Biometrika (—-), —-, —-, *pp*. 1–25
Printed in Great Britain

Efficient Evaluation of Natural Stochastic Policies in Offline Reinforcement Learning

BY NATHAN KALLUS

Department of Operations Research and Information Engineering and Cornell Tech, Cornell University
2 W Loop Rd, New York, NY, USA
kallus@cornell.edu

MASATOSHI UEHARA

Department of Computer Science and Cornell Tech, Cornell University
2 W Loop Rd, New York, NY, USA
mu223@cornell.edu

SUMMARY

We study the efficient off-policy evaluation of natural stochastic policies, which are defined in terms of deviations from the unknown behavior policy. This is a departure from the literature on off-policy evaluation that largely consider the evaluation of explicitly specified policies. Crucially, offline reinforcement learning with natural stochastic policies can help alleviate issues of weak overlap, lead to policies that build upon current practice, and improve policies' implementability in practice. Compared with the classic case of a pre-specified evaluation policy, when evaluating natural stochastic policies, the efficiency bound, which measures the best-achievable estimation error, is inflated since the evaluation policy itself is unknown. In this paper we derive the efficiency bounds of two major types of natural stochastic policies: tilting policies and modified treatment policies. We then propose efficient nonparametric estimators that attain the efficiency bounds under lax conditions and enjoy a partial double robustness property.

Some key words: Off-policy evaluation; Dynamic treatment regime; Semiparametric inference; Mobile health

1. INTRODUCTION

In many emerging application domains for reinforcement learning, exploration is highly limited and simulation unreliable, such as in healthcare (Gottesman et al., 2019; Kosorok and Moodie, 2015). In these domains, we must use offline reinforcement learning, where we evaluate and learn new sequential decision policies from existing observational data. A key task in offline reinforcement learning is that of off-policy evaluation, in which we evaluate a new policy from data logged by another behavior policy (Bibaut et al., 2019; Kallus and Uehara, 2019b; Murphy, 2003; Nachum et al., 2019; Robins, 2004; Zhang et al., 2013). In many applications such as mobile health, the horizon, or number of decision stages, is often long (Boruvka et al., 2018; Liao et al., 2020; Luckett et al., 2018; Shi et al., 2020). For example, in the trial of Liao et al. (2020), the horizon is 450 while the number of patients is 37. In such settings, the naïve sequential importance sampling estimator (Precup et al., 2000; Robins et al., 1999) suffers from the curse of horizon (Liu et al., 2018) in the sense that its mean squared error grows exponen-

tially in the horizon. Recent work in off-policy evaluation (Kallus and Uehara, 2019a, 2020a) has shown how efficiently leveraging problem structure, such as the Markov property, which means the reward and next state and action are independent of all past observations conditional on the current state and action, and time-homogeneity which means the dynamical system does not change over time, can significantly improve off-policy evaluation and address issues such as the curse of horizon, *i.e.*, showing that the mean squared error grows only polynomially in the horizon. Avoiding exponential dependence has been recognized as a central problem in offline reinforcement learning (Agarwal et al., 2019; Wang et al., 2020).

In most of the literature on off-policy evaluation, including the above, the policy to be evaluated is pre-specified, that is, it is a given and known function from states to a distribution over actions. In a departure from this, in this paper we consider the evaluation of natural stochastic policies, which may depend on the natural value of the action, that is, the treatment that would be observed in the absence of intervention (Haneuse and Rotnitzky, 2013; Muñoz and Van Der Laan, 2012; Shpitser and Pearl, 2012). Specifically, we consider policies defined as deviations from the behavior policy that generated the observed data (also known as propensity scores; Rosenbaum, 1983). In situations where we do not know behavior policies, as we typically see in observational studies, these policies cannot be pre-specified, which poses a new challenge to derive statistically efficient algorithms. Throughout this paper, we focus on this setting with an unknown behavior policy and unknown target evaluation policy defined in terms of the unknown behavior policy.

There are several primary advantages to natural stochastic policies. The first advantage is implementability. Subjects are often unable or reluctant to undertake an assigned treatment if the deviation from the treatment they would have naturally undertaken is large. For example, consider intervening on leisure-time physical activity to reduce mortality among the elderly (as in Díaz and van der Laan, 2018). An evaluation policy assigning $a + \delta$ minutes of weekly activity to an individual whose current physical activity level is a, also known as the natural value, would be a realistic intervention for small to moderate δ. On the other hand, evaluation policies assigning any arbitrary level of physical activity level ignoring the current level of physical activity is unrealistic and rarely implementable. Another example is intervening on air pollution levels to improve the health of children (as in Díaz and van der Laan, 2013). A possible evaluation policy is enforcing the pollution levels below a certain cutoff point if the observed pollution level (*i.e.*, the natural value) exceeds the threshold. A second advantage is interpretation. The value of certain natural stochastic policies can easily be interpreted as the impact of a directional change. For instance, in the tilting policies proposed by Kennedy (2019), evaluation policies are defined in such a way that a parameter δ controls the odds ratio of the evaluation policies and the current behavior policies. This parameter reflects the magnitude of the intervention's effect on the likelihood of receiving treatment. A third advantage is that we can relax or more easily satisfy the positivity assumption, which requires some overlap between the evaluation and behavior policies and is fundamentally necessary for off-policy evaluation. Often, we cannot know a priori whether the positivity assumption is satisfied for a given intervention in an observational study or how good is the overlap. We can, however, easily consider policies that only deviate slightly from the behavior policy, ensuring a good overlap and reliable evaluation by design. Because of these benefits, natural stochastic policies are commonly applied in medical settings (Díaz and van der Laan, 2013, 2018; Young et al., 2014, 2019).

In this paper, we derive efficiency bounds and develop efficient estimators for two key types of natural stochastic policies: tilting policies and modified treatment policies. We consider three longitudinal settings: (1) non-Markov decision processes, (2) time-varying Markov decision processes, and (3) time-homogeneous Markov decision processes. We derive efficiency bounds in each setting, which quantify the statistical limits of evaluation by establishing the best-achievable

asymptotic mean squared error in each setting. In particular, due to the unknown evaluation policy, the efficiency bounds are inflated in comparison with the case of a pre-specified evaluation policy. Irrespective of this inflation, we show the efficiency bound is still polynomial in the horizon in the case of time-varying Markov decision processes and Markov decision processes. Importantly, this indicates the curse of the horizon is surmountable. In fact, we proceed to develop estimators that achieve these efficiency bounds under lax conditions. Our estimator has a unique partial double robustness property, which is different from the usual double robustness observed when evaluating pre-specified policies (Jiang and Li, 2016; Kallus and Uehara, 2020a; Robins et al., 1999). We also demonstrate how efficient estimators for pre-specified policies (Kallus and Uehara, 2019a, 2020a) break if we just naïvely plug in an estimate of the evaluation policy.

