Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Nathan Kallus, Masatoshi Uehara

To cite this article:
Nathan Kallus, Masatoshi Uehara (2022) Efficiently Breaking the Curse of Horizon in Off-Policy Evaluation with Double Reinforcement Learning. Operations Research

Published online in Articles in Advance 23 Feb 2022

. https://doi.org/10.1287/opre.2021.2249

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–21
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

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

Nathan Kallus,[a] **Masatoshi Uehara**[a,*]

[a] Cornell University, New York, New York 10044
*Corresponding author
Contact: kallus@cornell.edu, https://orcid.org/0000-0003-1672-0507 (NK); mu223@cornell.edu, https://orcid.org/0000-0001-9017-3105 (MU)

Received: December 23, 2019
Revised: May 3, 2021; October 31, 2021
Accepted: November 25, 2021
Published Online in Articles in Advance: February 23, 2022

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2021.2249

Abstract. Off-policy evaluation (OPE) in reinforcement learning is notoriously difficult in long- and infinite-horizon settings due to diminishing overlap between behavior and target policies. In this paper, we study the role of Markovian and time-invariant structure in efficient OPE. We first derive the efficiency bounds and efficient influence functions for OPE when one assumes each of these structures. This precisely characterizes the curse of horizon: in time-variant processes, OPE is only feasible in the near-on-policy setting, where behavior and target policies are sufficiently similar. But, in time-invariant Markov decision processes, our bounds show that truly off-policy evaluation is feasible, even with only just one dependent trajectory, and provide the limits of how well we could hope to do. We develop a new estimator based on double reinforcement learning (DRL) that leverages this structure for OPE. Our DRL estimator simultaneously uses estimated stationary density ratios and q-functions and remains efficient when both are estimated at slow, nonparametric rates and remains consistent when either is estimated consistently. We investigate these properties and the performance benefits of leveraging the problem structure for more efficient OPE.

Funding: This work was supported by the National Science Foundation Division of Information and Intelligent Systems [1846210], and by the Masason Foundation.
Supplemental Material: The online appendices are available at https://doi.org/10.1287/opre.2021.2249.

Keywords: off-policy evaluation • Markov decision processes • infinite horizon • semiparametric efficiency

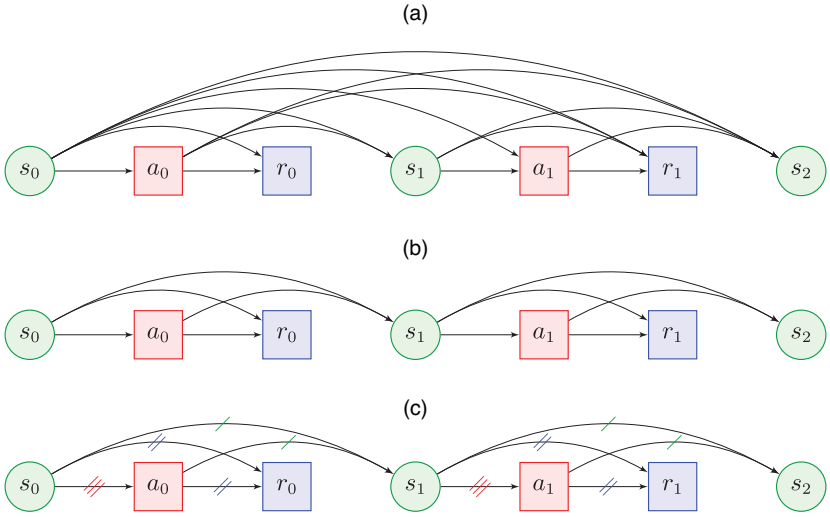
1. Introduction

Reinforcement learning (RL) in settings such as healthcare (Murphy 2003) and education (Mandel et al. 2014) is often limited to the offline or off-policy setting, where we only use existing observed data, due to the inability to simulate and the costliness of exploration. One important task in this setting is off-policy evaluation (OPE), where we want to estimate the mean reward of a candidate decision policy, known as the target policy using observed data generated by the log of another policy, known as the behavior policy (Precup et al. 2000, Mahmood et al. 2014, Li et al. 2015, Jiang and Li 2016, Munos et al. 2016, Thomas and Brunskill 2016, Liu et al. 2018b, Xie et al. 2019).[1] OPE, in particular, is a building block toward policy optimization from observational data (Huang and Jiang 2020, Kallus and Uehara 2020c). OPE, however, becomes increasingly difficult for problems with long and infinitely long horizons (Liu et al. 2018a). As the horizon grows, the overlap (i.e., density ratios) between trajectories generated by the target and behavior policies diminishes exponentially. This issue has, in particular, been noted as one of the key limitations for the applicability of RL in medical settings (Gottesman et al. 2019).

In this paper, we study the fundamental estimation limits for OPE in infinite-horizon settings, and we develop new estimators that leverage special problem structure to achieve these limits and enable efficient and effective OPE in these problem settings. Specifically, we first derive what is the best-possible asymptotic mean-squared error (MSE) that one can hope for in OPE in this setting, that is, we derive the efficiency bounds (van der Vaart 1998), which characterize the minimum limit of the square-root-scaled MSE (as we define in Section 1.3). To study the effect of problem structure, we separately consider three different models: non-Markov decision processes (NMDPs), time-varying Markov decision processes (TMDPs), and time-invariant Markov decision processes (MDPs). These models are illustrated in Figure 1 and precisely defined Section 1.2. Specifically, we focus on discounted bounded rewards. The differences between these bounds exactly characterizes the effect of taking into consideration additional problem structure on the feasibility of OPE.

Our bounds in the NMDP and TMDP models reveal an important phase transition: if the target and behavior policies are sufficiently similar (relative to the discount factor), then consistent estimation is feasible.

Figure 1. (Color online) Bayes Net Representation of the Independence Structure of the Truncated Trajectory Ending with s_2, $\mathcal{J}_{s_2}$, Under the Three Models: NMDP, TMDP, and MDP



Notes. Conditional on its parents, a node is independent of all other nodes. The congruency sign // indicates that the conditional probability function given parent nodes is equal.

Otherwise, there exist examples where it is infeasible. This can be understood as a phrase transition between being sufficiently close to on-policy that OPE is feasible even in infinite horizons and being sufficiently off-policy that it is hopeless. We show that adaptations of the doubly robust (DR) estimator in NMDPs (Jiang and Li 2016) and in MDPs (Kallus and Uehara 2020a) to the infinite horizon case achieve these bounds, that is, are efficient in the near-on-policy setting.

Our bounds in the MDP models, on the other hand, give hope for OPE in the truly off-policy setting. They show that by leveraging Markovian and time-invariant structure in RL problems, we can overcome the *curse of horizon* and indicate what it would mean to do so efficiently, that is, using all the data available optimally. The question is then how to achieve these bounds for efficient OPE. We propose an approach based on double reinforcement learning (Kallus and Uehara 2020a) and on simultaneously learning average visitation distributions and q-functions. And, we show that, unlike importance-sampling-based estimators (Liu et al. 2018a), our DRL estimator achieves the efficiency bound under certain mixing conditions. Thus, by carefully leveraging problem structure, we show how to efficiently break the curse of horizon in RL OPE.

1.1. Organization

The organization of papers is as follows. In Section 1.2, we define the decision process models and set up the OPE problem formally. In Section 1.3, we define the efficiency bounds formally, briefly reviewing semiparametric inference as it relates to our results. In Section 1.4, we review the relevant literature on OPE.

In Section 2, we derive the efficiency bounds under each of the models under consideration, NMDP, TMDP, and MDP. In Section 3, we analyze the asymptotic properties when we extend standard DR and DRL OPE estimators to infinite horizons and provide conditions for their efficiency in the NMDP and TMDP models. We note, however, that they are not efficient under the MDP model and have the wrong MSE scaling.

In Section 4, we propose the first efficient estimator for OPE under the MDP model and analyze its asymptotic properties as $T \to \infty$, including when our observations consist of a single trajectory, $n = 1$. This estimator is based on simultaneously learning q-functions and the ratio of average visitation distributions. In Section 6, we therefore discuss how to estimate the density ratio of average visitation distributions in an off-policy manner from a single (finite) trajectory. And, in Section 7, we discuss how to estimate q-functions in an off-policy manner from a single (finite) trajectory. In Section 8, we provide a numerical experiment to study the effects of leveraging problem structure efficiently. Finally, we conclude in Section 9.

1.2. Problem Setup and Notation

We consider a state space $\mathcal{S}$, action space $\mathcal{A}$, and reward space $\mathcal{R} \subset [0, R_{\max}]$, each a measurable space that may be continuous, discrete, or mixed.[2] We fix a base measure for each, $\lambda_{\mathcal{S}}, \lambda_{\mathcal{A}}, \lambda_{\mathcal{R}}$ (e.g., Lebesgue, counting, or other), focus on distributions on these spaces that are absolutely continuous with respect to (wrt) these, and identify them with their densities (Radon-Nikodym derivative wrt the base measure). A (time-invariant) Markov decision process (MDP) on $(\mathcal{S}, \mathcal{A}, \mathcal{R})$ is given by a reward distribution $p(r \mid s, a)$ for

the immediate reward after taking action a in state s and a transition distribution $p(s' \mid s,a)$ for the new state after taking action a in state s. A policy is a distribution $\pi(a \mid s)$ for the action to take in state s. We also associate with π an initial state distribution, $p_\pi^{(0)}(s_0)$.[3] Recall we identify distributions with densities so $p(r \mid s,a), p(s' \mid s,a), \pi(a \mid s), p_\pi^{(0)}(s_0)$ are densities with respect to $\lambda_{\mathcal{R}}, \lambda_{\mathcal{S}}, \lambda_{\mathcal{A}}, \lambda_{\mathcal{S}}$, respectively. Together, an MDP and a policy define a joint distribution over trajectories $\mathcal{J} = (s_0, a_0, r_0, s_1, a_1, r_1, \cdots)$. Namely, letting $\mathcal{J}_{s_{T+1}} = (s_0, a_0, r_0, \ldots, s_T, a_T, r_T, s_{T+1})$ be the length-$(T+1)$ trajectory up to s_{T+1}, we have that for any T, $\mathcal{J}_{s_{T+1}}$ has density $p_\pi^{(0)}(s_0)\pi(a_0 \mid s_0)p(r_0 \mid s_0,a_0)p(s_1 \mid s_0,a_0)\pi(a_1 \mid s_1)$ $p(r_1 \mid s_1,a_1) \cdots p(s_{T+1} \mid s_T,a_T)$. We also define $\mathcal{H}_{s_{T+1}} = (s_0, a_0, \ldots, s_T, a_T, s_{T+1})$ as the same length-$(T+1)$ trajectory but excluding reward variables, which has density $p_\pi^{(0)}(s_0)\pi(a_0 \mid s_0)p(s_1 \mid s_0,a_0) \cdots p(s_{T+1} \mid s_T,a_T)$, and we similarly denote by $\mathcal{H}_{a_T}$ the trajectory up to and including the variable a_T, excluding rewards. (We formally define MDP as a statistical model for the data-generating process in Definition 3.) We denote by $p_\pi^{(t)}(s_t)$ or $p_\pi^{(t)}(s_t,a_t,r_t,s_{t+1})$ the marginal distribution of s_t or of (s_t,a_t,r_t,s_{t+1}) (etc.) under p_π. We further define the γ-discounted average visitation frequency as

$$p_{\pi,\gamma}^{(\infty)}(s) = \lim_{T\to\infty} \frac{1}{\sum_{t=0}^T \gamma^t} \sum_{t=0}^T \gamma^t p_\pi^{(t)}(s).$$

Our ultimate goal is to estimate the average cumulative reward of the known target evaluation policy (and known initial state distribution), π_e, for a given discount factor $\gamma \in [0,1)$:

$$\rho^{\pi_e} = \lim_{T\to\infty} \rho_T^{\pi_e}, \quad \text{where}$$
$$\rho_T^\pi = c_T(\gamma)\, \mathrm{E}_{p_\pi}\left[\sum_{t=0}^T \gamma^t r_t\right],\ c_T(\gamma) = \left(\sum_{t=0}^T \gamma^t\right)^{-1}.$$

In particular, we wish to estimate ρ^{π_e} based on data generated by a different policy, π_b, known as the behavior policy and which may be known or unknown. (For brevity, we often use the subscript e or b to mean the subscript π_e or π_b, respectively.)

We will consider two data-generation settings.

Transition-Sampling Setting. In the transition-sampling setting, the data consists of n independent and identically distributed (iid) draws of state-action-reward-state quadruplets, $\mathcal{D} = \{(s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)})\}_{i=1}^n$, each drawn from $p_{\pi_b}(\mathcal{J}_{s_1})$. Note we do not assume stationarity in this setting, that is, the marginal densities of $p_{\pi_b}(\mathcal{J}_{s_1})$ wrt s and wrt s' can be different.

Trajectory-Sampling Setting. In the trajectory-sampling setting, the data consists of N observations of length-$(T+1)$ trajectories, $\mathcal{D} = \{(\mathcal{J}_{s_{T+1}}^{(i)}\}_{i=1}^N$, each drawn from $p_{\pi_b}(\mathcal{J}_{s_{T+1}})$. Here, we set $n = NT$ as we have n transitions, and also identify $\mathcal{D} = \{(s_t^{(j)}, a_t^{(j)}, r_t^{(j)}, s_{t+1}'^{(j)})\}_{j=1,t=0}^{N,T}$. Crucially, in this setting the transitions may be dependent. Unlike the transition-sampling setting, here we assume that the data are stationary: $p_{\pi_b}^{(t)} = p_{\pi_b}^{(t')}$ for any t,t'. That is, $p_{\pi_b}^{(0)}(s)$ is an invariant distribution under the state-transition kernel induced by the MDP and π_b. This appears strong but can be easily relaxed if we assume certain ergodicity so that the initial distribution does not in fact matter and we can allow any $p_{\pi_b}^{(0)}(s)$; we discuss this in Remark 8.

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

$$q(s_0,a_0) = \mathrm{E}_{p_{\pi_e}}\left[\sum_{k=0}^\infty \gamma^k r_k \mid s_0,a_0\right],$$
$$v(s_0) = \mathrm{E}_{p_{\pi_e}}\left[\sum_{k=0}^\infty \gamma^k r_k \mid s_0\right] = \mathrm{E}_{p_{\pi_e}}\left[q(a,s) \mid s_0\right].$$

Note that the very last expectation is taken only over $a_0 \sim \pi_e(a_0 \mid s_0)$. We define the policy, cumulative, marginal, and stationary density ratios, respectively, as

$$\eta(s,a) = \frac{\pi_e(a \mid s)}{\pi_b(a \mid s)}, \quad \nu_t(\mathcal{H}_{a_t}) = \prod_{k=0}^t \eta_k(s_k,a_k),$$
$$\mu_t(s_t,a_t) = \frac{p_{\pi_e}^{(t)}(s_t,a_t)}{p_{\pi_b}^{(t)}(s_t,a_t)}, \quad w(s) = \frac{p_{\pi_e,\gamma}^{(\infty)}(s)}{p_{\pi_b}^{(0)}(s)}.$$

In particular, in the latter, notice that we divide a γ-discounted average visitation frequency by an undiscounted marginal one. (In Remark 8, we discuss assuming ergodicity instead of stationarity in the trajectory-sampling setting, in which case we replace the denominator of $w(s)$ with the undiscounted stationary state distribution under $p_{\pi_b}(\mathcal{J})$.)

We can generalize the MDP setting in two ways. In TMDP, the reward, transition, and policy distributions can all depend on t, whereas the Markov assumption is still retained. Adding a t subscript to denote this, under TMDP, p_π is given by $p_\pi^{(0)}(s_0)\pi_0(a_0 \mid s_0)p_0(r_0 \mid s_0, a_0)p_1(s_1 \mid s_0,a_0)\ldots$. In NMDP, the reward, transition, and policy distributions can additionally all depend on the history of states and actions so that p_π is given by $p_\pi^{(0)}(s_0)\pi_0(a_0 \mid s_0)p_0(r_0 \mid s_0,a_0)p_1(s_1 \mid s_0,a_0)\ \pi_1(a_1 \mid \mathcal{J}_{s_1})$ $p_1(r_1 \mid \mathcal{J}_{a_1})\ldots$. In TMDP, q- and v-functions depend on t and are defined as the conditional expectations of $\sum_{k=0}^\infty \gamma^k r_{k+t}$ given s_t, a_t, and s_t, respectively, under p_{π_e}. In NMDP, we condition instead on $\mathcal{J}_{a_t}$ and $\mathcal{J}_{s_t}$, respectively. In both TMDP and NMDP, η is also t-dependent since the policies are. We only consider the trajectory-sampling setting under either TMDP and NMDP since, due to the time dependence, just observing length-1 trajectories would not be enough. (We formally define TMDP and NMDP as a statistical model for the data-generating process in Definitions 1 and 2.)

