simply $\hat{\rho}_{\text{IS}}^{\pi^e} = \mathrm{E}_n\left[\sum_{t=0}^T \hat{\nu}_t r_t\right]$ and $\hat{\rho}_{\text{DM}}^{\pi^e} = \mathrm{E}_n\left[\mathrm{E}_{\pi^e}\left[\hat{q}_0(s_0, a_0)|s_0\right]\right]$. We do not compare recent infinite horizon OPE estimators assuming the existence of stationary distributions (Liu et al., 2018; Nachum et al., 2019; Kallus & Uehara, 2019b) since it *cannot* be directly applied to these environments.

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. (7). For Cliff Walking, we use a histogram estimator for w as in Eq. (9). 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. 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 & Recht (2008).

We compare $\hat{\rho}_{\text{IS}}$, $\hat{\rho}_{\text{DRL(NMDP)}}$, $\hat{\rho}_{\text{DM}}$, $\hat{\rho}_{\text{MIS}}$, $\hat{\rho}_{\text{DRL(MDP)}}$. 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 find that the performance of $\hat{\rho}_{\text{DRL(MDP)}}$ is superior to all other estimators in either setting. This is especially true in Cliff Walking. The estimator $\hat{\rho}_{\text{DRL(MDP)}}$ also improves upon $\hat{\rho}_{\text{IS}}$ and $\hat{\rho}_{\text{DM}}$ but not as much as $\hat{\rho}_{\text{DRL(MDP)}}$. The estimator $\hat{\rho}_{\text{MIS}}$ offers a slight improvement over $\hat{\rho}_{\text{IS}}$, but is still outperformed by $\hat{\rho}_{\text{DRL(MDP)}}$, $\hat{\rho}_{\text{DRL(NMDP)}}$, and $\hat{\rho}_{\text{DM}}$.

That the improvement of $\hat{\rho}_{\mathrm{MIS}}$ over $\hat{\rho}_{\mathrm{IS}}$ and the overall improvements of $\hat{\rho}_{\mathrm{DRL(MDP)}}$ 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. 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.

Our DRL opens the door to many exciting future directions. One possible direction is how to apply these DRL to policy optimization. A possible way is combining DRL with policy gradient (Kallus & Uehara, 2020). We would be able to consider combining DRL with other policy optimization methods.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments and suggestions.

This material is based upon work supported by the National Science Foundation under Grant No. 1846210. Masatoshi Uehara was supported in part by MASASON Foundation.

References

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

Bibaut, A., Malenica, I., Vlassis, N., and van der Laan, M. 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*, pp. 654–663, 2019.

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

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

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

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68, 2018.

Dudik, M., Erhan, D., Langford, J., and Li, L. Doubly robust policy evaluation and optimization. *Statistical Science*, 29:485–511, 2014.

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

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

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. Guidelines for reinforcement learning in healthcare. *Nat Med*, 25: 16–18, 2019.

Hájek, J. A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14:323–330, 1970.

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

Kallus, N. and Uehara, M. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *Advances in Neural Information Processing Systems 32*, pp. 3320–3329. 2019a.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019b.

Kallus, N. and Uehara, M. Statistically efficient off-policy policy gradients. *arXiv preprint arXiv: 2002.04014*, 2020.

Le, H., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712, 2019.

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

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31*, pp. 5356–5366. 2018.

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

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

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems 29*, pp. 1054–1062. 2016.

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

Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems 32*. 2019.

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

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

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

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

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

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

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

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

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

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

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

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems 32*, pp. 9665–9675. 2019.