2. Related literature

Natural stochastic policies are widely studied in the non-sequential setting (Haneuse and Rotnitzky, 2013; Muñoz and Van Der Laan, 2012). However, they have not been extensively studied in the longitudinal setting but for a few exceptions. Kennedy (2019) considers off-policy evaluation with binary actions under a tilting policy in a non-Markov decision process. In comparison, we focus on the Markovian setting that is central to reinforcement learning and mobile health and where we can overcome the curse of horizon, and we also allow for a possibly continuous action space. Young et al. (2014) considers off-policy evaluation under a modified treatment policy in a non-Markov decision process using a parametric approach. In comparison, our methods are nonparametric and globally efficient, and we focus on the Markovian setting.

We emphasize that our setting is quite different from the standard setting in causal inference and policy evaluation, where the evaluation targets are pre-specified (Kallus and Uehara, 2019a; Narita et al., 2019; Robins et al., 1994). We discuss these differences in greater detail in Remarks 1 to 3 and Section 6.4. As in Haneuse and Rotnitzky (2013); Muñoz and Van Der Laan (2012), since the target functionals are more complicated due to the fact that evaluation policies are not pre-specified, we need much more careful analysis. Moreover, the efficient influence functions for evaluating natural stochastic policies do not generally have the usual doubly-robust structure observed in the pre-specified-policy case. In fact, our findings reveal that the efficient influence functions in our setting exhibit a partially doubly-robust structure. This means that certain nuisances need to be estimated consistently to ensure the off-policy evaluation is also consistent, albeit with possibly slow convergence rates. This special structure is explained in Theorem 3.

3. Setup and Background

3.1. Problem Setup and Definitions

Consider an H-long time-varying Markov decision process, with states, also known as covariates, $s_t \in \mathcal{S}_t$, actions $a_t \in \mathcal{A}_t$, rewards, also known as outcomes $r_t \in \mathbb{R}$, initial state distribution $p_1(s_1)$, transition distributions $p_{t+1}(s_{t+1} \mid s_t, a_t)$, and reward distributions $p_t(r_t \mid s_t, a_t)$, for $t = 1, \dots, H$. A policy $(\pi_t(a_t \mid s_t))_{t \le H}$ induces a distribution over trajectories $\mathcal{T} = (s_1, a_1, r_1, \dots, s_T, a_H, r_H, s_{H+1})$:

$$p_\pi(\mathcal{T}) = p_1(s_1) \prod_{t=1}^{H} \pi_t(a_t \mid s_t) p_t(r_t \mid s_t, a_t) p_{t+1}(s_{t+1} \mid s_t, a_t). \tag{1}$$

Given an evaluation policy π^{e}, which we consider as unknown in this paper, we are interested in its value, $J = \mathrm{E}_{p_{\pi^{\mathrm{e}}}}\left(\sum_{t=1}^{H} r_t\right)$, where the expectation is taken with respect to (with respect to

.) the density induced by the evaluation policy, $p_{\pi^{\mathrm{e}}}$. In the off-policy setting, our data consists of trajectory observations from some fixed policy, π^b, known as the behavior policy:

$$\mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(n)} \sim p_{\pi^b}, \quad \mathcal{T}^{(i)} = (S_1^{(i)}, A_1^{(i)}, R_1^{(i)}, \cdots, S_H^{(i)}, A_H^{(i)}, R_H^{(i)}). \qquad \text{(Off-policy data)}$$

In observational studies, as we consider herein, π^{b} is unknown, and the observed action $A_j^{(i)}$ is considered the natural value of the action in the sense that it is the one naturally observed in the absence of our intervention. Our goal is to estimate J from the observed data $\{\mathcal{T}^{(i)}\}_{i=1}^n$. Finally, while we do not make use of explicit counterfactual notation, our approach is equivalent to the counterfactual value of following π^{e} instead of π^{b} if we had employed potential outcomes and assumed the typical sequential ignorability and consistency assumptions (Ertefaie and Strawderman, 2018; Luckett et al., 2018).

We define the q- and v-functions for π^{e}, respectively, as

$$q_t = \mathrm{E}_{p_{\pi^{\mathrm{e}}}}\left(\textstyle\sum_{k=t}^H r_t \mid s_t, a_t\right),\ v_t = \mathrm{E}_{p_{\pi^{\mathrm{e}}}}\left(\textstyle\sum_{k=t}^H r_t \mid s_t\right).$$

Further, define the instantaneous (or, one-step), cumulative (or, multi-step), and marginal state density ratios and marginal state-action density ratio, respectively, as

$$\eta_t = \tfrac{\pi_t^e(a_t|s_t)}{\pi_t^b(a_t|s_t)}, \quad \lambda_t = \textstyle\prod_{k=1}^t \eta_k, \quad w_t = \tfrac{p_{\pi^e}(s_t)}{p_{\pi^b}(s_t)}, \quad \mu_t = \eta_t w_t,$$

where $p_\pi(s_t)$ is the marginal density at s_t under p_π. Here, we observe that η_t and λ_t are fundamental components of causal inference, often referred to as inverse probability weights, with their counterparts in longitudinal settings being described in Young et al. (2014, Equation 11). We note that in many causal inference settings, evaluation policies of interest are deterministic constant-action policies that remain invariant with respect to s_t. The marginal ratio w_t, in contrast, is special to reinforcement learning, where on assumes Markovianness, and it plays an important role in constructing statistically efficient estimators in time-varying Markov decision processes. We assume throughout the paper that $0 \le r_t \le R_{\max}$, $\eta_t \le C$, $w_t \le C'$, $\forall t \le H$. The latter two bounds constitute an overlap assumption.

Given trajectory data, $\mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(n)}$, we define the empirical expectation as $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(\mathcal{T}^{(i)})$. Unless otherwise noted, all expectations, variances, and probabilities are with respect to $p_{\pi^{\mathrm{b}}}$. Define the L_2 norm by $\|f\|_2 = [\mathrm{E}\{f^2(\mathcal{T})\}]^{1/2}$. We denote convergence in probability and distribution by $\xrightarrow{p}$ and $\xrightarrow{d}$, respectively.