To streamline notation, when no subscript is denoted, all expectations $\mathrm{E}[\cdot]$ and variances $\mathrm{var}[\cdot]$ are

taken wrt the behavior policy π_b, that is, p_{π_b}. At the same time, recall that v- and q-functions are for the target policy, π_e. For a function f of (parts of) a trajectory we often write f to mean the random variable $f(\mathcal{J})$. The L^p norm is defined as $\|f\|_p = \mathrm{E}[|f|^p]^{1/p}$. For example, we write $v_t = v_t(\mathcal{J}_{a_t})$, $\mu_t = \mu_t(s_t, a_t)$, etc. In the transition-sampling setting, for any function of s, a, r, s', we define its empirical average as

$$\mathbb{P}_n f = \mathbb{P}_n[f(s,a,r,s')] = n^{-1}\sum_{i=1}^{n} f(s^{(i)}, a^{(i)}, r^{(i)}, s'(i)).$$

When f also depends on the index I, we write $\mathbb{P}_n f(s, a, r, s', i) = n^{-1}\sum_{i=1}^{n} f(s^{(i)}, a^{(i)}, r^{(i)}, s'(i), i)$. In the trajectory-sampling setting, we define the time average as

$$\mathbb{P}_T f = \mathbb{P}_T[f(s,a,r,s')] = (T+1)^{-1}\sum_{t=0}^{T} f(s_t, a_t, r_t, s_{t+1}),$$

and for any function of a trajectory, we define the empirical average as

$$\mathbb{P}_N f = \mathbb{P}_N[f(\mathcal{J})] = N^{-1}\sum_{i=1}^{N} f(\mathcal{J}^{(i)}).$$

Thus, for a function of (s, a, r, s'), we have:

$$\mathbb{P}_N\mathbb{P}_T f = N^{-1}(T+1)^{-1}\sum_{t=0}^{T}\sum_{j=1}^{N} f(s_t^{(j)}, a_t^{(j)}, r_t^{(j)}, s_{t+1}^{(j)}),$$

which we also denote by $\mathbb{P}_n = \mathbb{P}_N\mathbb{P}_T$ and also allow functions that depend on the index $i = (j, t)$. Table EC.1 in the online appendix summarizes our notation.

1.3. Efficiency Bounds

In this section, we define formally what we mean by the best-possible asymptotic MSE. We focus on computing efficiency bounds in settings where the data are iid and its distribution fully identifying of the estimand (transition sampling for MDP and infinitely long trajectory sampling for TMDP and NMDP) so that we can apply standard semiparametric theory (Bickel et al. 1998, Tsiatis 2006, Kosorok 2008). After establishing these bounds, we will actually show they can be achieved by estimators both in these ideal settings and even in more complex sampling settings, such as a single growing trajectory. Here, we give a general overview of semiparametric theory as it pertains to our results and provide more technical detail and precise definitions in online Appendix B.2.

Suppose our data consists of n iid observations, each drawn from a distribution p, $O_1, \dots, O_n \sim p$. Let us fix p_0 as the true, unknown distribution. Whereas we do not know p_0, we assume it belongs to a model $\mathcal{M}$, that is, a set of possible data-generating processes. Given a parameter of interest $R : \mathcal{M} \to \mathbb{R}$, we want to estimate $R(p_0)$ using some estimator $\hat{R}(O_1, \dots, O_n)$. For example, in the transition-sampling setting under MDP, we will let $\mathcal{M}$ be all distributions $p_{\pi_b}(\mathcal{J}_1)$ for any choice of MDP and behavior policy, subject to certain minimal regularity and identifiability conditions that ensure the policy value is in fact a function of $p_{\pi_b}(\mathcal{J}_1)$.

The limiting law of $\hat{R}$ is the distributional limit of $\sqrt{n}(\hat{R} - R(p_0))$ and the asymptotic mean-squared error (AMSE) is the second moment of the limiting law, which in turn lower bounds the scaled limit infimum of the mean-squared error (MSE), $\liminf n\mathrm{E}[(\hat{R} - R(p_0))^2]$, by the portmanteau lemma. Roughly, we say $\hat{R}$ is regular wrt $\sqrt{n}$ if its limiting law is invariant to vanishing perturbations to p_0 that remain inside $\mathcal{M}$ (see Definition EC.7 in the online appendix for precise definition). This type of regularity is common and is often considered desirable, as otherwise the estimator may behave erratically under completely undetectable changes (see van der Vaart 1998, section 8.1). If $\sqrt{n}(\hat{R} - R(p_0)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i) + o_p(1/\sqrt{n})$ with $\mathbb{E}\phi(O) = 0$ then $\hat{R}$ is said to be asymptotically linear (AL) with influence function ϕ, and it follows its limiting law is $\mathcal{N}(0, \mathbb{E}\phi^2(O))$ at p_0.

Every gradient of R wrt $\mathcal{M}$ at $p = p_0$ is a Gâteaux derivative for all paths through p_0 that remain in $\mathcal{M}$, which is a p_0-measurable random variable $\phi(O)$. See Definition EC.6 in the online appendix for precise definition. The influence function of any regular AL (RAL) estimator is such a gradient (Theorem EC.1 in the online appendix). The gradient ϕ_{eff} with least second moment (if such exists) is called the *efficient influence function* (EIF). This motivated by the fact (Theorem EC.2) that

$$\mathrm{EffBd}(\mathcal{M}) = \mathrm{E}_{p_0}\phi_{\mathrm{eff}}^2,$$

which we call the efficiency bound, lower bounds the AMSE of any estimator that is regular wrt $\mathcal{M}$. An efficient estimator (at p_0) is a regular estimator (at p_0) with AMSE equal to $\mathrm{EffBd}(\mathcal{M})$.

If we have $\mathrm{EffBd}(\mathcal{M}) < \infty$ (i.e., the estimand is differentiable) and an estimator is shown to be AL with the EIF as its influence function, then, in addition, it is also regular and hence RAL and efficient, and conversely every efficient estimator is RAL (van der Vaart 1998, lemma 25.23). This also suggests an estimation strategy: try to approximate $\hat{\psi}(O) \approx \phi_{\mathrm{eff}}(O) + R(p)$ and use $\hat{R} = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}(O_i)$. 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.

Notice the efficiency bound depends on both p_0 and $\mathcal{M}$. We use $\mathrm{EffBd}(\mathcal{M})$ to highlight the latter dependence. Indeed, if $p_0 \in \mathcal{M} \subseteq \mathcal{M}'$, then $\mathrm{EffBd}(\mathcal{M}) \le \mathrm{EffBd}(\mathcal{M}')$ since estimators that are regular in $\mathcal{M}'$ are also regular in $\mathcal{M}$, even though p_0 is the same. Standard results (e.g., van der Vaart 1998, theorem 25.21) further establish that the efficiency lower bound also applies to all estimators (not just regular ones) in a local minimax fashion, where the local worst-case neighborhoods of p_0 are restricted to

remain in $\mathcal{M}$. The efficiency bound is infinite when the estimand is not pathwise differentiable wrt $\mathcal{M}$, in which case no regular estimators exist (Newey 1990).

1.4. Summary of Literature on OPE

OPE is a central problem in both RL and in the closely related dynamic treatment regimes (DTRs; Murphy et al. 2001). OPE is also equivalent to estimating the total treatment effect of some dynamic policy in a causal inference setting. Although we do not explicitly use counterfactual notation (potential outcomes or *do*-calculus), if we assume the usual sequential ignorability conditions (Ertefaie 2014), the estimands are the same and our results immediately apply.

In RL, one usually assumes that the (time-invariant) MDP model $\mathcal{M}_3$ holds. Nonetheless, with some exceptions that we review next, OPE methods in RL have largely not leveraged the additional independence and time-invariance structure of $\mathcal{M}_3$ 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. One can estimate the q-function by a value iteration in a finite-state-and-action-space setting utilizing an approximated MDP based on the empirical distribution (Bertsekas 2012). More generally, modeling the transition and reward probabilities and using the MDP approximated by the estimates is called the model-based approach (Sutton and Barto 2018). When the sample space and action space are continuous, we can apply some functional approximation to q-function modeling and use the temporal difference method (Lagoudakis and Parr 2004) or fitted Q–iteration (Antos et al. 2008). Once we have an estimate $\hat{q}$, the DM estimate is simply

$$\hat{\rho}_{\mathrm{DM}} = (1-\gamma)\,\mathbb{P}_N\big[\mathrm{E}_{\pi_e}[\hat{q}_0|s_0]\big],$$

where the inner expectation is simply over $a_0 \sim \pi_e(\cdot|s_0)$ 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. For DM, we can leverage the structure of $\mathcal{M}_3$ by simply restricting the q-function we learn to be the same for all t and solving the fixed point of the Bellman equation. However, DM can fail to be efficient and 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 $\hat{v}_t$ of v_t (or, $\hat{v}_t = v_t$ if the behavior policy is known), the IS estimate is simply

$$\hat{\rho}_{\mathrm{IS}} = c_T(\gamma)\,\mathbb{P}_N\left[\sum_{t=0}^{T}\gamma^t\hat{v}_t r_t\right].$$

A common variant is the self-normalized IS (SNIS), where we divide the tth summand by $\mathbb{P}_N\big[\gamma^t\hat{v}_t\big]$. Recall that T here denotes the finite length of the N trajectories in our data. In finite-horizon problems (i.e., when the estimand is $\rho_T^{\pi_e}$), when the behavior policy is known, IS is unbiased and consistent but its variance tends to be large and it is inefficient (Hirano et al. 2003). In infinite-horizon problems, we need T to grow for consistent estimation. But even if $T=\infty$ (i.e., our data consists of full trajectories), IS can have infinite variance because of diminishing overlap, known as the curse of horizon (Liu et al. 2018a). Our results (Table 1) in $\mathcal{M}_1$, $\mathcal{M}_2$ characterize more precisely when this curse applies or not.

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 and Li 2016). Under $\mathcal{M}_1$, the DR estimate has the form

$$\hat{\rho}_{\mathrm{DR}} = c_T(\gamma)\,\mathbb{P}_N\left[\sum_{t=0}^{T}\gamma^t\Big(\hat{v}_t(r_t-\hat{q}_t)+\hat{v}_{t-1}\mathrm{E}_{\pi_e}\big[\hat{q}_t\,|\,s_t\big]\Big)\right].$$

In finite-horizon problems, DR is known to be efficient under $\mathcal{M}_1$ (Kallus and Uehara 2020a). In infinite horizons, we derive the additional conditions needed for efficiency in $\mathcal{M}_1$ in Section 3.

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 $\hat{q}(s,a)$ to minimize variance rather than use a plug-in. Kallus and Uehara (2019) propose a variant that is similarly locally efficient but further ensures asymptotic MSE no worse than DR, IS, and SNIS under misspecification and stability properties similar to self-normalized IS.

However, all of the aforementioned IS and DR estimators do not leverage Markov structure and fail to be efficient under $\mathcal{M}_2$. Recently, in finite horizons, Kallus and Uehara (2020a) derived the efficiency bound of $\rho_T^{\pi_e}$ under $\mathcal{M}_2$ and provided an efficient estimator termed *double reinforcement learning* (DRL), taking the form

$$\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)} = c_T(\gamma)\,\mathbb{P}_N\left[\sum_{t=0}^{T}\gamma^t\Big(\hat{\mu}_t^{(i)}(r_t-\hat{q}_t^{(i)})+\hat{\mu}_{t-1}^{(i)}\mathrm{E}_{\pi_e}\big[\hat{q}_t^{(i)}\,|\,s_t\big]\Big)\right],$$

where $\hat{\mu}^{(i)}, \hat{q}^{(i)}$ can either be estimated in-sample ($\hat{q}_t^{(i)} = \hat{q}_t$ and assuming a Donsker condition) or cross-fitting (the sample is split and $\hat{q}_t^{(i)}$ is fit on the fold that

Table 1. Asymptotic Order of the Best-Achievable MSE in Each Model When Observing N Length-$(T+1)$ Trajectories

Model	Characteristics	MSE scaling	Required conditions
NMDP	Non-Markov, time variant	$\mathcal{O}(1/N)$	$N\to\infty$, $T=\omega(\log N)$, $\|\nu_t\|_\infty=\mathcal{O}(\gamma^{-t})$
TMDP	Markov, time variant	$\mathcal{O}(1/N)$	$N\to\infty$, $T=\omega(\log N)$, $\|\mu_t\|_\infty=\mathcal{O}(\gamma^{-t})$
MDP	Markov, time invariant	$\mathcal{O}(1/(NT))$	$T\to\infty$, $N\geq 1$, mixing, $\|w\|_\infty=O(1)$

Note. The variables η_t, ν_t, μ_t, w are the instantaneous, cumulative, marginal, and stationary density ratios, respectively (see Section 1.2 for definitions).

excludes *i*). DRL's efficiency depends only on the rates of convergence of these estimates, which can be as slow as $N^{-1/4}$, thus enabling the use of black box machine learning methods. In infinite horizons, we derive the additional conditions needed for efficiency in $\mathcal{M}_2$ in Section 3.

However, again, all of the aforementioned IS, DR, and DRL estimators do not leverage time-invariance and fail to be efficient under $\mathcal{M}_3$. Our results extend the notion of the curse of dimension and demonstrate that even estimators in $\mathcal{M}_2$, such as the efficient $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$, can fail to be consistent as μ_t can also explode just like ν_t. In contrast, in $\mathcal{M}_3$, regardless of the rate of growth of μ_t, ν_t, consistent evaluation is possible from even just a single trajectory and knowledge of the initial distribution.

Recently, Liu et al. (2018a) proposed a variant of the IS estimator for $\mathcal{M}_3$ that uses the ratio of the stationary distributions in hopes of overcoming the curse of horizon. We describe this estimator in detail in Section 6.1. Its asymptotic MSE was not previously studied. We provide some results in the parametric setting. The properties in the nonparametric setting are not known. In particular, as we discuss in Section 6.1, its lack of doubly robust structure and its not being an empirical average of martingale differences make analysis particularly challenging. At the same time, these issues also suggest that the estimator is inefficient.

2. Efficiency Bounds in Infinite Horizons

The efficiency bounds for $\rho_T^{\pi_e}$ in finite horizons under NMDP and TMDP are derived in Kallus and Uehara (2020a). First, we extend these results to infinite horizons, focusing in particular on when the bounds are infinite. Then—and more importantly—we study the efficiency bound in MDP.

2.1. Efficiency Bounds in Non-Markov and Time-Variant Markov Decision Processes

First, we formally define NMDP and TMDP as statistical models for our data-generating process. As data, we consider observing N (infinitely long) trajectories $\mathcal{J}$ from the behavior-policy-induced distribution $p_{\pi_b}(\mathcal{J})$. The model is the set of possibilities for $p_{\pi_b}(\mathcal{J})$. The NMDP model is given by (almost) all arbitrary distributions on the sequence $\mathcal{J}$.

Definition 1 (NMDP Models $\mathcal{M}_1, \mathcal{M}_{1,b}$). The NMDP model $\mathcal{M}_1$ is defined by all distributions $p_{\pi_b}(\mathcal{J})$ such that the conditional distribution of each variable in $\mathcal{J}$ given the past is absolutely continuous wrt the respective base measure (so it has a density) and the conditional distribution of action given history, $\pi_{b,t}$, is such that the (known and fixed) evaluation policy, $\pi_{e,t}$, is absolutely continuous wrt it. We also define the model $\mathcal{M}_{1,b}$ where we assume the behavior policy is known; that is, $\pi_{b,t}$ and $p^{(0)}_{\pi_b}$ are fixed at their known value and not allowed to vary.

The last restriction in the definition of $\mathcal{M}_1$ is known as weak overlap and it is equivalent to saying ν_t exists. It is necessary so to ensure that ρ^{π_e} is a function of $p_{\pi_b}(\mathcal{J})$, that is, is identifiable from the data (Khan and Tamer 2010). Observing infinitely long trajectories is also necessary for identifiability, but when constructing estimators we will show it suffices to observe trajectories of modestly growing length. Then, ρ^{π_e} is a function of $p_{\pi_b}(\mathcal{J})$ given by $(1-\gamma)\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t\nu_t r_t\right]$, that is, it is a well-defined map $\mathcal{M}_1\to\mathbb{R}$.