3.2. *Natural Stochastic Policies*

In off-policy evaluation, π^{e} is often pre-specified. Our focus is the case where π^{e} depends on the natural value of the treatment in an observational study. Importantly, in this setting, both π^{e} and π^{b} are unknown. In this paper, we consider two types of natural stochastic policies: modified treatment policies and tilting policies. These constructions are inspired by previous work focusing on the non-dynamic and non-Markov decision process setting (Díaz and Hejazi, 2020; Haneuse and Rotnitzky, 2013), which implicitly require that the horizon is short (logarithmic in the amount of data) due to the curse of horizon mentioned in Section 1. We focus on time-varying Markov decision process and Markov decision process settings, which permit long horizons. This is practically significant since long horizons is a common feature of modern mobile health applications, which target chronic diseases such as obesity (Thomas and Bond, 2015), alcohol abuse (Gustafson et al., 2014), and nicotine addiction (Riley et al., 2008).

DEFINITION 1 (TILTING POLICY). *Given $u_t : \mathcal{A}_t \to \mathbb{R}$, a tilting policy is defined as*

$$\pi_t^{\mathrm{e}}(a_t \mid s_t) = u_t(a_t)\pi_t^{\mathrm{b}}(a_t \mid s_t) \Big/ \int u_t(\tilde{a}_t)\pi_t^{\mathrm{b}}(\tilde{a}_t \mid s_t)\mathrm{d}\tilde{a}_t. \tag{2}$$

Tilting policies tilt the behavior policy slightly toward actions with higher values of u_t. For example, for a binary action, letting $u_t(1) = \delta$, $u_t(0) = 1$ yields

$$\pi_t^{\mathrm{e}}(a_t \mid s_t) = \mathrm{I}(a_t = 1)\frac{\delta\pi_t^{\mathrm{b}}(1 \mid s_t)}{1 + (\delta - 1)\pi_t^{\mathrm{b}}(1 \mid s_t)} + \mathrm{I}(a_t = 0)\frac{\delta^{-1}\pi_t^{\mathrm{b}}(0 \mid s_t)}{1 + (\delta^{-1} - 1)\pi_t^{\mathrm{b}}(0 \mid s_t)}, \tag{3}$$

as considered by Kennedy (2019) in the binary-action non-Markov decision process setting. For $\delta = 1$ we get $\pi^{\mathrm{e}} = \pi^{\mathrm{b}}$; as δ shrinks, we tilt toward action 0; and, as δ grows, we tilt toward action 1. The parameter δ directly controls the amount of overlap; specifically $\pi_t^{\mathrm{e}}(a_t \mid s_t)/\pi_t^{\mathrm{b}}(a_t \mid s_t) \leq \max(\delta, \delta^{-1})$. For the general case in Definition 1, we have that $\pi_t^{\mathrm{e}}(a_t \mid s_t)/\pi_t^{\mathrm{b}}(a_t \mid s_t) \leq \max_{\tilde{a}_t} u_t(\tilde{a}_t)/\min_{\tilde{a}_t} u_t(\tilde{a}_t)$ so that the variation in u_t can directly control the overlap. Tilting policies ensure that $\pi_t^{\mathrm{e}}(\cdot \mid s_t)$ is absolutely continuous with respect to $\pi_t^{\mathrm{b}}(\cdot \mid s_t)$ so that the density ratio always exists. Thus, the overlap assumption is automatically satisfied. In contrast, if π_t^{e} is pre-specified and π_t^{b} is unknown, we cannot always ensure that the density ratio exists.

DEFINITION 2 (MODIFIED TREATMENT POLICY). *A modified treatment policy is specified by the maps $\tau_t : \mathcal{S}_t \times \mathcal{A}_t \to \mathcal{A}_t$ and assigns the action $\tau_t(s_t, \tilde{a}_t)$ in state s_t where $\tilde{a}_t$ is the natural action value distributed as $\tilde{a}_t \sim \pi_t^{\mathrm{b}}(\cdot \mid s_t)$.*

For example, if for each s_t, $\tau_t(s_t, \cdot)$ has a differentiable inverse $\tilde{\tau}_t(s_t, \cdot)$, then $\pi_t^{\mathrm{e}}(a_t \mid s_t) = \pi_t^{\mathrm{b}}(\tilde{\tau}_t(s_t, a_t) \mid s_t)\tilde{\tau}_t'(s_t, a_t)$, where $'$ denotes a differentiation with respect to a_t. The simplest example of a modified treatment policy is $\tau_t(s_t, a_t) = a_t + b_t(s_t)$ for some function $b_t(s_t)$, for which $\pi_t^{\mathrm{e}}(a_t|s_t) = \pi_t^{\mathrm{b}}(a_t - b_t(s_t) \mid s_t)$. The function $b_t(s_t)$ quantifies the deviation from the natural value. Keeping $b_t(s_t)$ small ensures implementability. While tilting policies are more descriptive, easily interpretable as an incremental policy change in a particular direction, modified treatment policies can be more prescriptive, corresponding to a direct change to current behavior.

3.3. Off-Policy Evaluation

Step-wise importance sampling (IS; Precup et al., 2000) and direct estimation of q-functions (DM; Ernst et al., 2005) are two common approaches for off-policy evaluation. However, the former is known to suffer from the high variance and the latter from model misspecification. To alleviate this, the doubly robust estimate combines the two (Jiang and Li, 2016; Murphy et al., 2001; Thomas and Brunskill, 2016). However, the asymptotic mean squared error of these can still grow exponentially in the horizon H. Kallus and Uehara (2020a) show that the efficiency bound in the time-varying Markov decision process case is actually polynomial in H and gives an estimator achieving it. When we additionally assume time invariance on time-varying Markov decision processes, Kallus and Uehara (2019a) show an orders-smaller efficiency bound and develop an efficient estimator leveraging time invariance.

All the above methods focus on the case where π^{e} is given explicitly. If the behavior policy is known, then natural stochastic policies can also be regarded as given explicitly and these still apply. When π^{b} is unknown, as in observational studies, we can still operationalize these methods for evaluating natural stochastic policies by first estimating π^{b} from the data, plugging this into π^{e} , and then treating π^{e} as specified by this estimate. However, this will fail to be efficient, as we discuss in Section 5. In fact, the efficiency bounds for evaluating natural stochastic policies are different than the pre-specified case.

Several variants of tilting policies have been widely employed in RL (Agarwal et al., 2021; Schulman et al., 2015). In these papers, the data is experimental but not observational. Thus, behavior and evaluation policies are known. As mentioned, the off-policy evaluation in these cases is covered by existing off-policy evaluation works. We focus on the case where the data is observational, and the behavior and evaluation policies are not known a priori, which is common in medical settings.

3.4. Curse of Horizon

In the non-Markov decision process model the trajectory distribution is $p_\pi(\mathcal{T}) = p_1(s_1)\prod_{t=1}^H \pi_t(a_t \mid \not{s}_{s_t}) p_t(r_t \mid \not{s}_{a_t}) p_{t+1}(s_{t+1} \mid \not{s}_{r_t})$, where $\not{s}_{a_t} = (s_1, a_1, r_1, \cdots, a_t)$, $\not{s}_{s_t} = (s_1, a_1, r_1, \cdots, s_t)$ and $\not{s}_{r_t} = (s_1, a_1, r_1, \cdots, r_t)$. This is in contrast to our time-varying Markov decision process model where the conditional densities can only depend on the most recent state and action. The non-Markov decision process setting is more standard in the classical literature of dynamic treatment regimes when the horizon H is short (Díaz et al., 2020; Kennedy, 2019; Robins et al., 1994; Zhang et al., 2013). Among these, Kennedy (2019) and Díaz et al. (2020) consider the evaluation of natural stochastic policies.

We can formally show that efficiency bounds for off-policy evaluation under the non-Markov decision process model grows exponentially in the horizon, which is often referred to as the curse of horizon in reinforcement learning. In our work, by leveraging the Markovian assumption in time-varying Markov decision processes, we aim to show how we can circumvent the curse of horizon. For completeness and comparison, in Appendix B, we derive efficiency bounds under non-Markov decision process and compare them to Kennedy (2019) and the concurrent work of Díaz et al. (2020).

4. EFFICIENCY BOUNDS

4.1. Tilting Policies

We calculate the efficient influence function and efficiency bounds for evaluating natural stochastic policies in reinforcement learning with respect to nonparametric models $\mathcal{M}_{\mathrm{TMDP}}$, which are induced by time-varying Markov decision process distributions. After deriving them for off-policy evaluation of tilting policies, we turn to the case of modified treatment policies. The efficiency bound is the smallest-possible error we can hope to achieve in estimating J. This is the gold standard to measure the optimality of estimators in causal inference. See van der Laan and Robins, 2003; van der Vaart, 1998 for more details. Crucially, any efficient estimator is asymptotically linear with influence function equal to the efficient influence function. Thus, efficient influence functions also play important role in constructing efficient estimators.

The following theorem reveals the efficient influence function and efficiency bound for tilting policies.

THEOREM 1. *Let π^{e} be as in Definition 1. Then the efficient influence function and efficiency bound of J with respect to the model $\mathcal{M}_{\mathrm{TMDP}}$ are, respectively,*

$$-J + \textstyle\sum_{t=1}^H \{\mu_t(r_t - v_t) + \mu_{t-1} v_t\}, \quad \Upsilon_{\mathrm{TI1}} = \sum_{t=0}^H \mathrm{E}[\mathrm{var}\{\mu_t(r_t + v_{t+1}) \mid s_t\}],$$

where $\mu_0 = 1, v_0 = r_0 = 0$. Moreover, Υ_{TI1} is upper bounded by $CC' R_{\max}^2 H^2$.

The function u_t that specifies the tilting policy is implicit in the variables μ_t, v_t above, which depend on π_t^{e}. While the efficiency bound is larger than in the case of a pre-specified evaluation policy (Kallus and Uehara, 2020a), the overall order, $CC' R_{\max}^2 H^2$, is the same. This indicates we can possibly circumvent the curse of horizon by using an efficient estimator.

Remark 1 *(Comparison to pre-specified evaluation policy).* In the pre-specified evaluation policy case, Kallus and Uehara (2020a) show that the efficient influence function and efficiency

bound are, respectively,

$$-J+\sum_{t=1}^{H}\left\{\mu_t(r_t-q_t)+\mu_{t-1}v_t\right\},\quad \sum_{t=0}^{H}\mathrm{E}\left\{\mu_t^2\mathrm{var}\left(r_t+v_{t+1}\mid s_t,a_t\right)\right\}. \tag{4}$$

Specifically, this is also the efficiency bound for evaluating a tilting policy when π^{b} is known and π^{e} is also known. To derive our result, compared to Kallus and Uehara (2020a) who deal with pre-specified policies, we additionally calculate the derivative with respect to behavior policies. Recall our estimand also depends on behavior policies. Since this additional term is $\sum_t \mu_t(q_t-v_t)$, by adding it to the efficient influence function in (4), the efficient influence function in Theorem 1 is derived. Compared with the efficiency bound in (4), Υ_{TI1} is larger by $\sum_{t=1}^{H}\mathrm{E}\left\{w_t^2\mathrm{var}\left(\eta_t q_t\mid s_t\right)\right\}$. Hence, the more the q-function $q_t(s_t,a_t)$ varies over actions a_t at any one state, the greater the difference.

Remark 2 *(Non-dynamic case).* When $H=1$ and the evaluation policy is pre-specified, the efficient influence function is given by the familiar doubly robust influence function $\eta_1\left\{r_1-q_1(s_1,a_1)\right\}+v_1(s_1)-J$ (Dudik et al., 2014; Robins et al., 1999). In contrast, for $H=1$, the efficient influence function in Theorem 1 for evaluating tilting policies is instead $\eta_1\left\{r_1-v_1(s_1)\right\}+v_1(s_1)-J$, where we used $\mu_1=\eta_1$. The difference to the familiar doubly robust influence function is $\eta_1(q_1(s_1,a_1)-v_1(s_1))$, which is exactly the derivative of J with respect to π_b, that is, the component of the efficient influence function accounting for the uncertainty in π_b (which is 0 in usual off-policy evaluation where policy value has no dependence on π_b). Intuitively, we can understand it as an optimal control variate for $v_1(s_1)$.

4.2. *Modified Treatment Policies*

We next handle the case of modified treatment policies. Again, we will see that we can potentially circumvent the curse of horizon with an efficient estimator.

THEOREM 2. *Let π^{e} be as in Definition 2. Then the efficient influence function and efficiency bound of J with respect to the model $\mathcal{M}_{\mathrm{TMDP}}$ are, respectively,*

$$-J+\sum_{t=1}^{H}\left\{\mu_t(r_t-q_t)+\mu_{t-1}q_t^{\tau}\right\},\quad \Upsilon_{\mathrm{MO1}}=\sum_{t=0}^{H}\mathrm{E}\left\{\mu_t^2\mathrm{var}\left(r_t+q_{t+1}^{\tau}\mid s_t,a_t\right)\right\} \tag{5}$$

where $q_t^{\tau}(s_t,a_t)=q_t(s_t,\tau_t(s_t,a_t))$. Moreover, Υ_{MO1} is upper bounded by $CC'R_{\max}^2H^2$.