The TMDP model is obtained by restricting the NMDP model to satisfy the Markovian condition.

Definition 2 (TMDP Models $\mathcal{M}_2, \mathcal{M}_{2,b}$). The TMDP model $\mathcal{M}_2$ is defined by restricting the model $\mathcal{M}_1$ so that s_{t+1} is conditionally independent of $\mathcal{J}_{r_{t-1}}$ given s_t, a_t, and r_t is conditionally independent of $\mathcal{J}_{r_{t-1}}$ given s_t, a_t, and a_t is conditionally independent of $\mathcal{J}_{r_{t-1}}$ given s_t. Similarly, we define $\mathcal{M}_{2,b}$ by fixing $\pi_{b,t}$ and $p^{(0)}_{\pi_b}$ at their known value.

All of our models are nonparametric in the sense that we do not further restrict these distributions in any way beyond requiring densities and overlap.

We now proceed to compute the efficiency bounds for ρ^{π_e} in these models. By slightly modifying the results of Kallus and Uehara (2020a), we obtain the following theorems.

Theorem 1 (EB Under NMDP).

$$\begin{aligned}\mathrm{EB}(\mathcal{M}_1) &= \mathrm{EB}(\mathcal{M}_{1,b}) \\ &= (1-\gamma)^2 \sum_{k=1}^{\infty} \mathrm{E}[\gamma^{2(k-1)} v_{k-1}^2(\mathcal{J}_{a_{k-1}}) \mathrm{var}(r_{k-1} \\ &\quad + v_k(\mathcal{J}_{s_k}) \mid \mathcal{J}_{a_{k-1}})]. \end{aligned} \tag{1}$$

Theorem 2 (EB Under TMDP).

$$\begin{aligned}\mathrm{EB}(\mathcal{M}_2) &= \mathrm{EB}(\mathcal{M}_{2,b}) \\ &= (1-\gamma)^2 \sum_{k=1}^{\infty} \mathrm{E}[\gamma^{2(k-1)} \mu_{k-1}^2(s_{k-1}, a_{k-1}) \\ &\quad \mathrm{var}(r_{k-1} + v_k(a_k, s_k) \mid a_{k-1}, s_{k-1})]. \end{aligned} \tag{2}$$

Remark 1. Equations (1) and (2) are almost the same as the limit as $T \to \infty$ of $c_T^2(\gamma)$ times the finite-horizon efficiency bounds derived by Kallus and Uehara (2020a). They are the same if we replace the lower summation limit with $k = 0$ instead of $k = 1$ in Equations (1) and (2). This is because we here assume $p_{\pi_e}^{(0)}$ is known whereas in Kallus and Uehara (2020a) the assumption is that $p_{\pi_b}^{(0)} = p_{\pi_e}^{(0)}$ are unknown, the uncertainty due to which increases the efficiency bound.

Remark 2. Notice the crucial role of the model in the notion of efficiency, even though p_{π_b} is a given single distribution. The efficiency bounds in Theorems 1 and 2 are local: given a single p_{π_b}, the bounds correspond to the best MSE we can achieve if we are regular under small perturbations of p_{π_b} that remain inside the model. In particular, even if $p_{\pi_b} \in \mathcal{M}_2$ happens to be TMDP, the efficiency bound is different whether we allow perturbations that remain TMDP or just NMDP. That is, if our estimator "works" for NMDPs (i.e., is regular in $\mathcal{M}_1$), it will suffer the larger bound (Equation (1)) even if the particular instance encountered happens to be a TMDP.

Theorems 1 and 2 show that, when Equations (1) and (2) are finite, the best-achievable leading term in the MSE of any regular estimator in NMDP or TMDP is $\mathrm{EB}(\mathcal{M}_1)/N$ or $\mathrm{EB}(\mathcal{M}_2)/N$, respectively. It also shows that the knowledge of $\pi_b, p_{\pi_b}^{(0)}$ does not improve the bound. The intuitive reason for this is that ρ^{π_e} is only a function of the transition- and reward-distribution parts of $p_{\pi_b}(\mathcal{J})$ so that $\pi_b, p_{\pi_b}^{(0)}$ are ancillary. When the efficiency bound takes an infinite value, the estimand is not pathwise differentiable wrt the model and no regular $\sqrt{n}$-consistent estimator exists (Newey 1990).

Corollary 1 (Sufficient Conditions for Existence of Efficiency Bounds). *If* $\|v_k\|_\infty = o(\gamma^{-k})$, *then* $\mathrm{EB}(\mathcal{M}_1) < \infty$. *If* $\|\mu_k\|_\infty = o(\gamma^{-k})$, *then* $\mathrm{EB}(\mathcal{M}_2) < \infty$. *Moreover, if* $p_{\pi_b} \in \mathcal{M}_2$ *and* $\mathrm{EB}(\mathcal{M}_1) < \infty$, *then* $\mathrm{EB}(\mathcal{M}_2) < \infty$.

Remark 3 (The Curse of Horizon in $\mathcal{M}_1$, Extended). To demonstrate the curse of horizon, Liu et al. (2018a) gave an example where the IS estimator has a diverging variance as horizon grows. But it is not clear if—and without assuming MDP structure—there might be another estimator that would not suffer from this. Our results show that in fact there is not. If we take any example where $\mathrm{var}(r_{k-1} + v_k \mid \mathcal{J}_{a_{k-1}})$ are uniformly lower bounded (i.e., state transitions and reward emissions are nondegenerate), then as long as $\mathrm{E}[\log(\eta_k)] \geq -\log(\gamma)$ for all k, we will necessarily have that $\mathrm{EB}(\mathcal{M}_1) = \infty$. (Notice that $\mathrm{E}[\log(\eta_k)]$ is exactly the expected Kullback-Leibler divergence.) In this case, as long as we are not restricting the model beyond $\mathcal{M}_1$, we simply cannot break the curse of horizon and it affects all (regular) estimators, not just IS.

Remark 4 (The Curse of Horizon in $\mathcal{M}_2$, a Milder Version of the Original). Our results further extend the curse of horizon to $\mathcal{M}_2$, providing another refinement of the notion. The curse is milder in $\mathcal{M}_2$ than in $\mathcal{M}_1$, since the EBs are necessarily ordered. It is, in fact, much milder. In particular, rather than involve the growth of the cumulative density ratios, whether $\mathrm{EB}(\mathcal{M}_2)$ converges or diverges depends on the growth of the marginal density ratios. These, of course, can also grow and $\mathrm{EB}(\mathcal{M}_2)$ can diverge. However, whereas we can easily make $\mathrm{EB}(\mathcal{M}_1) = \infty$ even with a simple MDP example, to make $\mathrm{EB}(\mathcal{M}_2)$ diverge we need a more pathological example. It can be verified that if p_{π_b} is actually stationary (or, nonstationary but ergodic) and the stationary distributions overlap, then we will necessarily have $\|\mu_k\|_\infty = O(1)$.

This means that, for an MDP, we can overcome the curse of horizon that affects estimators such as $\hat{\rho}_{\mathrm{DR}}$ and $\hat{\rho}_{\mathrm{IS}}$ by using estimators that are efficient under $\mathcal{M}_2$, the first of which was proposed by Kallus and Uehara (2020a), that is, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$. However, this is still not efficient in an MDP case. In fact, this is not just a matter of constants: this will not even yield the right scaling of the MSE.

2.2. Efficiency Bounds in Time-Invariant Markov Decision Processes

Next, we consider the MDP model. For the efficiency bound computation, we focus on the transition-sampling setting, where we observe n draws from $p_{\pi_b}(\mathcal{J}_{s_1})$. We next formally define an MDP as a statistical model for our data.

Definition 3 (MDP Models $\mathcal{M}_3, \mathcal{M}_{3,b}$). The MDP model, $\mathcal{M}_3$, is given by all distributions $p_{\pi_b}(\mathcal{J}_{s_1})$ on (s, a, r, s') such that the distribution of s' is independent of r given s, a, the conditional distribution of each variable given the past is absolutely continuous wrt the respective base measure (so it has a density), and further the base measure is absolutely continuous wrt the distribution of s. As before, we define $\mathcal{M}_{3,b}$ by fixing $p_{\pi_b}^{(0)}$ and π_b at their known value.

The last restriction ensures $w(s)$ exists without putting additional restrictions on the MDP itself. The existence of $w(s)$ is the analogue of overlap for the MDP setting and is necessary for identifiability. Then, ρ^{π_e} is a functional of $p_{\pi_b}(\mathcal{J}_{s_1})$ given by $\mathbb{E}[w(s)\eta(s,a)r]$, that is, it is a well-defined map $\mathcal{M}_3 \to \mathbb{R}$.

Theorem 3 (EB Under MDP). *The EIF in either $\mathcal{M}_3$ or $\mathcal{M}_{3,b}$ is*

$$\phi_{\text{eff}}(s,a,r,s') = w(s)\eta(s,a)(r+\gamma v(s') - q(s,a)).$$

The efficiency bound in either model is therefore

$$\text{EB}(\mathcal{M}_3) = \text{E}\Big[w^2(s)\eta^2(a,s)(r+\gamma v(s') - q(s,a))^2\Big]. \quad (3)$$

Theorem 3 shows that the lower bound of the first order asymptotic MSE is $\text{EB}(\mathcal{M}_3)/n$. It also shows that the knowledge of $\pi_b, p^{(0)}_{\pi_b}$ does not improve the bound. This suggests that in MDP, the MSE should scale inversely with the number of transitions (n) we observe, not the number of trajectories (T). Whereas standard efficiency analysis does not apply to the trajectory-sampling setting in MDP since the transitions are dependent, we will show in Section 4 that we can nonetheless achieve the same efficiency bound with a scaling of $n = N(T+1)$ under certain mixing assumptions. Thus, the achievable MSE under MDP is a factor of T faster than under NMDP and TMDP. In this sense, efficiency in $\mathcal{M}_3$ corresponds to improvement in the rate, not just the constant, relative to efficiency in $\mathcal{M}_1$ or $\mathcal{M}_2$. This is in contrast to the comparison between $\mathcal{M}_1$ and $\mathcal{M}_2$, which have efficiency bounds that are on the same scale and only differ in the leading coefficient.

Remark 5 (Unknown $p^{(0)}_{\pi_e}$). In our setup, we assumed $p^{(0)}_{\pi_e}$ is known, but our results can be extended to the case where $p^{(0)}_{\pi_e}$ is unknown but we see samples from it. In particular, suppose $p^{(0)}_{\pi_e}$ is allowed to vary arbitrarily in the model (but remains a density wrt the state base measure) and our data consists of n iid draws of (s_0,s,a,r,s') from $p^{(0)}_{\pi_e}(s_0)p_{\pi_b}(\mathcal{J}_1)$. Then, a modification of Theorem 3 shows that the EB (whether we know the behavior policy or not) is

$$\text{var}_{p^{(0)}_{\pi_e}}[v(s_0)] + \text{E}\Big[w^2(s)\eta^2(a,s)(r+\gamma v(s') - q(s,a))^2\Big].$$

Compared with Equation (3), we have an additional term corresponding to the variance wrt $p^{(0)}_{\pi_e}$.

When $\gamma = 0$, this reduces to the bound in the no-horizon bandit OPE setting (Robins et al. 1994):

$$\text{var}_{p^{(0)}_{\pi_e}}[v(s)] + \text{E}\Big[\eta^2(a,s)(r - q(s,a))^2\Big],$$

where here $q(s,a) = \mathbb{E}[r \mid s,a]$ becomes simply the outcome regression function.

3. Efficient Estimators for Infinite Horizons Under NMDP and TMDP

Before turning to developing an efficient estimator under the MDP model, we briefly review how we can extend the efficient finite-horizon DRL estimators of Kallus and Uehara (2020a) to be efficient in the infinite-horizon NMDP and TMDP settings.

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. For example, metric entropy or Donsker assumptions can be avoided by using a cross-fitting strategy (Klaassen 1987, Zheng and van Der Laan 2011, Chernozhukov et al. 2018). We proceed by presenting the infinite-horizon extensions of the DRL estimators of Kallus and Uehara (2020a) and their properties. Again, the two DRL estimators present here are not efficient under $\mathcal{M}_3$.

3.1. Non-Markov Decision Process

The infinite-horizon extension of the DRL estimator under $\mathcal{M}_1$ is as follows. We consider the trajectory-sampling setting where we observe N trajectories. Fix some horizon truncation ω_N. Let $q^{\omega_N}_t = \mathbb{E}_{\pi_e}\big[\sum_{k=t}^{\omega_N} \gamma^{t-k} r_t \mid \mathcal{J}_{a_t}\big]$, $v^{\omega_N}_t = \mathbb{E}_{\pi_e}\big[\sum_{k=t}^{\omega_N} \gamma^{t-k} r_t \mid \mathcal{J}_{s_t}\big]$. Then the estimator is given by

$$\hat{\rho}_{\text{DRL}(\mathcal{M}_1)} = c_{w_N}(\gamma)\frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{\omega_N}\gamma^t\Big(\hat{v}^{(i)}_t(\mathcal{H}^{(i)}_{a_t})\Big(r^{(i)}_t - \hat{q}^{(i)}_t(\mathcal{H}^{(i)}_{a_t})\Big) + \hat{v}^{(i)}_{t-1}(\mathcal{H}^{(i)}_{a_{t-1}})\hat{v}^{(i)}_t(\mathcal{H}^{(i)}_{s_t})\Big),$$

where $\hat{v}^{(i)}_t$, $\hat{q}^{(i)}_t$ are some plug-in estimates of v_t, $q^{\omega_N}_t$ to be used for the ith observation and $\hat{v}^{(i)}_t(\mathcal{H}_{s_t}) = \text{E}_{a_t \sim \pi_e(\cdot \mid \mathcal{J}_{s_t})}[\hat{q}^{(i)}_t(\mathcal{H}_{a_t}) \mid \mathcal{J}_{s_t}]$ is the corresponding v-estimate. Notice that $\hat{v}^{(i)}_t$ is computable as it is an integral wrt the known measure $\pi_e(\cdot \mid \mathcal{J}_{s_t})$ (e.g., it is a simple sum if $\mathcal{A}$ is finite).

We can consider two cases. In the adaptive version, we construct functional estimators $\hat{v}_t, \hat{q}_t$ based on the whole data and then set $\hat{v}^{(i)}_t = \hat{v}_t$, $\hat{q}^{(i)}_t = \hat{q}_t$. The adaptive version of $\hat{\rho}_{\text{DRL}(\mathcal{M}_1)}$ is exactly the DR estimator, $\hat{\rho}_{\text{DR}}$.

In the cross-fitting version, the sample is evenly split into two folds and $\hat{v}^{(i)}_t$, $\hat{q}^{(i)}_t$, $\hat{v}^{(i)}_t$ are computed on estimates fit on the opposite fold so that they are independent of data point i. Namely, the cross-fitting procedure is:

- Split the data set into two disjoint data sets, $\mathcal{D}_0$ and $\mathcal{D}_1$. Let $j(i)$ be such that $\mathcal{J}^{(i)} \in \mathcal{D}_{j(i)}$.
- Using only the trajectories in $\mathcal{D}_0$, construct the functional estimators $\hat{v}^{[0]}_t, \hat{q}^{[0]}_t$ for $t \le \omega_N$. And, using only the trajectories in $\mathcal{D}_1$, construct the functional estimators $\hat{v}^{[1]}_t, \hat{q}^{[1]}_t$ for $t \le \omega_N$.
- Set $\hat{v}^{(i)}_t = \hat{v}^{[j(i)]}_t$, $\hat{q}^{(i)}_t = \hat{q}^{[j(i)]}_t$.

Kallus and Uehara (2020a, section 6) discusses the estimation of $v_t, q_t^{\omega_N}$, that is, q-functions for finite-horizon problems. In particular, if the behavior policy is known, we can simply let $\hat{v}_t^{(i)} = v_t$. As we make formal later, our q-estimates need only estimate $q_t^{\omega_N}$ and not q_t, which depends on all future rewards ad infinitum. This can be done using only the truncated trajectory $\mathcal{J}_{r_{\omega_N}}^{(i)}$ (e.g., using regression), and given q-estimates, the estimator similarly only depends on $\mathcal{J}_{r_{\omega_N}}^{(i)}$. Therefore, whereas we can consider it as an estimator in the $\mathcal{M}_1$ model where we observe the infinitely long $\mathcal{J}^{(i)}$, it is in fact implementable even if we just see finite trajectories of length at least ω_N.

We can now state a straightforward infinite-horizon extension of the efficiency result of (Kallus and Uehara 2020a, theorems 4 and 6) under $\mathcal{M}_1$ in finite horizons. Essentially, we just need to be careful about choosing ω_N. We focus on the analysis of the cross-fitting version.

Theorem 4 (Asymptotic Property of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$). *Define κ_N^v, κ_N^q such that $\|\hat{v}_t^{[j]} - v_t\|_2 \le \kappa_N^v$, $\|\hat{q}_t^{[j]} - q_t^{\omega_N}\|_2 \le \kappa_N^q$ for $0 \le t \le \omega_N, j = 0, 1$. Assume* (4a) *$v_t \le C^t$ and $\gamma C < 1$ for some $C > 0$,* (4b) *$0 \le \hat{q}_t^{[j]} \le (1-\gamma)^{-1} R_{\max}$ and $0 \le \hat{v}_t^{[j]} \le C^t$ for the aforementioned C and $0 \le t \le \omega_N, j = 0, 1$,* (4c) *$(\kappa_N^v \vee \kappa_N^q)\omega_N = \mathrm{o}_p(1)$,* (4d) *$\omega_N = \omega(\log N)$,* (4e) *$\kappa_N^v \kappa_N^q \omega_N = \mathrm{o}_p(N^{-1/2})$. Then, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ is RAL and efficient; in particular, $\sqrt{N}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)} - \rho^{\pi_e}) \xrightarrow{d} \mathcal{N}(0, \mathrm{EB}(\mathcal{M}_1))$.*

Each assumption has the following interpretation. Condition (4a) is sufficient to guarantee that the EB is finite (see Corollary 1). Conditions (4b), (4c) are required to control a term related to a stochastic equicontinutiy condition. In particular, even if we observe infinitely long trajectories ($T = \infty$) we cannot set $\omega_N = \infty$. Notably, with cross-fitting, we make no assumptions about our nuisance estimates except for rates, meaning we can use black box machine learning methods that may not satisfy strong metric entropy conditions. Without cross-fitting, the same theorem would hold if we additionally impose a Donsker condition on $\hat{v}_t, \hat{q}_t$ but such would be restrictive on the types of estimators allowed (see Definition EC.1 in the online appendix for a definition of Donsker). Condition (4d) is needed so that $\rho_{\omega_N}^{\pi_e} = \rho^{\pi_e} + o(1/\sqrt{N})$. Condition (4e) is needed to show the inflation in variance due to using plug-in estimates is $\mathrm{o}_p(N^{-1/2})$, that is, the asymptotic variance is not changed because of the plug-in. Because of the mixed bias property (Rotnitzky et al. 2021) of the influence function, the rate is multiplicative in the two estimators' convergence rate. Finally, note that if we know the behavior policy we can take $\kappa_N^v = 0$ so the conditions on κ_N^q are very lax. If the behavior policy is not known, we can still allow very slow rates; for example, if $\omega_N = \log^{1+\epsilon} N$, $\kappa_N^v = N^{-\zeta_v}$, $\kappa_N^q = N^{-\zeta_q}$, then we only need the rates to satisfy $\zeta_v + \zeta_q > \frac{1}{2}$, $\zeta_v \vee \zeta_q > 0$, $\epsilon > 0$.

3.2. Time-Variant Markov Decision Process

In finite horizons, Kallus and Uehara (2020a) proposed the first efficient OPE estimator under TMDP. We now repeat the process in the previous section and show the results can be easily extended to the infinite-horizon case. Fix some horizon truncation ω_N. Let $q_t^{\omega_N} = \mathbb{E}_{\pi_e}\left[\sum_{k=t}^{\omega_N} \gamma^{t-k} r_t \mid s_t, a_t\right]$, $v_t^{\omega_N} = \mathbb{E}_{\pi_e}\left[\sum_{k=t}^{\omega_N} \gamma^{t-k} r_t \mid s_t\right]$. The estimator is given by

$$\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)} = c_{w_N}(\gamma) \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{\omega_N} \gamma^t (\hat{\mu}_t^{(i)}(s_t^{(i)}, a_t^{(i)})(r_t^{(i)} - \hat{q}_t^{(i)}(s_t^{(i)}, a_t^{(i)})) + \hat{\mu}_{t-1}^{(i)}(s_{t-1}^{(i)}, a_{t-1}^{(i)}) \hat{v}^{(i)}(s_t^{(i)})),$$

where $\hat{\mu}_t^{(i)}, \hat{q}_t^{(i)}$ are some plug-in estimates of $\mu_t, q_t^{\omega_N}$ to be used for the ith observation and $\hat{v}_t^{(i)}(s_t) = \mathrm{E}_{a_t \sim \pi_e(\cdot \mid s_t)}\left[\hat{q}_t(s_t, a_t) \mid s_t\right]$, which is an integral over $a \sim \pi_e(\cdot \mid s_t^{(i)})$, which is known. Again, $\mu_t^{(i)}, q_t^{(i)}$ can be estimated adaptively or using cross-fitting as in Section 3.1. Kallus and Uehara (2020a, section 6) discuss strategies for estimating $\mu_t, q_t^{\omega_N}$, that is, q-functions for finite-horizon problems.

We can again state a straightforward infinite-horizon extension of the efficiency result of Kallus and Uehara (2020a, theorem 9) under $\mathcal{M}_2$ in finite horizons. We focus on the analysis of the cross-fitting version.

Theorem 5 (Asymptotic Property of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$). *Define κ_N^μ, κ_N^q such that $\|\hat{\mu}_t^{[j]} - \mu_t\|_2 \le \kappa_N^\mu$, $\|\hat{q}_t^{[j]} - q_t^{\omega_N}\|_2 \le \kappa_N^q$ for $0 \le t \le \omega_N, j = 0, 1$. Assume* (5a) *$\mu_t \le C'^t$ and $\gamma C' < 1$ for some $C' > 0$,* (5b) *$0 \le \hat{q}_t^{[j]} \le (1-\gamma)^{-1} R_{\max}$ and $0 \le \hat{\mu}_t^{[j]} \le C'^t$ for the aforementioned C' and $0 \le t \le \omega_N, j = 0, 1$.* (5c) *$(\kappa_N^\mu \vee \kappa_N^q)\omega_N = \mathrm{o}_p(1)$,* (5d) *$\omega_N = \omega(\log N)$,* (5e) *$\kappa_N^\mu \kappa_N^q \omega_N = \mathrm{o}_p(N^{-1/2})$. Then, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$ is RAL and efficient; in particular, $\sqrt{N}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)} - \rho^{\pi_e}) \xrightarrow{d} \mathcal{N}(0, \mathrm{EB}(\mathcal{M}_2))$.*

Again, the estimate is feasible as long as we observe trajectories of length $\omega(\log N)$, and the cross-fitted version makes no assumption on nuisance estimates except rates. And, again, we can allow very slow rates: if $\omega_N = \log^{1+\epsilon} N$, $\kappa_N^\mu = N^{-\zeta_\mu}$, $\kappa_N^q = N^{-\zeta_q}$, then we only need the rates to satisfy $\zeta_\mu + \zeta_q > \frac{1}{2}$, $\zeta_\mu \vee \zeta_q > 0$, $\epsilon > 0$.

3.3. Inefficiency Under MDP

The methods in this section could be applied to an MDP. In fact, many papers using DR-type methods such as $\hat{\rho}_{\mathrm{DR}}$ (equal to the adaptive version of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$) assume that the underlying distribution is MDP when estimating q-functions: that is, they fit q-functions that depend only on s_t, a_t and that are time invariant. However, using this additional structure in order to produce better q-function estimates does not improve the asymptotic variance. Indeed, even if we used the oracle q-functions and oracle density ratios, we still only obtain the efficiency bounds in Theorems 4 and 5. Thus, even though we might use a total of $\mathcal{O}(NT)$

transition observations to get better q-function estimates, if we use standard DR-type methods, this will get washed out, at least asymptotically, and our variance will only vanish as $\mathcal{O}(1/N)$.

4. Efficient Estimator for Markov Decision Process

In this section, we propose an estimator that is efficient under the MDP model by leveraging the EIF obtained in Theorem 3. To our knowledge it is the first such estimator. We consider both the transition-sampling and trajectory-sampling settings and show that, under appropriate conditions in each setting, we achieve the same efficiency bound derived in Theorem 3 asymptotically. Specifically, the conditions in the trajectory-sampling setting include certain sufficient mixing so that dependent-data observations that sufficiently far apart appear near independent. We nonetheless need to develop a special sample-splitting procedure to handle the dependent data in this setting.

For brevity, we focus here on the case where the behavior policy is known, which is more relevant in RL. That is, we have that $\eta(s,a)$ is known. Our results can easily be extended to the unknown behavior policy case as well (see Remark 6).

4.1. Efficient Estimation Under Transition Sampling

The key to our estimator is the following estimating function, defined for a given w- and q-function:

$$\psi(s,a,r,s';w',q') = (1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[v'(s_0)] + w'(s)\eta(s,a)\big(r+\gamma v'(s')-q'(s,a)\big),$$

where we use w', q' to denote dummy such functions and use the shorthand that, given any q', we let $v'(s) = \mathrm{E}_{\pi_e}[q'(s,a)\mid s] = \int_a q'(s,a)\pi_e(a\mid s)d\lambda_{\mathcal{A}}(a)$, which is computable as an integral of q' wrt the known π_e (a sum if $\mathcal{A}$ is finite). Similarly, given q', the term $(\mathrm{E}_{p^{(0)}_{\pi_e}}[v'(s_0)])$ in the equation is also computable as both $p^{(0)}_{\pi_e}$ and π_e are known. Notice this term is also constant wrt (s,a,r,s'). This estimating function is derived from the EIF in Theorem 3: when $q'=q$, $w'=w$, we have $\psi(s,a,r,s';w,q) = \rho^{\pi_e} + \phi_{\mathrm{eff}}(s,a,r,s')$.

Based on this estimating function, our estimator is

$$\begin{aligned}\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} &= \mathbb{P}_n[\psi(s,a,r,s';\hat{w}^{(i)},\hat{q}^{(i)})]\\ &= \frac{1}{n}\sum_{i=1}^{n}(1-\gamma)\mathrm{E}_{s_0\sim p^{(0)}_{\pi_e}}[\hat{v}^{(i)}(s_0)]\\ &\quad+\frac{1}{n}\sum_{i=1}^{n}\hat{w}^{(i)}(s^{(i)})\eta(a^{(i)},s^{(i)})(r^{(i)}+\gamma\hat{v}^{(i)}(s'^{(i)})\\ &\quad-\hat{q}^{(i)}(s^{(i)},a^{(i)})),\end{aligned} \tag{4}$$

where $\hat{w}^{(i)}, \hat{q}^{(i)}$ are some plug-in estimates of w, q to be used for the ith observation. Recall $\hat{v}^{(i)}$ is defined in terms of $\hat{q}^{(i)}$ by taking expectations over $a\sim\pi_e(\cdot\mid s)$. Again, we consider two cases. First, we consider an adaptive version, where we let $\hat{w}^{(i)}=\hat{w}$, $\hat{q}^{(i)}=\hat{q}$ be shared among all data points and be estimated on the whole data set of n observations of (s,a,r,s'). Second, we consider a cross-fitting estimator, where we split the n observations into two even folds and $\hat{w}^{(i)}$, $\hat{q}^{(i)}$ are shared by all points i in the same fold and are estimated on data only on the opposite fold. The specific steps of the cross-fitting procedure are as in Section 3.1. Namely, we have four estimators: $\hat{w}^{[0]}$, $\hat{q}^{[0]}$, $\hat{w}^{[1]}$, $\hat{q}^{[1]}$. The first two are fit on one half of the data and the latter two on the other, and $\hat{w}^{(i)}, \hat{q}^{(i)}$ are set to those fit on the half not containing i. Unless otherwise specified, we always refer to the cross-fitting version.

The key to showing efficiency of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ is establishing the doubly robust (or, mixed bias) structure of $\psi(s,a,r,s';w',q')$, namely, that its expectation remains ρ^{π_e} whether just $w'=w$ or just $q'=q$. Suppose that $q'=q$. Then,

$$\begin{aligned}&\mathbb{E}[\mathbb{P}_n[\psi(s,a,r,s';w',q)]]\\ &\quad=(1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[v(s_0)]\\ &\qquad+\mathrm{E}_{p_{\pi_b}}[w'(s)\eta(s,a)\{r-q(s,a)+\gamma v(s')\}]\\ &\quad=(1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[v(s_0)]=\rho^{\pi_e}.\end{aligned} \tag{5}$$

Heuristically, this suggests that if $\hat{q}^{(i)}\to q$ and $\hat{w}^{(i)}\to w'$, where generally $w'\neq w$, then we expect that $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}\to\rho^{\pi_e}$. This viewpoint paints the estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ as given by taking the direct method and adding a control variate term.

On the other hand, if $w'=w$, then we have that

$$\begin{aligned}&\mathbb{E}[\mathbb{P}_n[\psi(s,a,r,s';w,q')]]\\ &\quad=\mathrm{E}_{p_{\pi_b}}[w(s)\eta(s,a)r]\\ &\qquad+\mathrm{E}_{p_{\pi_b}}[w(s)\{-\eta(s,a)q'(s,a)+\gamma\eta(s,a)v'(s')\}]\\ &\qquad+(1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[v'(s_0)]\end{aligned} \tag{6}$$

$$\begin{aligned}&\quad=\mathrm{E}_{p_{\pi_b}}[w(s)\eta(s,a)r]\\ &\qquad+\mathrm{E}_{p_{\pi_b}}[w(s)\{-\eta(s,a)q'(s,a)+v'(s)\}]\end{aligned} \tag{7}$$

$$\quad=\mathrm{E}_{p_{\pi_b}}[w(s)\eta(s,a)r]=\rho^{\pi_e}. \tag{8}$$

Note that from Equation (6) to Equation (7), we have used that for any $f_w(s)$ (see Lemma 1):

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

Heuristically, this suggests that if $\hat{w}^{(i)}\to w$ and $\hat{q}^{(i)}\to q'$, where generally $q'\neq q$, then we expect that $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}\to\rho^{\pi_e}$. Together, Equations (5) and (8) show that $\mathbb{P}_n[\psi(s,a,r,s';w',q')]$ has zero Gâteaux derivative in w',q' in any direction at $w'=w, q'=q$, a property known as Neyman orthogonality (Chernozhukov et al. 2018).

We now proceed to prove formally the efficiency and double robustness of our estimator.

Theorem 6 (Efficiency of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ Under Transition Sampling: Cross-Fitting). *Define κ_n^w, κ_n^q such that $\|\hat{w}^{[j]}-w\|_2\le\kappa_n^w$ and $\|\hat{q}^{[j]}-q\|_2\le\kappa_n^q$ for $j=0,1$. Assume* (6a) *there exist constants $C_w, C_{S'}>0$ such that $w\le C_w$ and $p_{b,S'}(\cdot)/$*

$p_{b,S}(\cdot) \le C_{S'}$, *where* $p_{b,S'}(\cdot)$ *and* $p_{b,S}(\cdot)$ *are marginal densities of* $p_{\pi_b}(s,a,r,s')$ *wrt* s' *and* s, *(6b)* $0 \le \hat{q}^{[j]} \le (1-\gamma)^{-1}R_{\max}$ *and* $0 \le \hat{w}^{[j]} \le C_w$ *for* $j = 0, 1$, *(6c)* $\kappa_n^w \vee \kappa_n^q = o_p(1)$, *and (6d)* $\kappa_n^w \kappa_n^q = o_p(n^{-1/2})$. *Then,* $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ *is RAL and efficient; in particular,* $\sqrt{n}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} - \rho^{\pi_e}) \xrightarrow{d} \mathcal{N}(0, \mathrm{EB}(\mathcal{M}_3))$.

The result essentially follows by showing that $|\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} - \mathbb{P}_n[\psi(s,a,r,s';w,q)]| = \mathcal{O}_p(\kappa_n^w \kappa_n^q) + o_p(n^{-1/2})$. Under the previous rate assumptions, the right-hand side is $o_p(n^{-1/2})$ and the result is immediately concluded from the central limit theorem (CLT). Here, using cross-fitting, we are able to completely avoid any restriction on our plug-in estimators, except for requiring a slow rate. In particular, the rate can be subparametric, that is, slower than square root. Crucially, this allows us to potentially use any nonparametric black box machine learning method, whether we can ensure good metric entropy conditions or not.

The adaptive version requires additional metric entropy conditions on the estimators. Let $\mathcal{N}(\tau, \mathcal{F}, \|\cdot\|_\infty)$ be the τ-covering number of $\mathcal{F}$ wrt L_∞ norm.