Remark 3 *(Comparison to pre-specified evaluation policy).* To derive our result, compared to Kallus and Uehara (2020a) who deal with pre-specified policies, we additionally calculate the derivative with respect to behavior policies. This additional derivative is $\sum_t(-v_t(s_t)+q_t^{\tau}(s_t,a_t))$. Hence, by adding it to the efficient influence function in Theorem 1, the efficient influence function in Theorem 2 is derived. Compared with the efficiency bound for a pre-specified evaluation policy, Υ_{MO1} is larger by $\sum_{t=0}^{H}\mathrm{E}\left\{\mu_t^2\mathrm{var}\left(q_{t+1}^{\tau}\mid s_{t+1}\right)\right\}$. Hence, the more the q-function $q_t^{\tau}(s_t,a_t)$ varies over actions a_t at any one state, the greater the difference.

Remark 4 *(Non-dynamic case).* When $H=1$, the efficient influence function is $\eta_1\left\{r_1-q_1(s_1,a_1)\right\}+q_1^{\tau}(s_1,a_1)-J$. This matches the results in Díaz and van der Laan (2013, 2018), where we used $\mu_1=\eta_1$. The difference to the familiar doubly robust influence function is $q_1^{\tau}(s_1,a_1)-v_1(s_1)$, which is exactly the derivative of J with respect to π_b. Intuitively, we can understand $q_1^{\tau}(s_1,a_1)$ as an unbiased surrogate for $v_1(s_1)=\mathrm{E}_{a_1\sim\pi_1^b(\cdot|s_1)}[q_1(s_1,\tau(s_1,a_1))\mid s_1]$, which does not require that we know π^b, but, at the same, has higher variance.

5. EFFICIENT AND (PARTIALLY) DOUBLY ROBUST ESTIMATION

5.1. *Tilting Policies*

Staring with tilting policies, we next propose efficient estimators for natural stochastic policies based on the obtained efficient influence functions. Recalling a general theorem in semiparametric theory that efficient estimators are asymptotically linear with efficient influence functions,

Algorithm 1 Efficient Off-Policy Evaluation for Natural Stochastic Policies

Take a K-fold random partition of $\{1,\dots,n\} = I_1 \cup \cdots \cup I_K$ such that the size of each fold, $|I_k|$, is within 1 of n/K; set $\mathcal{U}_k = \{\mathcal{T}^{(i)} : i \in I_k\}$, $\mathcal{L}_k = \{\mathcal{T}^{(i)} : i \notin I_k\}$

for $k \in \{1, \cdots, K\}$ **do**

Using only $\mathcal{L}_k$ as data, construct nuisance estimators $\hat{w}_t^{(k)}$, $\hat{\pi}_t^{b,(k)}, \hat{q}_t^{(k)}$ for $t \le H$

Set $\hat{\pi}_t^{e,(k)}(a_t \mid s_t) = u_t(a_t)\hat{\pi}_t^{b,(k)}(a_t \mid s_t) / \int u_t(\tilde{a}_t)\hat{\pi}_t^{b,(k)}(\tilde{a}_t \mid s_t)\mathrm{d}\tilde{a}_t$,

$\hat{\eta}_t^{(k)}(s_t, a_t) = \hat{\pi}_t^{e,(k)}(a_t \mid s_t)/\hat{\pi}_t^{b,(k)}(a_t \mid s_t)$, $\hat{v}_t^{(k)}(s_t) = \int \hat{q}_t^{(k)}(s_t, a_t)\hat{\pi}_t^{e,(k)}(a_t \mid s_t)\mathrm{d}a_t$

Set $\hat{J}_k = \frac{1}{|I_k|}\sum_{\mathcal{T}\in\mathcal{U}_k} \hat{\phi}^{(k)}(\mathcal{T})$, where

$$\hat{\phi}^{(k)}(\mathcal{T}) = \sum_{t=1}^{H} \hat{w}_t^{(k)}(s_t)\hat{\eta}_t^{(k)}(s_t, a_t)\left\{r_t - \hat{v}_t^{(k)}(s_t)\right\} + \hat{w}_{t-1}^{(k)}(s_{t-1})\hat{\eta}_{t-1}^{(k)}(s_{t-1}, a_{t-1})\hat{v}_t^{(k)}(s_t) \quad (6)$$

end for

Return $\hat{J}_{\mathrm{TI1}} = \frac{1}{n}\sum_{k=1}^{K} |I_k|\hat{J}_k$

one natural way to obtain efficient estimators is to take empirical averages of approximations of efficient influence functions.

We propose an estimator $\hat{J}_{\mathrm{TI1}}$ for tilting policies in Algorithm 1. This is a meta-algorithm given estimators for the nuisances w_t, π_t^b, q_t, which we discuss how to estimate in Section 5.3. This estimator is derived by taking an empirical average of the approximation of the efficient influence function given by plugging in estimators for the unknown nuisance functions.

We next prove $\hat{J}_{\mathrm{TI1}}$ is efficient under nonparametric rate conditions on nuisance estimators, which crucially can be slower than $\mathcal{O}_p(n^{-1/2})$. The use of cross-fitting allows us to avoid metric entropy conditions (Chernozhukov et al., 2018; Zheng and van Der Laan, 2011). In Appendix D, we provide a parallel result for the algorithm without cross-fitting but imposing additional conditions on nuisance estimates.

THEOREM 3 (EFFICIENCY). *Suppose* $\forall k \le K, \forall j \le H$, $\|\hat{\pi}_j^{b,(k)}(a_j \mid s_j) - \pi_j^b(a_j \mid s_j)\|_2 \le \alpha_1$, $\|\hat{w}_j^{(k)}(s_j) - w_j(s_j)\|_2 \le \alpha_2$, $\|\hat{q}_j^{(k)}(s_j, a_j) - q_j(s_j, a_j)\|_2 \le \beta$, *where* $\alpha_2 = \mathcal{O}_p(n^{-1/4})$, $\alpha_1 = \mathrm{o}_p(n^{-1/4})$, $\beta = \mathcal{O}_p(n^{-1/4})$, $\alpha_2\beta = \mathrm{o}_p(n^{-1/2})$. *Then,* $\sqrt{n}(\hat{J}_{\mathrm{TI1}} - J) \xrightarrow{d} \mathcal{N}(0, \Upsilon_{\mathrm{TI1}})$.

The result essentially follows by showing that $|\hat{J}_{\mathrm{TI1}} - J - \mathbb{P}_n\{\phi(\mathcal{T})\}| \le \alpha_1\alpha_2 + \alpha_1\beta + \alpha_2\beta + \alpha_1^2 + o_p(n^{-1/2})$, where $\phi(\mathcal{T})$ is the efficient influence function. Under the above rate assumptions, the right-hand side is $\mathrm{o}_p(n^{-1/2})$ and the result is concluded from central limit theorem. Importantly, in this situation, the error is upper-bounded by terms consisting of products of two (possibly the same) L2 errors of nuisance estimators. If the right-hand side did not consist of product terms, we cannot allow for nonparametric rates of nuisance estimators to ensure $\sqrt{n}$-consistency and efficiency.