Theorem 7 (Efficiency of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ Under Transition Sampling: Adaptive). *Define* κ_n^w, κ_n^q *such that* $\|\hat{w} - w\|_2 \le \kappa_n^w$ *and* $\|\hat{q} - q\|_2 \le \kappa_n^q$. *Suppose the conditions of Theorem 6 hold and that in addition* $\hat{w} \in \mathcal{F}_w, \hat{q} \in \mathcal{F}_q$ *such that* $\log \mathcal{N}(\tau, \mathcal{F}_w, \|\cdot\|_\infty) = O(1/\tau^2)$, $\log \mathcal{N}(\tau, \mathcal{F}_q, \|\cdot\|_\infty) = O(1/\tau^2)$. *Then,* $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ *is RAL and efficient; in particular,* $\sqrt{n}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} - \rho^{\pi_e}) \xrightarrow{d} \mathcal{N}(0, \mathrm{EB}(\mathcal{M}_3))$.

Next, we formalize the notion of double robustness, which ensures our estimate is consistent even if we inconsistently estimate one of the components.

Theorem 8 (Double Robustness of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$). *Assume only conditions* (6a)–(6b) *of Theorem 6 hold. Assume further that* $\|\hat{w}^{[j]} - w^\dagger\|_2 = o_p(1)$ *and* $\|\hat{q}^{[j]} - q^\dagger\|_2 = o_p(1)$ *for some* $w^\dagger, q^\dagger$. *Then, as long as either* $w^\dagger = w$ *or* $q^\dagger = q$, *then we have that* $\mathrm{plim}_{n\to\infty} \hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} = \rho^{\pi_e}$. *The same holds for the adaptive version, if we further assume the metric entropy condition in Theorem 7.*

Theorem 8 does not provide a rate or an asymptotic distribution. We next strengthen the result (and, correspondingly, the conditions) to ensure a rate. This kind of double robustness is sometimes called model double robustness because the rates needed essentially correspond to parametric estimation and therefore the conditions essentially refer to whether these parametric models are well specified (Smucler et al. 2019).

Theorem 9 (Model Double Robustness of $\hat{\rho}_{DRL(\mathcal{M}_3)}$). *Assume only conditions* (6a)–(6b) *of Theorem 6 hold. If either* $\|\hat{q}^{[j]} - q^\dagger\|_2 = o_p(1)$, $\|\hat{w}^{[j]} - w\|_2 = \mathcal{O}_p(n^{-1/2})$ *or* $\|\hat{q}^{[j]} - q\|_2 = \mathcal{O}_p(n^{-1/2})$, $\|\hat{w}^{[j]} - w^\dagger\|_2 = o_p(1)$ *holds for* $j = 0,1$, *then* $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} = \rho^{\pi_e} + \mathcal{O}_p(n^{-1/2})$. *The same holds for the adaptive version, if we further assume the metric entropy condition in Theorem 7.*

Remark 6 (Unknown Behavior Policy). All of results are easily extended to the case where the behavior policy is unknown by replacing $\hat{w}(s)\eta(s,a)$ with $\hat{w}(s)\hat{\eta}(s,a)$, where $\hat{\eta}(s,a)$ is some estimator for $\eta(s,a)$, for example, $\pi_e(a \mid s)/\hat{\pi}^b(a \mid s)$, where $\hat{\pi}^b(a \mid s)$ is some estimator for the behavior policy. All of the results stay the same where conditions on $\|\hat{w} - w\|_2$ are simply replaced with the same conditions on $\|\hat{w}\hat{\eta} - w\eta\|_2 = \mathcal{O}(\|\hat{w} - w\|_2 + \|\hat{\eta} - \eta\|_2)$ instead.

Remark 7. After the first posted version of this paper, Tang et al. (2020) proposed a doubly robust-style estimator for the infinite-horizon MDP setting, which is given by taking a sample average of $\tilde{\psi}(s,a,r,s';\hat{w},\hat{v})$, where

$$\tilde{\psi}(s,a,r,s';w',v') = (1-\gamma)\mathrm{E}_{p_{\pi_e}^{(0)}}[v'(s_0)] + w'(s)\eta(s,a)(r + \gamma v'(s') - v'(s)),$$

and $\hat{w}, \hat{v}$ are adaptively estimated. The asymptotic behavior was not fully characterized, but following our work, (Kallus and Uehara 2020b, theorem 19) proved that if we impose Donsker conditions or if we use cross-fold estimates and under appropriate estimation rates (or, even if we plug-in oracle w, v), we can obtain that it is asymptotically normal with variance $\mathrm{var}[\tilde{\psi}(s,a,r,s';w,v)]$. This, however, is larger than $\mathrm{EB}(\mathcal{M}_3)$ by $\mathrm{E}[w^2(s)\mathrm{var}[\eta(s,a)\{r + \gamma v(s')\} \mid s]]$ (see Kallus and Uehara 2020b, section 6.3). That is, this estimator is not efficient, even in ideal oracle-nuisance settings. Moreover, it is only partially doubly robust in that it requires that π_b be well specified. In comparison, our estimator is in fact efficient and fully doubly robust.

4.2. Efficient Estimation Under Trajectory Sampling

We next study the trajectory-sampling setting and show that we can achieve the very same efficiency bound even though the transition data are dependent. All of our results apply to the asymptotic regime $T \to \infty$, where $N \ge 1$ is arbitrary, bounded, or growing. In particular, we can consider just a single, long trajectory ($N = 1$). Since the data are dependent, the standard notions of regular estimation do not apply; therefore, our efficiency statements are phrased solely in terms of showing that we can achieve the same asymptotic distribution of centered normal with variance equal to the efficiency bound corresponding to iid observations from the same stationary distribution.

Indexing the data as $\{(s_t^{(j)}, a_t^{(j)}, r_t^{(j)}, s_{t+1}'^{(j)})\}_{j=1,t=0}^{N,T}$ and identifying each (j, t) with a corresponding $i = 1, \dots, n$, where $n = N(T+1)$, we define our estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} = \mathbb{P}_N \mathbb{P}_T [\psi(s,a,r,s';\hat{w}^{(i)},\hat{q}^{(i)})]$. That is, the same as in Equation (4), taking an average of ψ over transitions with estimated w- and q-functions, but the transitions now are actually dependent observations. Because of this, we restrict our attention to the case where there is nonetheless sufficient mixing. We also need to be more careful when constructing cross-fitting estimates.

Letting $x_t^{(j)} = (s_t^{(j)}, a_t^{(j)}, r_t^{(j)}, s_t^{(j+1)})$, recall that we assume that $x_0^{(j)}, x_1^{(j)}, \dots$ forms a stationary process for each $j = 1, \dots, N$, that is, whereas these are dependent, the marginal distribution of each has an identical distribution. In the results that follow, we further assume that far-apart observations are less dependent, that is, the effect of earlier states gets washed away the farther ahead we look. To measure the level of such dependence we use the standard mixing coefficients $\alpha_m, \beta_m, \phi_m, \rho_m$, each of which measures the dependence between $x_0^{(j)}, \dots, x_t^{(j)}$ and $x_{t+m}^{(j)}, x_{t+m+1}^{(j)}, \dots$ using different metrics of dependence (taking worst-case over t). For example, α_m is the total variation distance between the joint distribution of the two subsequences and the product of their marginals. Since these are standard we relegate their definitions to online Appendix B.1. The coefficients are related via $2\alpha_m \le \beta_m \le \phi_m$, $4\alpha_m \le \rho_m \le 2\phi_m^{1/2}$, so α_m is weakest and ϕ_m is (almost) strongest (Bradley 2005).

Before we proceed to discuss feasible estimators, we show that despite dependent data, our estimating function retains its efficiency structure under sufficient mixing.

Theorem 10 (Efficiency Structure Under Mixing). *Suppose* $\sum_{m=1}^{\infty} \alpha_m < \infty$ *and* $w \le C_w$ *for some* $C_w > 0$. *Then we have* $\sqrt{NT}(\mathbb{P}_N \mathbb{P}_T[\psi(s,a,r,s';w,q)] - \rho^{\pi_e}) \xrightarrow{d} \mathcal{N}(0, \mathrm{EB}(\mathcal{M}_3))$.

The α-mixing condition in Theorem 10 is used in order to invoke a stationary-process CLT (Ibragimov and Linnik 1971, theorem 18.5.4). However, such a CLT still involves covariances across time, which would inflate the asymptotic variance. The key structural aspect of $\mathbb{P}_N \mathbb{P}_T[\psi(s,a,r,s';w,q)]$ that enables the result is that when we use the oracle q-function, the variables being time-averaged in the second term in Equation (4) form a martingale difference sequence, which ensures zero covariances across time. This occurs by virtue of the fact that the conditional expectation of the term inside the parentheses is zero by the definition of q. This essentially yields the result after some algebra. In terms of showing efficiency of a feasible (rather than oracle) estimator, what remains is to show that our estimator is equal to the oracle average up in Theorem 10 to errors that are $o_p((NT)^{-1/2})$.

Remark 8 (Relaxing Stationarity by Ergodicity). Assuming that $p_{\pi_b}^{(0)}$ is invariant so that $x_0^{(j)}, x_1^{(j)}, \dots$ is stationary is purely technical. It can easily be replaced by assuming ergodicity instead, so that the initial state distribution is irrelevant and we only approach stationarity. Namely, note $x_0^{(j)}, x_1^{(j)}, \dots$ forms a Markov chain. If it is a positive Harris chain (for definition, see Meyn and Tweedie 2009), then proposition 17.1.6 in Meyn and Tweedie (2009) guarantees that any CLT that holds when the initial state distribution is invariant also holds for any initial state distribution. This is simply because ergodicity means the initial state distribution gets washed away, asymptotically. All our results in this section proceed by showing $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} = \mathbb{P}_N \mathbb{P}_T [\psi(s,a,r,s';w,q)] + o_p((NT)^{-1/2})$ and then applying a mixing-process CLT on the dependent but stationary process in the first term. Each time, per that proposition, we can assume a positive Harris chain instead of stationarity, let the denominator of w be the invariant distribution, and define all mixing coefficients wrt the chain starting from the invariant distribution, and then this CLT will still hold and our characterizations of the asymptotic distribution of the estimator will still hold (see also Jones 2004, remark 6). Since this can always be done, we focus our analysis on stationary processes for generality.

We next analyze such feasible estimators, considering three cases: adaptive, cross-fitted with $N > 1$, and cross-fitted with $N = 1$. The difficulty with the latter case is that the data consists of a single, long trajectory, so any way we split the data, we will still have some dependence between the folds, undermining the standard cross-fitting technique. For each cross-fitting estimator, we define a segmentation of our n observations into folds and estimate w- and q-functions separately in each fold. If $N \ge 2$, we can split our data into folds across trajectories. Let $\mathcal{D}_0, \mathcal{D}_1$ be a random even partition of $\{1, \dots, N\}$ and fit $\hat{w}^{[j]}, \hat{q}^{[j]}$ in each fold separately (see Figure 2(a)). We then set $\hat{w}^{(i)}, \hat{q}^{(i)}$ to the estimates $\hat{w}^{[1-j]}, \hat{q}^{[1-j]}$ fit only on $\mathcal{D}_{1-j}$ where j is such that $\mathcal{D}_j$ contains the trajectory for observation i. We refer to this case as *cross-trajectory-fitting*. The benefit of this approach is that we have perfect independence across the folds because trajectories are independent. Recall that we used a similar strategy in the transition-sampling setting. Unfortunately, this is not possible when $n = 1$. In this case, we propose the following alternative. Let $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ be a random even partition of $\{0, \dots, T\}$ and fit $\hat{w}^{[j]}, \hat{q}^{[j]}$ in each fold separately (see Figure 2(b)). We then set $\hat{w}^{(i)}, \hat{q}^{(i)}$ to the estimates $\hat{w}^{[(j+2) \bmod 4]}, \hat{q}^{[(j+2) \bmod 4]}$ fit only on $\mathcal{T}_{(j+2) \bmod 4}$ where j is

such that $t \in \mathcal{T}_j$. Thus, we always use nuisances estimated on a fold that is not adjacent to the tth data point. We refer to this case as *cross-time-fitting*. Although we do not have perfect independence between folds, under sufficient mixing, nonadjacent folds will be sufficiently near independent, asymptotically.

First, we analyze the case of the cross-trajectory-fitting version, where we can avoid complex metric entropy assumptions by virtue of the unique structure of our estimator.

Theorem 11 (Efficiency of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ with Cross-Trajectory-Fitting). *Define* κ_n^w, κ_n^q *such that* $\|\hat{w}^{[j]} - w\|_2 \le \kappa_n^w$ *and* $\|\hat{q}^{[j]} - q\|_2 \le \kappa_n^q$ *for* $j = 0, 1$. *Assume* (11a) $\sum_{k=1}^{\infty} \rho_k < \infty$, (11b) $w \le C_w$ *for some* $C_w > 0$, (11c) $0 \le \hat{q}^{[j]} \le (1-\gamma)^{-1} R_{\max}$ *and* $0 \le \hat{w}^{[j]} \le C_w$, (11d) $\kappa_n^w \vee \kappa_n^q = o_p(1)$, *and* (11e) $\kappa_n^w \kappa_n^q = o_p(n^{-1/2})$. *Then,* $\sqrt{NT}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} - \rho^{\pi_e}) \xrightarrow{d} \mathcal{N}(0, \mathrm{EB}(\mathcal{M}_3))$.

Notice that condition (11a) is slightly stronger than the mixing condition in Theorem 10. The other conditions match Theorem 6.

Cross-trajectory-fitting is only feasible for $N \ge 2$ (although N need not grow). If $N = 1$, we instead proposed cross-time-fitting, which we analyze next.

Theorem 12 (Efficiency of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ with Cross-Time-Fitting). *Define* κ_n^w, κ_n^q *such that* $\|\hat{w}^{[j]} - w\|_2 \le \kappa_n^w$ *and* $\|\hat{q}^{[j]} - q\|_2 \le \kappa_n^q$ *for* $j = 0, 1, 2, 3$. *Assume* (12a) $\phi_t^{1/2} = O(1/t^{2+\epsilon})$ *for some* $\epsilon > 0$, (12b) $w \le C_w$ *for some* $C_w > 0$, (12c) $0 \le \hat{q}^{[j]} \le (1-\gamma)^{-1} R_{\max}$ *and* $0 \le \hat{w}^{[j]} \le C_w$, (12d) $\kappa_n^w \vee \kappa_n^q = o_p(1)$, *and* (12e) $\kappa_n^w \kappa_n^q = o_p(n^{-1/2})$. *Then,* $\sqrt{NT}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} - \rho^{\pi_e}) \xrightarrow{d} \mathcal{N}(0, \mathrm{EB}(\mathcal{M}_3))$.

In both Theorems 11 and 12, we are able to avoid strong conditions on the plug-in estimators we use aside from requiring a slow, subparametric convergence rate. We only require slightly stronger mixing conditions than the oracle case in Theorem 10.

Finally, for the adaptive version of our estimator, we need to control the metric entropy of our plug-in estimators. In particular, we suppose that we are given some class $\mathcal{F}_\psi$ that almost surely contains $\psi(\cdot, \cdot, \cdot, \cdot; \hat{w}, \hat{q})$. We let $J_{[]}(\infty, \mathcal{F}_\psi, L_p)$ be the bracketing integral wrt the L_p norm (for definition, see Kosorok 2008).

Theorem 13 (Efficiency of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ with In-Sample Fitting). *Define* κ_n^w, κ_n^q *such that* $\|\hat{w} - w\|_2 = \kappa_n^w$ *and* $\|\hat{q} - q\|_2 = \kappa_n^q$ *and fix some* $p > 2$. *Assume* (13a) $\sum_{m=1}^{\infty} m^{2/(p-2)} \beta_m < \infty$, (13b) $w \le C_w$ *for some* $C_w > 0$, (13c) $0 \le \hat{q} \le (1-\gamma)^{-1} R_{\max}$ *and* $0 \le \hat{w} \le C_w$, (13d) $\kappa_n^w \vee \kappa_n^q = o_p(1)$, (13e) $\kappa_n^w \kappa_n^q = o_p(n^{-1/2})$, *and* (13f) $J_{[]}(\infty, \mathcal{F}_\psi, L_p(p_{\pi_b}^{\infty})) < \infty$. *Then,* $\sqrt{NT}(\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)} - \rho^{\pi_e}) \xrightarrow{d} \mathcal{N}(0, \mathrm{EB}(\mathcal{M}_3))$.