Notice that if we knew the behavior policy so that $\alpha_1 = 0$, this becomes simply $\alpha_2\beta + o_p(n^{-1/2})$ and we recover the doubly robust structure of the pre-specified case (Kallus and Uehara, 2020a): the estimator is consistent if either w_t or q_t is consistently estimated. In this case, the error is upper-bounded by terms consisting of products of two different L2 errors of nuisance estimators. Unlike this case, in our general setting with $\alpha_1 \neq 0$, because of the term α_1^2, the consistent estimation of π^{b} is required to estimate J consistently. So, we have a partial double robustness in the sense that the estimator is consistent as long as π^{b} and either w or q are consistently estimated.

THEOREM 4 (PARTIAL DOUBLE ROBUSTNESS). *Suppose* $\forall k \leq K, \forall j \leq H$, *for some* $w_j^\dagger, q_j^\dagger$, $\|\hat{\pi}_j^{b,(k)}(a_j|s_j) - \pi_j^b(a_j|s_j)\|_2 = \mathrm{o}_p(1)$, *and* $\|\hat{q}_j^{(k)}(s_j, a_j) - q_j^\dagger(s_j, a_j)\|_2 = \mathrm{o}_p(1)$, *and* $\|\hat{w}_j^{(k)}(s_j) - w_j^\dagger(s_j)\|_2 = \mathrm{o}_p(1)$. *As long as either* $q_j^\dagger = q_j$ *or* $w_j^\dagger = w_j$, *we have* $\hat{J}_{\mathrm{TI1}} \xrightarrow{p} J$.

Unlike standard double robustness as described in Hahn (1998); Lawless et al. (1999); Robins et al. (1994), the partial double robustness above arises because behavior policies are no longer ancillary for the functional of interest. Due to similar reasons, partial double robustness properties also appear in other contexts, such as certain policy evaluation problems with instrumental variables (Tchetgen Tchetgen and Vansteelandt, 2013) and dynamic discrete choice models (Chernozhukov et al., 2019).

Remark 5 *(Why we cannot simply rely on the efficient influence function for pre-specified policies).* Since we have to estimate π^{b} and π^{e} consistently for our estimator to work, a careful reader might wonder whether we might as well plug in the estimated π^{e} implied from this estimation into estimators that are efficient for the pre-specified case such as that of Kallus and Uehara (2020a). But, this must necessarily fail. Consider the case of $H = 1$ for simplicity and replace Eq. (1) in Algorithm 1 with $\phi^{(k)}(\mathcal{T}) = \phi^{\mathrm{pre}}(s_1, a_1, r_1; \hat{\pi}_1^{b,(k)}, \hat{q}_1^{(k)}, \hat{\pi}_1^{e,(k)})$ where $\phi^{\mathrm{pre}}(s_1, a_1, r_1; \tilde{\pi}_1^b, \tilde{q}_1, \tilde{\pi}_1^e) = \frac{\tilde{\pi}_1^e(s_1,a_1)}{\tilde{\pi}_1^b(s_1,a_1)}(r_1 - \tilde{q}_1(s_1, a_1)) + \int \tilde{q}_1(s_1, \tilde{a}_1)\tilde{\pi}_1^e(\tilde{a}_1 \mid s_1)\mathrm{d}\tilde{a}_1$. That is, plug in our nuisance estimates into the familiar doubly robust formula. With a specified evaluation policy, guarantees for this estimator (and its dynamic version in Kallus and Uehara, 2020a) leverage that the value of any evaluation policy $\tilde{\pi}_1^e$ can be written $J(\tilde{\pi}_1^e) = \mathrm{E}\phi^{\mathrm{pre}}(s_1, a_1, r_1; \pi_1^b, \tilde{q}_1, \tilde{\pi}_1^e) = \mathrm{E}\phi^{\mathrm{pre}}(s_1, a_1, r_1; \tilde{\pi}_1^b, q_1, \tilde{\pi}_1^e)$. That is, this formulation has zero derivative in perturbations to π_1^b, q_1, with $\tilde{\pi}_1^e$ fixed, so estimation errors in these translate to negligible downstream errors. However, $J(\tilde{\pi}_1^e)$ certainly does not have zero derivative in $\tilde{\pi}_1^e$ at π_1^e. Therefore, even small errors in π_1^e must directly propagate to the off-policy evaluation estimate, inflating variance, possibly introducing $\sqrt{n}$-order bias, and imperiling rates if we use nonparametric estimates. In contrast, in Theorem 3, we leveraged an intermediate result that the special structure of our estimator affords it evaluation errors that are quadratic in error for π_t^b, so small errors therein will become negligible.

Remark 6 *(Estimation of v-functions).* Although $\hat{q}_t^{(k)}$ does not explicitly appear in Eq. (1), we do need to estimate $\hat{q}_t^{(k)}$ first and then compute $\hat{v}_t^{(k)}$ based on this estimate as done in Algorithm 1, instead of directly estimating v_t, to ensure partial double robustness with respect to $w_t, \pi_t^{\mathrm{b}}, q_t$. That is, all nuisance estimates have to use the same estimate of π_t^b in order for errors to multiply.

5.2. *Modified Treatment Policies*

We similarly define the estimator $\hat{J}_{\mathrm{MO1}}$ for the case of modified treatment policies by taking Algorithm 1 and (a) replacing $\hat{\pi}_t^{e,(k)}(a_t \mid s_t)$ by $\hat{\pi}_t^{e,(k)}(a_t \mid s_t) = \hat{\pi}_t^{b,(k)}(\tilde{\tau}_t(s_t, a_t) \mid s_t)\tilde{\tau}_t'(s_t, a_t)$ and (b) replacing Eq. (1) by

$$\hat{\phi}^{(k)}(\mathcal{T}) = \textstyle\sum_{t=1}^H \hat{w}_t^{(k)}(s_t)\hat{\eta}_t^{(k)}(s_t, a_t)\left\{r_t - \hat{q}_t^{(k)}(s_t, a_t)\right\} + \hat{w}_{t-1}^{(k)}(s_{t-1})\hat{\eta}_{t-1}^{(k)}(s_{t-1}, a_{t-1})\hat{q}_t^{(k)}(s_t, \tau_t(s_t, a_t)).$$

We then have the following efficiency and full double robustness results.