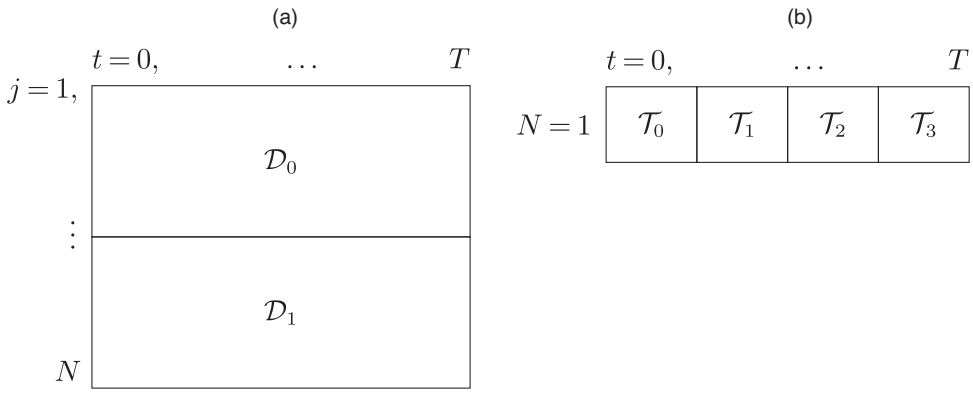
To prove this, we invoke a uniform central limit theorem for β-mixing sequences (Kosorok 2008, theorem 11.24). Because of in-sample fitting, we require condition (13f) in order to control a term corresponding to a stochastic equicontinuity condition.

Remark 9 (When Stationarity Fails). In this section, we assumed the data are stationary, or at least eventually stationary as in Remark 8. But such may not apply to problems with absorbing states, as we study in Section 8.2. But even without stationarity, we can still view the data as transitions $(s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)})$, $i = 1, \ldots, NT$, drawn (nonindependently) from:`

$$\left(\frac{1}{T}\sum_{t=1}^{T} p_b^{(t)}(s,a)\right) p(r|s,a) p(s'|s,a).$$

If the effective state-action distribution $\frac{1}{T}\sum_{t=1}^{T} p_b^{(t)}(s,a)$ has good coverage and $N \to \infty$, we should still expect convergence, and our DRL estimator is still using the best estimating function in the sense that it is still the least-norm gradient of the estimand, as a function of the T-long trajectories. Nonetheless, due to the dependence of transitions in the same trajectory and without

Figure 2. Arrangement of Folds for Cross-Fitting of Nuisances for DRL in $\mathcal{M}_3$



Note. (a) Two folds over $N \ge 2$ trajectories; (b) four folds over a single trajectory.

stationarity and mixing, it is difficult to theoretically characterize the rate of the MSE in T.

The remaining question is how to consistently estimate q and w, especially from a single trajectory. We discuss how to estimate w in Section 6 and how to estimate q in Section 7. We first discuss how our results in this section lend themselves directly to constructing confidence intervals.

5. Asymptotically Valid Confidence Intervals

We are often interested in confidence intervals in addition to point estimates. Our asymptotic normality results lend themselves directly to the construction of such. Namely, all we have to do is consistently estimate the asymptotic variance. If an estimator $\hat{\rho}_n$ satisfies $\sqrt{n}(\hat{\rho}_n - \rho^{\pi_e}) \to \mathcal{N}(0, V)$ and we have a consistent variance estimator $\hat{V}_n \to V$, then we will always have that $\mathbb{P}\left(|\hat{\rho}_n - \rho^{\pi_e}| \le \Phi^{-1}(1-\alpha/2)\sqrt{\hat{V}/n}\right) \to 1-\alpha$, where Φ^{-1} is the inverse cumulative distribution function of the standard normal (e.g., for $\alpha = 0.05$, $\Phi^{-1}(1-\alpha/2) \approx 1.96$). This means that the confidence interval $\left[\hat{\rho}_n - \Phi^{-1}(1-\alpha/2)\sqrt{\hat{V}/n}, \hat{\rho}_n + \Phi^{-1}(1-\alpha/2)\sqrt{\hat{V}/n}\right]$ has asymptotic coverage exactly $1-\alpha$. By Theorems 4, 5, 6, 11, 12, and 13, it then suffices to estimate $\mathrm{EB}(\mathcal{M}_1)$, $\mathrm{EB}(\mathcal{M}_2)$, $\mathrm{EB}(\mathcal{M}_3)$ to construct asymptotically valid confidence intervals.

Focusing on $\mathrm{EB}(\mathcal{M}_3)$ and the transition-sampling setting, we propose the following estimator:

$$\widehat{\mathrm{EB}}(\mathcal{M}_3) = \mathbb{P}_n[(\psi(s,a,r,s';\hat{w}^{(i)},\hat{q}^{(i)}) - \hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)})^2],$$

that is, the sample variance of $\psi(s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)}; \hat{w}^{(i)}, \hat{q}^{(i)})$. This estimate is consistent under the same conditions as in Theorem 6.

Theorem 14. *Under the conditions of Theorem 6,*

$$\widehat{\mathrm{EB}}(\mathcal{M}_3) \xrightarrow{p} \mathrm{EB}(\mathcal{M}_3).$$

A similar result holds in $\mathcal{M}_1$ and $\mathcal{M}_2$. In each case, our estimators, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$ and $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$, were constructed as sample averages of cross-fitted estimates of the corresponding EIF plus the estimand. Taking the sample variance corresponding to this sample average, we again obtain a consistent variance estimator that we can use to construct asymptotically valid confidence intervals.

Note that since our estimators are efficient, one cannot improve on these confidence intervals, asymptotically. More formally, a test based on an efficient estimator is automatically locally uniformly powerful in the sense that the power function defined in a neighborhood of the true data-generating process attains the upper bound (see van der Vaart 1998, lemma 25.45).

6. Modeling the Ratio of Average Visitation Distributions

Our DRL estimator in $\mathcal{M}_3$ relied on having an estimator for the ratio of average visitation distributions, $w(s)$. In this section, we discuss its estimation from semiparametric inference perspective. These estimates can then be plugged into $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$.

6.1. Importance Sampling Using Stationary Density Ratios

Before discussing how to estimate $w(s)$, we consider an IS-type estimator for MDPs using $w(s)$. We can transform our DRL estimator to an IS-type estimator by simply choosing $\hat{q}^{(i)} = 0$. This leads to the marginalized importance sampling (MIS) estimator

$$\hat{\rho}_{\mathrm{MIS}} = \mathbb{P}_n[\eta(s,a)\hat{w}(s)r], \quad \hat{w}(s) \approx w(s). \tag{9}$$

where "$\approx$" means "estimating."

Note that this is different from the IS estimator proposed by Liu et al. (2018a), which is defined as an empirical approximation of

$$\mathrm{E}_{p^{(\infty)}_{\pi_b,\gamma}}[\eta(s,a)\hat{\tilde{w}}(s)r], \quad \hat{\tilde{w}}(s) \approx \tilde{w}(s) = \frac{p^{(\infty)}_{\pi_e,\gamma}(s)}{p^{(\infty)}_{\pi_b,\gamma}(s)}. \tag{10}$$

The difference between the two methods is that we use $p^{(0)}_{\pi_b}(s)$ instead of $p^{(\infty)}_{\pi_b,\gamma}(s)$ in the denominator of the density ratio. In the transition-sampling setting, $p^{(0)}_{\pi_b}(s)$ in Equation (9) can be anything. In the trajectory-sampling setting, the denominator is an invariant distribution, or is the stationary distribution $p^{(\infty)}_{\pi_b}(s)$ if we consider the ergodic case (see Remark 8), which is still different from $p^{(\infty)}_{\pi_b,\gamma}(s)$. There are a few benefits to this. Intuitively, since we see samples from $p^{(\infty)}_{\pi_b}$, using Equation (9) can be more efficient because, to get a sample from the distribution $p^{(\infty)}_{\pi_b,\gamma}$, we would essentially have to throw away $(1-\gamma)$ fraction of our samples. Indeed, the performance of Equation (10) behaves badly when $\gamma < 1$ (Liu et al. 2018a, figure 3(d)).

Nonetheless, unlike $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ as in Section 4.2, the estimator $\hat{\rho}_{\mathrm{MIS}}$ does not have a martingale difference structure. This means that the covariance terms across the time in the CLT do not drop out, potentially inflating the variance of the $\mathbb{P}_T$ average in the trajectory-sampling setting. Moreover, because it lacks a doubly robust structure, there is an inflation term due to the plug-in of an estimate, $\hat{w}$, of w, unlike $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$. This occurs even if the estimate has a parametric rate, $\|\hat{w} - w\|_2 = \mathcal{O}_p(n^{-1/2})$, because there is no mixed bias structure to cancel it out. These two reasons make it difficult to analyze the asymptotic MSE of $\hat{\rho}_{\mathrm{MIS}}$. They also suggest the estimator is not efficient.

6.2. Efficient Semiparametric Estimation

The remaining question is how to estimate $w(s) = p^{(\infty)}_{\pi_e,\gamma'}(s)/p^{(0)}_{\pi_b}(s)$. Here, we take a semiparametric approach. First, we consider a characterization of $w(s)$ by modifying theorem 4 in Liu et al. (2018a). We obtain the following lemma.

Lemma 1 (Characterization of $w(s)$). *Define*

$$L(w',f_w) = \mathrm{E}[\gamma w'(s)\eta(s,a)f_w(s') - w'(s)f_w(s)] + (1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[f_w(s)]. \tag{11}$$

Then, for $w' = w$, we have $L(w',f_w) = 0$ for any f_w. Conversely, if $L(w',f_w) = 0$ for all λ_S-square-integrable functions f_w and there is a unique solution g to the integral equation

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

then $w'(s) = w(s)$.

Again, this holds for any $p^{(0)}_{\pi_b}(s)$. This is the difference from (Liu et al. 2018a, theorem 4), which only holds for $p^{(0)}_{\pi_b}(s) = p^{(\infty)}_{\pi_b,\gamma'}(s)$. When $p^{(0)}_{\pi_b}(s)$ is an invariant distribution, as in the trajectory-sampling setting, $L(w',f_w)$ is equal to

$$\mathrm{E}[\gamma w'(s)\eta(s,a)f_w(s') - w'(s')f_w(s')] + (1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[f_w(s)]. \tag{12}$$

Thus, in this case, the condition that $L(w',f_w) = 0$ for all f_w is equivalent to the conditional moment equation:

$$\mathrm{E}\left[w(s)\eta(s,a) - w(s') + (1-\gamma)\frac{p^{(0)}_{\pi_e}(s')}{p^{(0)}_{\pi_b}(s')}\,\middle|\, s'\right] = 0. \tag{13}$$

Note this is not a standard moment equation since it still depends on the unknown quantity $p^{(0)}_{\pi_b}(s)$. This is closely related to a similar key relation of $\mu_k(s_k)$ used in Section 3.2, namely, $\mathrm{E}[\nu_{k-1} \mid s_k] = \mu_k(s_k)$, which implies

$$\mathrm{E}[\mu_{k-1}(s_{k-1})\eta(a_{k-1},s_{k-1}) - \mu_k(s_k) \mid s_k] = 0. \tag{14}$$

For derivation, refer to (Kallus and Uehara 2020a, section 3). Heuristically, taking a limit as $k \to \infty$, replacing $\lim_{k\to\infty}\mu_k(s)$ with $w(s)$, and setting $\gamma = 1$, we get Equation (13). Notice that in Equation (14), we obtain μ_k from μ_{k-1}, whereas in Equation (13) we obtain w from itself, that is, it solves a fixed-point equation. This change is analogous to the change in q-equations between the time-variant finite-horizon problem and the time-invariant infinite-horizon problem.

Suppose first that we assume a parametric model $w(s) = w(s;\beta^*)$. Then, β^* can be estimated as a solution to an empirical approximation of Equation (11), that is,

$$\mathbb{P}_n[\gamma w(s;\beta)\eta(s,a)f_w(s') - w(s;\beta)f_w(s)] + (1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[f_w(s)] = 0, \tag{15}$$

for some vector-valued function f_w. We denote the estimator as $\hat{\beta}_{f_w}$. Note $\mathrm{E}_{p^{(0)}_{\pi_e}}[f_w(s)]$ can be exactly calculated because $p^{(0)}_{\pi_e}$ is known.

Example 1 (Linear Regression Approach). Consider a case when our model is linear in some features of s, that is, $w(s;\beta) = \beta^\top\psi(s)$. Then, as in linear regression, a natural choice for $f_w(s)$ is $\psi(s)$. The estimator of $\hat{\beta}_\psi$ is constructed as the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\psi(s^{(i)})(\gamma\eta(s^{(i)},a^{(i)})\psi^\top(s(i)) - \psi^\top(s^{(i)}))\beta + (1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[\psi(s)] = 0.$$

In the finite-state-space setting, we can use $\psi(s) = (\mathrm{I}(s^{*1} = s),\ldots,\mathrm{I}(s^{*d} = s))^\top$, where $\mathcal{S} = \{s^{*1},\ldots,s^{*d}\}$.

More generally, for a linear or nonlinear model, under the correct specification assumption, that is, there exists β^* such that $w(s) = w(s;\beta^*)$, we have the following efficient estimation result. We focus on the transition-sampling setting.

Theorem 15 (Efficient Estimation of $w(s;\beta^*)$ Under Transition Sampling). *Define*

$$\Delta_{f_w}(s,a,s';\beta) = w(s;\beta)\{\gamma\eta(s,a)f_w(s') - f_w(s)\}.$$

Suppose $\mathbb{E}\sup_{\beta\in\Theta_\beta}\|\Delta_{f_w}(s,a,s';\beta)\| < \infty$, where Θ_β is a parameter space for β. Assume $w(s) = w(s;\beta^*)$ for some $\beta^* \in \Theta_\beta$ and that a vector-valued f_w is given such that $L(w(s;\beta),f_w) = 0 \iff \beta = \beta^*$. Further assume standard regularity conditions: Θ_β is compact, β^* is in its interior, $w(s;\beta)$ is a C^2-function with respect to β with first and second derivatives uniformly bounded, and for any α with $\|\alpha\| = 1$, we have $\mathrm{E}[|\alpha^\top\Delta_{f_w}(s,a,s';\beta)|^{2+\epsilon}]|_{\beta=\beta^*} < \infty$ for some $\epsilon > 0$. Then, the asymptotic variance of $\hat{\beta}_{f_w}$ is

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

Importantly, regardless of the choice of f_w, the rate of $\|w(s;\hat{\beta}_{f_w}) - w(s)\|_2$ will be $\mathcal{O}_p(n^{-1/2})$. Compared with the usual conditional moment equation setting (Chen 2007), the efficient choice of f_w to minimize the asymptotic variance here is unclear because $p^{(0)}_b(s)$ is unknown.

Because of the doubly robust structure of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$, it did not matter how we estimated w as long as we had a (subparametric) rate. This is not true for $\hat{\rho}_{\mathrm{MIS}}$. We can, however, derive its asymptotics for the particular estimation approach given in Equation (15).

Theorem 16 (Asymptotic Property of $\hat{\rho}_{\mathrm{MIS}}$). *Suppose the conditions of Theorem 15 hold and that $\mathbb{G}_n[r\eta(s,a)w(s;\hat{\beta}_{f_w})] - \mathbb{G}_n[r\eta(s,a)w(s;\beta^*)] = \mathrm{o}_p(1)$, where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathrm{E})$ is the empirical process. Then, $\sqrt{n}(\hat{\rho}_{\mathrm{MIS}} - \rho^{\pi_e}) \xrightarrow{d} \mathcal{N}(0, V_{\mathrm{MIS}})$*

where

$$V_{\mathrm{MIS}} = \mathrm{var}[w(s;\beta)\eta(s,a)r + \mathrm{E}[\nabla_{\beta^\top} w(s;\beta)\eta(s,a)r]\mathrm{E}[\nabla_{\beta^\top}\Delta_{f_w}(s,a,s';\beta)]^{-1} \Delta_{f_w}(s,a,s';\beta)]|_{\beta=\beta^*}. \quad (16)$$

Note the technical condition $\mathbb{G}_n[r\eta(s,a)w(s;\hat{\beta}_{f_w})] - \mathbb{G}_n[r\eta(s,a)w(s;\beta^*)] = \mathrm{o}_p(1)$ can potentially be verified as in the proofs of Theorems 11 and 13.

7. Modeling the *q*-Function

In this section, we discuss from a semiparametric inference perspective how to estimate the q-function in an off-policy manner, potentially from only one trajectory. Our approach can be seen as a generalization of LSTDQ (Lagoudakis and Parr 2004). The estimated q-function we obtain can be used in our estimator, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$.

By definition, the q-function is characterized as a solution to

$$q(s,a) = \mathrm{E}[r \mid s,a] + \gamma \mathrm{E}[\mathrm{E}_{a'\sim\pi_e}[q(s',a') \mid s'] \mid s,a].$$

Assume a parametric model for the q-function, $q(s,a) = q(s,a;\beta)$. Then, the parameter β can be estimated using the following recursive estimating equation:

$$\mathrm{E}\big[e_{\mathrm{q}}(s,a,r,s';\beta) \mid s,a\big] = 0,$$
$$\text{where } e_{\mathrm{q}}(s,a,r,s';\beta) = r + \gamma \mathrm{E}_{a'\sim\pi_e}\big[q(s',a';\beta) \mid s'\big] - q(s,a;\beta).$$

This implies that for any function $f_q(s,a)$,

$$\mathrm{E}[f_q(s,a)e_{\mathrm{q}}(s,a,r,s';\beta)] = 0. \quad (17)$$

More specifically, given a vector-valued $f_q(s,a)$, we can define an estimator $\hat{\beta}_{f_q}$ as the solution to

$$\mathbb{P}_n[f_q(s,a)e_{\mathrm{q}}(s,a,r,s';\beta)] = 0. \quad (18)$$

Example 2 (LSTDQ). When $q(s,a;\beta) = \beta^\top\psi(s,a)$ and $f_q(s,a) = \psi(s,a)$, this leads to the LSTDQ method (Lagoudakis and Parr 2004):

$$\left(\sum_{i=1}^n \psi(s^{(i)},a^{(i)})[\psi^\top(s^{(i)},a^{(i)}) - \gamma \mathrm{E}_{a\sim\pi_e}\{\psi^\top(s'^{(i)},a) \mid s^{(i)}\}]\right)^{-1} \left\{\sum_{i=1}^n r^{(i)}\psi(s^{(i)},a^{(i)})\right\} = 0.$$

More generally, for a linear or nonlinear model, under the correct specification assumption, that is, that there exists some β^* such that $q(s,a) = q(s,a;\beta^*)$, we have the following result. We again focus on the transition-sampling setting.

Theorem 17 (Efficient Estimation of $q(s,a;\beta)$ Under Transition Sampling). *Suppose* $\mathbb{E}\sup_{\beta\in\Theta_\beta}\|e_q(s,a,r,s';\beta) f_q(s,a)\| < \infty$, *where* Θ_β *is a parameter space for* β. *Assume* $q(s,a) = q(s,a;\beta)$ *for some* $\beta \in \Theta_\beta$ *and that a vector-valued* f_q *is given such that (Equation (17) holds)* $\iff \beta = \beta^*$. *Further, assume standard regularity conditions:* Θ_β *is compact,* β^* *is in its interior,* $q(s,a;\beta)$ *is* C^2*-function with respect to* β *with first and second derivatives uniformly bounded, and for any* α *with* $\|\alpha\| = 1$, *we have* $\mathrm{E}[|e_q(s,a,r,s';\beta)\alpha^\top f_q(s,a)|^{2+\epsilon}]|_{\beta=\beta^*} > 0$ *for some* $\epsilon > 0$. *The lower bound for the asymptotic MSE for estimating* β^* *scaled by* n *is*

$$V_\beta = \mathrm{E}[\nabla_\beta m_q(s,a;\beta)v_q^{-1}(s,a;\beta)\nabla_{\beta^\top} m_q(s,a;\beta)]^{-1}|_{\beta=\beta^*},$$

where $m_q(s,a;\beta) = \mathrm{E}[e_q(s,a,r,s';\beta) \mid s,a]$, $v_q(s,a) = \mathrm{var}[e_q(s,a,r,s';\beta) \mid s,a]$.

This bound is achieved when

$$f_q(s,a) = \nabla_\beta m_q(s,a;\beta)v_q^{-1}(s,a;\beta)|_{\beta=\beta^*}. \quad (19)$$

Importantly, regardless of the choice of f_q, the rate $\|q(\cdot,\cdot;\hat{\beta}_{f_q}) - q\|_2$ is $\mathcal{O}_p(n^{-1/2})$. Nonetheless, efficient estimation is preferred. Practically, we do not know the efficient f_q in Equation (19). One way is parametrically estimating it and another way is a sieve generalized method of moments (GMM) estimator, using a basis expansion for f_q (Hahn 1997).

We can also extend the approach to achieve nonparametric estimation of q. This most easily done by extending the LSTDQ approach in Example 2. We simply let $q(s,a;\beta_n) = \sum_{j=1}^{d_n}\beta_j\psi_j(s,a)$, where $\psi_1,\psi_2,\ldots$ is a basis expansion of L^2 and $d_n \to \infty$ as we collect more data. Given regularity conditions and smoothness conditions on q, we can obtain rates on $\|q(\cdot,\cdot;\hat{\beta}_n) - q\|_2$ without assuming correct parametric specification (Chen and Shen 1998). This provides a means to estimate q for $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$, either parametrically or nonparametrically.

If we use q estimated parametrically as in Equation (18), we can also establish the asymptotic behavior of $\hat{\rho}_{\mathrm{DM}}$. Again, as in the case of $\hat{\rho}_{\mathrm{MIS}}$, because $\hat{\rho}_{\mathrm{DM}}$ lacks the doubly robust structure, we must have parametric rates on q-estimation in order to achieve $1/n$ MSE scaling in Theorem 18, unlike the case of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ where q-estimation can have slow nonparametric rates.

Theorem 18 (Asymptotic Property of $\hat{\rho}_{\mathrm{DM}}$). *Let* $\hat{\rho}_{\mathrm{DM}} = (1-\gamma)\mathrm{E}_{p^{(0)}_{\pi_e}}[\mathrm{E}_{\pi_e}\{q(s,a;\hat{\beta}_{f_q}) \mid s\}]$. *Suppose the assumptions of Theorem 17 hold. Then,* $\sqrt{n}(\hat{\rho}_{\mathrm{DM}} - \rho^{\pi_e}) \xrightarrow{d} \mathcal{N}(0, V_{\mathrm{DM}})$ *where*

$$V_{\mathrm{DM}} = (1-\gamma)^2\mathrm{E}_{p^{(0)}_{\pi_e}}[\mathrm{E}_{\pi_e}[\nabla_{\beta^\top}q(s,a;\beta) \mid s]] V_\beta \mathrm{E}_{p^{(0)}_{\pi_e}}[\mathrm{E}_{\pi_e}[\nabla_\beta q(s,a;\beta) \mid s]]|_{\beta=\beta^*}.$$

Interestingly, this is smaller than or equal to the efficiency bound in $\mathcal{M}_3$. This is not a contradiction since $\hat{\rho}_{\mathrm{DM}}$ as in Theorem 18 is not regular wrt $\mathcal{M}_3$ as it

assumes the well-specification of the parametric model $q(s,a;\beta)$, which leads to the smaller model than $\mathcal{M}_3$.

Lemma 2. $V_{\mathrm{DM}} \le \mathrm{EB}(\mathcal{M}_3)$.

This result is well-known in the bandit setting when we use a binary deterministic policy (Tan 2007). Our result can be seen as its generalization to the more complex MDP setting.

Remark 10. Ueno et al. (2011) and Luckett et al. (2020) considered related semiparametric estimation techniques for the v-function. Compared with that, our focus is a q-function estimation rather than a value function estimation. Note many traditional TD-type methods (Sutton and Barto 2018), including LSTD(λ) (Boyan 1999, Nedić and Bertsekas 2003), gradient temporal difference (GTD) learning (Sutton et al. 2009a), temporal difference learning with gradient correction (TDC) (Sutton et al. 2009b), and off-policy LSTD (Yu 2012) are also defined as the solution to estimating equations as in Equation (17). For details, refer to Ueno et al. (2011) and Yu et al. (2018). The asymptotic MSEs of these methods can be calculated as in Theorem 17.

8. Experimental Results

In this section, we conduct experiments to compare our method with existing off-policy evaluation methods. We consider a simpler setting that perfectly fits the theory and a more challenging setting that requires some function approximation.

8.1. Taxi Environment

First we consider the taxi environment and focus on simple w- and q-estimators in order to illustrate the doubly robust property of our method. For detail on this environment, see Liu et al. (2018a).

We set our target evaluation policy to be the final policy $\pi_e = \pi^*$ after running q-learning for 1,000 iterations. We set another policy π_+ as the result after 150 iterations. The behavior policy is then defined as $\pi_b = \alpha\pi^* + (1-\alpha)\pi_+$, where we range α to vary the overlap. We show results for $\alpha = 0.2, 0.6$ here and provide additional results for $\alpha = 0.4, 0.8$ in online Appendix D. We consider the case with the behavior policy known and set $\gamma = 0.98$. Note that this π^*, π_+ are fixed in each setting.

We estimate all w-functions following Example 1. For q-functions, we use a value iteration for the approximated MDP based on the empirical distribution. Then, we compare $\hat{\rho}_{\mathrm{IS}}$, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$, $\hat{\rho}_{\mathrm{MIS}}$, $\hat{\rho}_{\mathrm{DM}}$, and $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$. We consider observing a single trajectory ($n = 1$) of increasing length T, $T \in [50{,}000, 1{,}00{,}000, 2{,}00{,}000, 4{,}00{,}000]$. For each, we consider 200 replications. Note that we use adaptive (in-sample) fitting and not cross-fitting because $n = 1$. In addition, we do not compare with a marginalized importance sampling estimator or to $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$ because μ_t cannot be estimated with $n = 1$ (e.g., the empirical estimated marginal importance $\hat{\mu}_t$ is just ν_t).

To study the effect of double robust property, we consider three settings:

1. Both w-model and q-model are correct.
2. Only w-model is correct: we add noise $\mathcal{N}(1.0, 1.0)$ to $\hat{q}(s,a)$.
3. Only q-model is correct: we add noise $\mathcal{N}(1.0, 1.0)$ to $\hat{w}(s)$.

8.1.1. Results and Discussion.

We report the resulting MSE over the replications for each estimator in each setting in Figures 3–8.

First, we note that the estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ handily outperforms the standard IS and DR estimators, $\hat{\rho}_{\mathrm{IS}}$, $\hat{\rho}_{\mathrm{DR}}$, in every setting. This is owed to the fact that these do not leverage the MDP structure. The competitive comparison is of course to DM and MIS.

We find that, in the large-sample regime, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ dominates all other estimators across all settings. First,

Figure 3. (Color online) Setting (1) with $\alpha = 0.2$

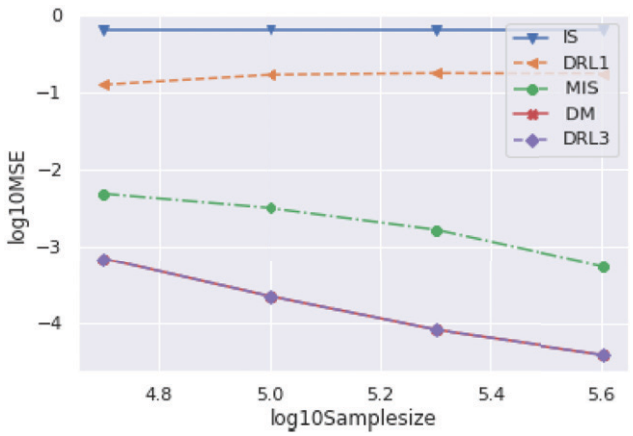


Figure 4. (Color online) Setting (1) with $\alpha = 0.6$

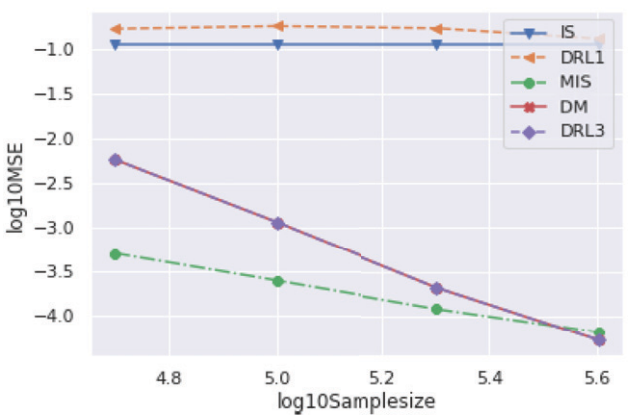


Figure 5. (Color online) Setting (2) with $\alpha = 0.2$

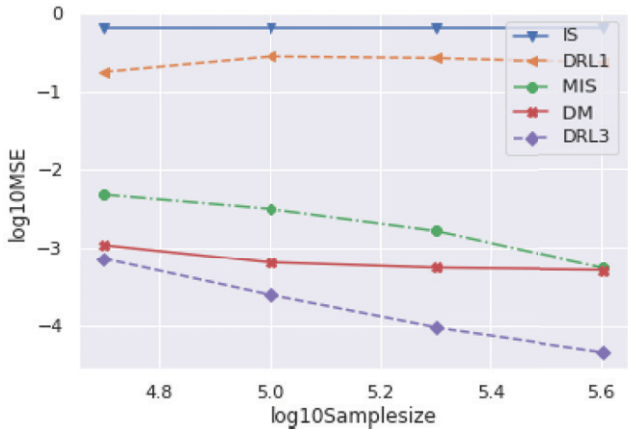
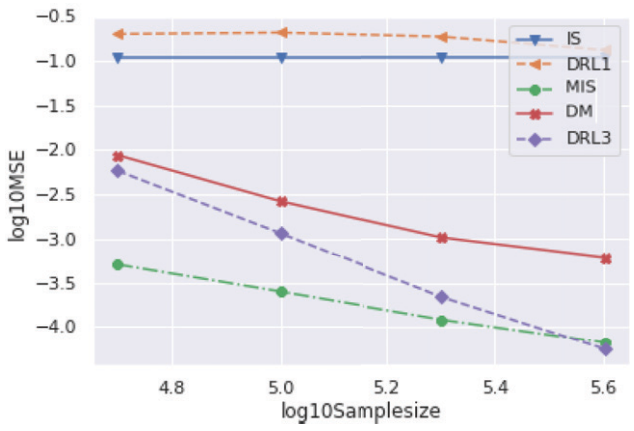


Figure 6. (Color online) Setting (2) with $\alpha = 0.6$



for $T = 4{,}00{,}000$, it has the lowest MSE among all estimators for each setting. Second, whereas in some settings it has MSE similar to another method, it beats it handily in another setting. Compared with DM, the MSE is similar when the q-function is well specified but $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ does much better when q is ill-specified. Compared with MIS, the MSE is similar when both the w-function is well specified and there is good overlap but $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ performs much better when either specification or overlap fails. This is of course owed to the doubly robust structure and the efficiency of $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$.

In the small-to-medium sample regime, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ performs the best among all estimators except when overlap is good ($\alpha = 0.6$) and w is well specified (settings (2) and (3)). In these cases, for the small-to-medium sample regime, MIS performs better. However, as in the large-sample regime, it performs much worse in small-to-medium samples too when overlap is bad or when w is misspecified. In particular, in setting (2) with $\alpha = 0.2$, $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ has performance much better than all other estimators across the sample-size regimes.

Because having either parametric misspecification or nonparametric rates for $\hat{w}$ and $\hat{q}$ is unavoidable in practice (for continuous state-action spaces), the estimator $\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_3)}$ is superior. This is doubly true when overlap can be weak.

8.2. CartPole Environment

We next conduct an experiment in the CartPole environment based on the implementation of OpenAI gym (Brockman et al. 2016). In the CartPole environment, the state space is continuous and four-dimensional and the action space is binary. Thus, we require flexible models for w and q and may not be able to guarantee their precise convergence. Moreover, the environment has an absorbing state and therefore our trajectories are highly nonstationary, yet we show our method still works in practice as suggested by Remark 9.