THEOREM 5 (EFFICIENCY). *Suppose* $\forall k \leq K, \forall j \leq H$, $\|\hat{\pi}_j^{b,(k)}(a_j|s_j) - \pi_j^{b,(k)}(a_j|s_j)\|_2 \leq \alpha_1$, $\|\hat{w}_j^{(k)}(s_j) - w_j(s_j)\|_2 \leq \alpha_2$, $\|\hat{q}_j^{(k)}(s_j, a_j) - q_j(s_j, a_j)\|_2 \leq \beta$ *where* $(\alpha_2 + \alpha_1)\beta = \mathrm{o}_p(n^{-1/2})$, $\max\{\alpha_2, \alpha_1, \beta\} = \mathrm{o}_p(1)$. *Then,* $\sqrt{n}(\hat{J}_{\mathrm{MO1}} - J) \xrightarrow{d} \mathcal{N}(0, \Upsilon_{\mathrm{MO1}})$.

THEOREM 6 (DOUBLE ROBUSTNESS). *Assume* $\forall k \leq K, \forall j \leq H$, *for some* $\pi_j^{b\dagger}, q_j^\dagger, w_j^\dagger$, $\|\hat{\pi}_j^b(a_j|s_j) - \pi_j^{b\dagger}(a_j|s_j)\|_2 = \mathrm{o}_p(1)$, $\|\hat{q}_j^{(k)}(s_j, a_j) - q_j^\dagger(s_j, a_j)\|_2 = \mathrm{o}_p(1)$, $\|w_j^{(k)}(s_j) - w_j^\dagger(s_j)\|_2 = \mathrm{o}_p(1)$. *Then as long as either* $q_j^\dagger = q_j$ *or* $\pi_j^{b\dagger} = \pi_j^b, w_j^\dagger = w_j$, *we have* $\hat{J}_{\mathrm{MO1}} \xrightarrow{p} J$.

These theorems arise from the bias structure $|\hat{J}_{\mathrm{MO1}} - J - \mathbb{P}_n\{\phi(\mathcal{T})\}| \leq (\alpha_2 + \alpha_1)\beta + o_p(n^{-1/2})$. Hence, the error is upper-bounded by products of L2 errors for nuisance estimators. This ensures efficiency even if nuisances converge at nonparametric rates slower than $\mathcal{O}_p(n^{-1/2})$.

The conditions on nuisance estimates in these theorems are weaker than the ones for tilting policies. Comparing Theorem 3 and 5, the condition in Theorem 5 is satisfied even if some of $\alpha_2, \alpha_1, \beta$ are slower than $\mathrm{o}_p(n^{-1/4})$. Comparing Theorems 4 and 6, the condition in Theorem 6 can be satisfied even if the behavior policy model is misspecified. The intuitive reason is that for a modified treatment policy, J can be specified in a form not depending on π^{b}, while this is not true for tilting policies. We again emphasize that plugging in an estimate of π^{e} into the method of Kallus and Uehara, 2020a can fail to be efficient and even $\sqrt{n}$-consistent. In particular, where Kallus and Uehara (2020a) uses $\hat{v}_t^{(k)}(s_t)$, our proposal instead uses $\hat{q}_t^{(k)}(s_t, \tau_t(s_t, a_t))$.

5.3. *Nuisance Estimation*

Our estimators for both types of stochastic policies require that we estimate $\pi_t^{\mathrm{b}}, w_t, q_t$, possibly at some slow rate. Here we discuss some standard ways to estimate these nuisance functions.

First, estimating π_t^{b} amounts to conditional density estimation. Once we fit π_t^{b}, we also immediately have an estimate of π_t^{e}. We can then use standard methods for estimating w_t and q_t that assume π_t^{e} is given by plugging in our estimate for it as follows. Generally speaking, if the estimate for q_t or w_t would have had some convergence rate r_n if π_t^{e} were given exactly, then this rate does not deteriorate as long as the plugged-in estimate for π_t^{e} also has rate at least r_n.

Next, we discuss the estimation of q-functions. In the tabular case (*i.e.*, finite state and action spaces), a model-based approach is the most common way to estimate q-functions from off-policy data, wherein we directly estimate the transition and reward distribution and then compute q-functions from these. In the non-tabular case, we have to rely on some function approximation. In particular, we can rely on the Bellman equation: $q_t(s_t, a_t) = \mathrm{E}\left\{r_t + q_{t+1}(s_{t+1}, \pi^{\mathrm{e}}) \mid s_t, a_t\right\}$, where $q_t(s_t, \pi) = \int q_t(s_t, a_t)\pi(a_t \mid s_t)\mathrm{d}a_t$. One of the most common ways to operationalize this is using fitted q-iteration (Antos et al., 2008; Duan and Wang, 2020; expressed here using an estimated evaluation policy, $\hat{\pi}^e$): set $\hat{q}_{H+1} \equiv 0$, and for $t = H, \ldots, 1$ estimate $\hat{q}_t$ by regressing $r_t + \hat{q}_{t+1}(s_{t+1}, \hat{\pi}^e)$ onto s_t, a_t using any given (possibly nonparametric) regression method.

Finally, we discuss estimating w_t. In tabular cases, we can use a model-based approach (Yin and Wang, 2020), computing the density ratios directly from the estimated transition densities. In non-tabular cases, we must rely on function approximation. Here, we can use the relation $w_t = \mathrm{E}(\eta_{t-1}w_{t-1} \mid s_t)$. Then, we can use an iterative procedure: for each $t = 2, \ldots,$ (1) estimate π_{t-1}^b using any flexible regression method and use it to construct an estimate $\hat{\eta}_{t-1}$, (2) estimate w_t by $\hat{w}_t$ by regressing $\hat{\eta}_{t-1}\hat{w}_{t-1}$ on s_t using any flexible regression method, with $\hat{w}_1 = 1$.

6. EXTENSION TO TIME-HOMOGENEOUS MARKOV DECISION PROCESSES

6.1. *Setting*

We next extend our results to time-homogeneous time-varying Markov decision processes where transitions, rewards, and policies do not depend on t, *i.e.*, $p_t(r|s,a) = p(r|s,a), p_t(s'|s,a) = p(s'|s,a), \pi_t^{\mathrm{b}} = \pi^{\mathrm{b}}, \tau_t = \tau$. Here, the estimand we consider is an average discounted reward, $J(\gamma) = (1-\gamma)\lim_{T\to\infty} \mathrm{E}_{\pi^{\mathrm{e}}}(\sum_{t=1}^T \gamma^{t-1} r_t)$ when the initial state distribution is $p_e^{(1)}(s)$. Although we can still apply methods developed for the time-varying Markov deci-

sion process in the Markov decision process case, if we correctly leverage the time-homogeneity of Markov decision process, we should do much better in that the asymptotic rate of mean squared error should be $\mathcal{O}(1/NH)$, not $\mathcal{O}(1/N)$, when we observe N trajectories of length H. Especially, this difference is significant when H is much larger than N as seen in modern mobile health applications. In this section, we derive the efficient influence function and efficiency bound of $J(\gamma)$ with respect to the Markov decision process model denoted by $\mathscr{M}_{\mathrm{MDP}}$. Then, we compare to the efficiency bound and estimators for pre-specified evaluation policies.