We set the target and behavior policy in the following way. First, we run deep Q-network (DQN) in an online interaction with the environment to learn q^*, following OpenAI's default implementation.[4] Then, based on q^*, we define a range of softmax policies given by a

Figure 7. (Color online) Setting (3) with $\alpha = 0.2$

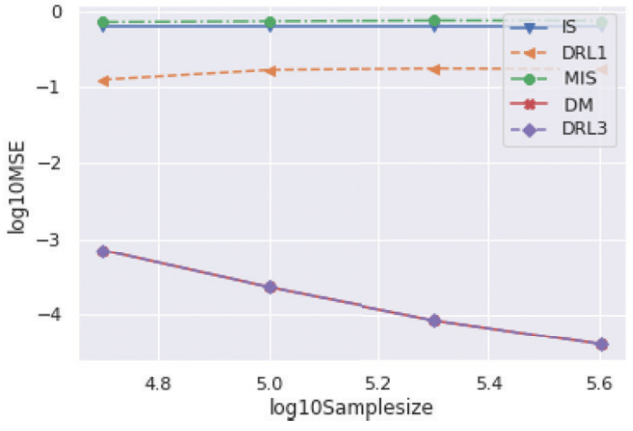


Figure 8. (Color online) Setting (3) with $\alpha = 0.6$

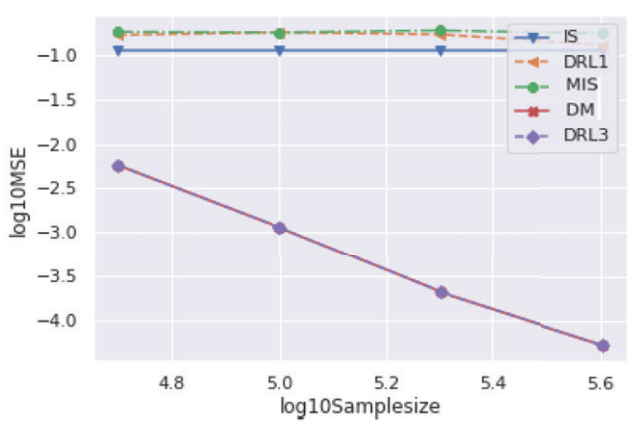


Figure 9. (Color online) CartPole: $\tau = 1.3$ and N Varying

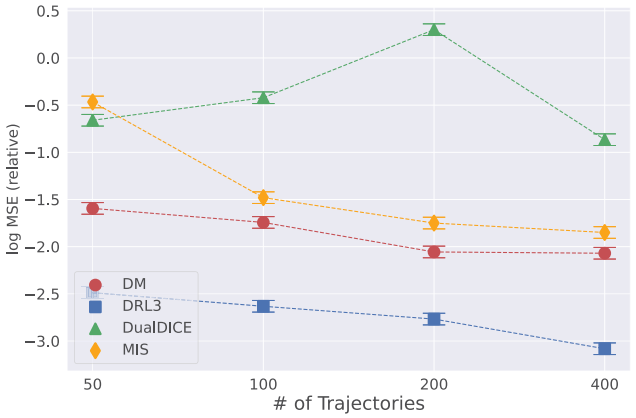
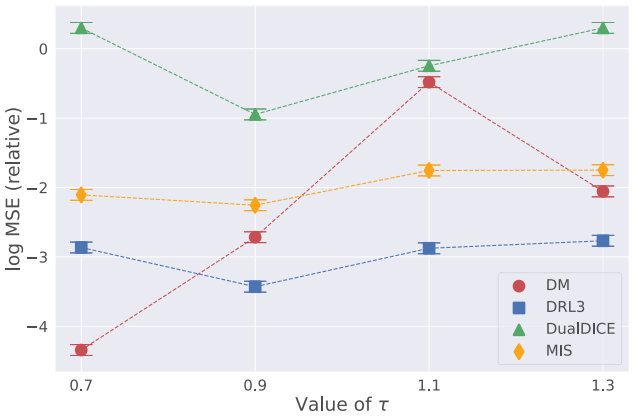


Figure 10. (Color online) CartPole: $n = 200$ and τ Varying



temperature parameter τ: $\pi(a \mid s:\tau) \propto \exp(Q(s,a)/\tau)$. We then set the behavior policy as $\pi_b(a \mid s) = \pi(a \mid s:1.0)$, and we consider a variety of evaluation policies $\pi_e(a \mid s) = \pi(a \mid s:\tau)$ for $\tau \in [0.7, 0.9, 1.1, 1.3]$. The training data set is generated by executing the behavior policy with a fixed horizon length $T = 1{,}000$. Specifically, if the agent visits the terminal absorbing states before 1,000 steps, the rest of the trajectory will consist of repeating the last state. We consider observing $N \in [50, 100, 200, 400]$ trajectories, that is, $n \in [50, 100, 200, 400] \times 1{,}000$ transitions.

We estimate w using a minimax approach leveraging Equation (11). Namely, we consider a model $w(s;\beta)$ given a neural network with 32 units in each, ReLU activations for hidden layers, and a softplus activation for the output to ensure nonnegative output. Then, we fit the weights β by minimizing the maximum of the left-hand side of Equation (15) over all f_w in the unit ball of the reproducing kernel Hilbert space (RKHS) with the Gaussian kernel $k(x_i, x_j) = \exp(-\|x_i - x_j\|^2/(2\sigma^2))$. We similarly estimate q leveraging Equation (18). We again use the same neural network architecture for $q(s,a;\beta)$ except that the input has one more dimension and we do not apply an activation to the output. We again consider f_q in the same RKHS unit ball (but with one more input dimension). For both w- and q-estimation, we normalize all data to have mean zero and unit variance and set the length-scale parameter σ to the median of pairwise distances in the data. We use Adam to optimize the neural networks and set the leaning rate to 0.005.

We compare MIS ($\hat{\rho}_{\mathrm{MIS}}$), DM ($\hat{\rho}_{\mathrm{DM}}$) and DRL3 ($\hat{\rho}_{\mathcal{M}_3}$) using these w- and q-estimators. We also compare these to DualDICE (Nachum et al. 2019), which is a variant of the MIS estimator. In DualDICE, the w estimator is based on a different minimax objective function using two neural networks. We choose hyperparameters to be the same as in the implementation of Uehara et al. (2020).

8.2.1. Results and Discussion. We run 40 replications of the experiment for each τ and N and consider the MSE of each algorithm relative to $(\rho^{\pi_e} - \rho^{\pi_b})^2$. To estimate the latter normalizer, we estimate each of $\rho^{\pi_e}, \rho^{\pi_b}$ as a simple sample average using 1,000 on-policy trajectories. This normalization enhances interpretability as we vary τ.

In Figure 9, we report the results for varying N and fixing $\tau = 1.3$ and in Figure 10 for varying τ and fixing $N = 200$. We show the relative MSEs on a logarithmic scale with 90% confidence intervals. We observe that DRL clearly outperforms the other estimators. This can be attributed to the fact that both w- and q-estimators are flexible, and hence have high variance, which influences the variance of both MIS and DM, respectively, whereas DRL is largely insensitive to the particular w- and q-estimators. One exception is $\tau = 0.7, N = 200$, where we see DM performs better than DR. On the other hand, MIS always performs worse than DR. This would suggest that w-estimation is more difficult than q-estimation in this environment, possibly because of the nonstationarity of the data. Finally, we note DualDICE performs consistently badly across the settings, which can be attributed to the instability of the minimax optimization of two neural networks involved in its w-estimation.

9. Conclusions

We established the efficiency bound for OPE in a time-invariant Markov decision process in the regime where N is (potentially) finite and $T \to \infty$. This novel lower bound quantifies how fast one could hope to estimate policy value in a model usually assumed in RL. According to our results, many IS and DR OPE estimators used in RL are in fact not leveraging this structure to the fullest and are inefficient. This leads to MSE that is suboptimal in rate, not just in leading coefficient. We instead proposed the first efficient estimator achieving the efficiency bound, and enjoy a double robustness property at the same time. We

hope our work inspires others to further develop estimators that build on ours by leveraging MDP structure as we have here and perhaps combining this with ideas such as balancing (Kallus 2018), stability (Kallus and Uehara 2019), or blending (Thomas and Brunskill 2016) that can improve the finite-sample performance in addition to our asymptotic efficiency.

Endnotes

[1] OPE can also sometimes refer to estimating the whole value or quality function of a policy; here we focus on estimating the mean reward.

[2] Whereas in control settings one often needs to restrict to standard Borel measurable spaces due to measurability issues when optimizing (Hernández-Lerma and Lasserre 2012), since we are only considering evaluation, we do not require such a restriction.

[3] In greatest generality, MDPs need not restrict the conditional next-state s' distributions to be absolutely continuous with respect to the same base measure for all state-action pairs s, a. The same is true for reward and policy distributions. We make this restriction here to be able to easily consider perturbations to the MDP distributions in a semiparametric framework.

[4] See https://github.com/openai/baselines.

References

Antos A, Szepesvári C, Munos R (2008) Fitted Q-iteration in continuous action-space MDPs. *Adv. Neural Inform. Processing Systems* 20:9–16.

Bertsekas DP (2012) *Dynamic Programming and Optimal Control*, 4th ed. Athena Scientific Optimization and Computation Series (Athena Scientific, Belmont, MA).

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

Boyan JA (1999) Least-squares temporal difference learning. *Proc. Internat. Conf. Machine Learn.* (ICML), 49–56.

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

Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W (2016) OpenAI gym. Preprint, submitted June 5, https://arxiv.org/abs/1606.01540.

Chen X (2007) Large sample sieve estimation of semi-nonparametric models. Heckman JJ, Leamer EE, eds. *Handbook of Econometrics* (Elsevier, New York), 6:5549–5632.

Chen X, Shen X (1998) Sieve extremum estimates for weakly dependent data. *Econometrica* 66:289–314.

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

Dudik M, Erhan D, Langford J, Li L (2014) Doubly robust policy evaluation and optimization. *Statist. Sci.* 29:485–511.

Ertefaie A (2014) Constructing dynamic treatment regimes in infinite-horizon settings. Preprint, submitted June 3, https://arxiv.org/abs/1406.0764.

Farajtabar M, Chow Y, Ghavamzadeh M (2018) More robust doubly robust off-policy evaluation. *Proc. 35th Internat. Conf. Machine Learn.*, 1447–1456.

Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, Celi LA (2019) Guidelines for reinforcement learning in healthcare. *Nature Medicine* 25:16–18.

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

Hernández-Lerma O, Lasserre JB (2012) *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, vol. 30 (Springer Science & Business Media, New York)

Hirano K, Imbens G, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71:1161–1189.

Huang J, Jiang N (2020) From importance sampling to doubly robust policy gradient. *Proc. Internat. Conf. Machine Learn.* (PMLR), 4434–4443.

Ibragimov IA, Linnik YV (1971) *Independent and Stationary Sequences of Random Variables* (Walters-Noordhof, The Netherlands).

Jiang N, Li L (2016) Doubly robust off-policy value evaluation for reinforcement learning. *Proc. 33rd Internat. Conf. Machine Learn.*, 652–661.

Jones GL (2004) On the Markov chain central limit theorem. *Probab. Surveys* 1:299–320.

Kallus N (2018) Balanced policy evaluation and learning. *Adv. Neural Inform. Processing Systems* 31:8895–8906.

Kallus N, Uehara M (2019) Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. *Adv. Neural Inform. Processing Systems* 32:3320–3329.

Kallus N, Uehara M (2020a) Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Machine Learn. Res.* 21(167):1–63.

Kallus N, Uehara M (2020b) Efficient evaluation of natural stochastic policies in offline reinforcement learning. Preprint, submitted June 3, https://arxiv.org/abs/2006.03886.

Kallus N, Uehara M (2020c) Statistically efficient off-policy policy gradients. *Proc. Internat. Conf. Machine Learn.* (PMLR), 5089–5100.

Khan S, Tamer E (2010) Irregular identification, support conditions, and inverse weight estimation. *Econometrica* 78:2021–2042.

Klaassen CAJ (1987) Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.* 15:1548–1562.

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

Lagoudakis M, Parr R (2004) Least-squares policy iteration. *J. Machine Learn. Res.* 4:1107–1149.

Li L, Munos R, Szepesvari C (2015) Toward minimax off-policy value estimation. *Proc. 18th Internat. Conf. Artificial Intelligence Statist.* (San Diego, CA), 608–616.

Liu Q, Li L, Tang Z, Zhou D (2018a) Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Adv. Neural Inform. Processing Systems* 31:5356–5366.

Liu Y, Gottesman O, Raghu A, Komorowski M, Faisal AA, Doshi-Velez F, Brunskill E (2018b) Representation balancing MDPs for off-policy policy evaluation. *Adv. Neural Inform. Processing Systems* 31:2644–2653.

Luckett DJ, Laber EB, Kahkoska AR, Maahs DM, Mayer-Davis E, Kosorok MR (2020) Estimating dynamic treatment regimes in mobile health using v-learning. *J. Amer. Statist. Assoc.* 115(530): 692–706.

Mahmood AR, van Hasselt HP, Sutton RS (2014) Weighted importance sampling for off-policy learning with linear function approximation. *Adv. Neural Inform. Processing Systems* 27: 3014–3022.

Mandel T, Liu Y, Levine S, Brunskill E, Popovic Z (2014) Off-policy evaluation across representations with applications to educational games. *Proc. 13th Internat. Conf. Autonomous Agents Multiagent Systems*, 1077–1084.

Meyn S, Tweedie RL (2009) *Markov Chains and Stochastic Stability*, 2nd ed. (Cambridge University Press, New York).

Munos R, Stepleton T, Harutyunyan A, Bellemare M (2016) Safe and efficient off-policy reinforcement learning. *Adv. Neural Inform. Processing Systems* 29:1054–1062.

Murphy SA (2003) Optimal dynamic treatment regimes. *J. Royal Statist. Soc. Ser. B Statist. Methodology* 65:331–355.

Murphy SA, Van Der Laan MJ, Robins JM (2001) Marginal mean models for dynamic regimes. *J. Amer. Statist. Assoc.* 96: 1410–1423.

Nachum O, Chow Y, Dai B, Li L (2019) DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. *Adv. Neural Inform. Processing Systems*, vol. 32.

Nedić A, Bertsekas DP (2003) Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynam. Systems* 13:79–110.

Newey WK (1990) Semiparametric efficiency bounds. *J. Appl. Econometrics* 5(2):99–135.

Precup D, Sutton R, Singh S (2000) Eligibility traces for off-policy policy evaluation. *Proc. 17th Internat. Conf. Machine Learn.*, 759–766.

Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* 89:846–866.

Rotnitzky A, Smucler E, Robins J (2021) Characterization of parameters with a mixed bias property. *Biometrika* 108(1):231–238.

Scharfstein D, Rotnizky A, Robins JM (1999) Adjusting for non-ignorable dropout using semi-parametric models. *J. Amer. Statist. Assoc.* 94:1096–1146.

Smucler E, Rotnitzky A, Robins JM (2019) A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. Preprint, submitted April 7, https://arxiv.org/abs/1904.03737v1.

Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).

Sutton RS, Maei HR, Szepesvári C (2009a) A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. *Adv. Neural Inform. Processing Systems* 21:1609–1616.

Sutton RS, Maei HR, Precup D, Bhatnagar S, Silver D, Szepesvári C, Wiewiora E (2009b) Fast gradient-descent methods for temporal-difference learning with linear function approximation. *Proc. 26th Annual Internat. Conf. Machine Learn.* (ICML), 993–1000.

Tan Z (2007) Comment: Understanding OR, PS and DR. *Statist. Sci.* 22(4):560–568.

Tang Z, Feng Y, Li L, Zhou D, Liu Q (2020) Harnessing infinite-horizon off-policy evaluation: Double robustness via duality. *ICLR 2020*, 1–20. https://iclr.cc/virtual_2020/poster_S1glGANtDr.html.

Thomas P, Brunskill E (2016) Data-efficient off-policy policy evaluation for reinforcement learning. *Proc. 33rd Internat. Conf. Machine Learn.* 2139–2148.

Tsiatis AA (2006) *Semiparametric Theory and Missing Data.* Springer Series in Statistics (Springer, New York).

Uehara M, Huang J, Jiang N (2020) Minimax weight and q-function learning for off-policy evaluation. *Proc. Internat. Conf. Machine Learn.*, 9659–9668.

Ueno T, Kawanabe M, Mori T, Maeda SI, Ishii S (2011) Generalized TD learning. *J. Machine Learn. Res.* 12:1977–2020.

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

Xie T, Ma Y, Wang YX (2019) Toward optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Adv. Neural Inform. Processing Systems* 32: 9665–9675.

Yu H (2012) Least squares temporal difference methods: An analysis under general conditions. *SIAM J. Control Optim.* 50(6): 3310–3343.

Yu H, Mahmood AR, Sutton RS (2018) On generalized Bellman equations and temporal-difference learning. *J. Machine Learn. Res.* 19(1):1864–1912.

Zheng W, van Der Laan MJ (2011) Cross-validated targeted minimum-loss-based estimation. van Der Laan MJ, Rose S, eds. *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer Series in Statistics (Springer, New York), 459–474.

Nathan Kallus is an assistant professor in the School of Operations Research and Information Engineering and Cornell Tech at Cornell University. His research interests include optimization, especially under uncertainty and informed by data; causal inference; sequential decision making; and algorithmic fairness.

Masatoshi Uehara is a PhD candidate in the Department of Computer Science at Cornell University. His research interests include reinforcement learning and causal inference.