Following the standard offline reinforcement learning setting (Agarwal et al., 2021), we consider the observed data to be n i.i.d. draws from the following offline distribution: for $i = 1, \ldots, n$,

$$(s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)}, a'^{(i)}) \sim p_b(s, a, r, s', a') = p_b(s)\pi^{\mathrm{b}}(a \mid s)p(s'|s, a)p(r|s, a)\pi^{\mathrm{b}}(a' \mid s').$$

Given N trajectories of length H, we can convert it into the above $n = NH$ transition tuples. We note that our theory holds for any $p_b(s)$. In practice, $p_b(s)$ is often taken to be the stationary distribution associated with the policy π^{b}. Taking $p_b(s)$ as such, we can handle the case where the data is not independent and identically distributed but instead comes from observing $N = n/H$ trajectories of length H, if we impose some additional mixing assumptions and let $H \to \infty$, following the approach in Kallus and Uehara (2019a).

In this section, we consider a fully nonparametric model $\mathscr{M}_{\mathrm{MDP}}$ in that we make no restrictions on the above distributions. We define $q(s, a) = \mathrm{E}_{p_{\pi^{\mathrm{e}}}}\left(\sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_1 = s, a_1 = a\right)$, $q^{\tau}(s, a) = q(s, \tau(s, a))$, $v(s) = \mathrm{E}_{\pi^{\mathrm{e}}(a|s)}\left\{q(s, a)|s\right\}$, $w^*(s) = p^{(\infty)}_{e,\gamma}(s)/p_b(s)$, and $\mu^*(s, a) = w^*(s)\eta(s, a)$.

Per Kallus and Uehara (2019a); Liu et al. (2018), we can rewrite $J(\gamma)$ as

$$J(\gamma) = \mathrm{E}_{s\sim p_b(s), a\sim\pi^{\mathrm{b}}(a|s), r\sim p(r|s,a)}[r\; p^{(\infty)}_{e,\gamma}(s)\pi^{\mathrm{e}}(a \mid s)/\{p_b(s)\pi^{\mathrm{b}}(a \mid s)\}], \tag{7}$$

where $p^{(\infty)}_{e,\gamma}(s)$ is the γ-discounted average state visitation distribution associated with the Markov decision process, policy π^{e}, and the initial state distribution $p^{(1)}_e(s)$.

6.2. Tilting Policies

THEOREM 7. *Let π^{e} be as in Definition 1. The efficient influence function and efficiency bound of $J(\gamma)$ with respect to $\mathscr{M}_{\mathrm{MDP}}$ are, respectively,*

$$\mu^*(s, a)\left\{r + \gamma v(s') - v(s)\right\}, \quad \Upsilon_{\mathrm{TI2}} = \mathrm{E}\left(\mathrm{var}\left[\mu^*(s, a)\left\{r + \gamma v(s')\right\} \mid s\right]\right).$$

Again this is different from the pre-specified-policy case (Kallus and Uehara, 2019a). We discuss the differences in detail in Section 6.4.

We can construct an efficient estimator by following a similar but slightly different cross-fitting strategy as before. With additional data $s^{(j)}_1 \sim p^{(1)}_e(s)$, $j = 1, \ldots, m$ where $m = \Omega(n)$ (if $p^{(1)}_e$ is known), and given nuisance estimators $\hat{\pi}^{b,(k)}, \hat{q}^{(k)}, \hat{w}^{*(k)}$, we propose the estimator $\hat{J}_{\mathrm{TI2}}$ for $J(\gamma)$ by taking Algorithm 1 and replacing $\hat{J}_k$ with

$$\begin{aligned}
\hat{J}_k &= \tfrac{1-\gamma}{m}\textstyle\sum_{j=1}^{m} \hat{v}^{(k)}(s^{(j)}_1) \qquad (8)\\
&\quad + \tfrac{1}{|I_k|}\textstyle\sum_{i\in I_k} \hat{w}^{*(k)}(s^{(i)})\hat{\eta}^{(k)}(s^{(i)}, a^{(i)})\{r^{(i)} + \gamma\hat{v}^{(k)}(s'^{(i)}) - \hat{v}^{(k)}(s^{(i)})\},\\
\hat{\pi}^{e,(k)}(a \mid s) &= u(a)\hat{\pi}^{b,(k)}(a \mid s)\big/ \textstyle\int u(\tilde{a})\hat{\pi}^{b,(k)}(\tilde{a} \mid s)\mathrm{d}\tilde{a},\\
\hat{\eta}^{(k)}(s, a) &= \hat{\pi}^{e,(k)}(a \mid s)/\hat{\pi}^{b,(k)}(a \mid s),\ \hat{v}^{(k)}(s) = \textstyle\int \hat{q}^{(k)}(s, a)\hat{\pi}^{e,(k)}(a \mid s)\mathrm{d}a. \qquad (9)
\end{aligned}$$

To estimate π^{b}, we can follow Section 5.3. To estimate w^*, q, we can solve the following set of moment conditions given test functions $\mathscr{F}, \mathscr{G}$ (cf. Kallus and Uehara, 2019a; Liu et al., 2018):

$$0 = (1-\gamma)\mathrm{E}_{s_1\sim p_e^{(1)}}\{f(s_1)\} + \mathrm{E}_{(s,a,s')\sim p_b}\left[\gamma w^*(s)\left\{\eta(s,a)f(s') - f(s)\right\}\right] \quad \forall f \in \mathscr{F}, \quad (10)$$

$$0 = \mathrm{E}_{(s,a,r,s')\sim p_b}\left[g(s,a)\left\{r + \gamma v(s') - q(s,a)\right\}\right] \quad \forall g \in \mathscr{G}. \quad (11)$$

Then, by leveraging the idea of minimax estimation, with function classes $\mathscr{Q} \subset [\mathcal{S}\times\mathcal{A}\to\mathbb{R}]$, $\mathscr{G}\subset[\mathcal{S}\times\mathcal{A}\to\mathbb{R}]$, we can estimate $q(\cdot)$ by

$$\hat{q} = \arg\min_{q\in\mathscr{Q}} \max_{g\in\mathscr{G}} \tfrac{1}{n}\textstyle\sum_{i=1}^n \left[g(s^{(i)},a^{(i)})\left\{r^{(i)} + \gamma q(s'^{(i)},\hat{\pi}^e) - q(s^{(i)},a^{(i)})\right\}\right].$$

The estimation of $w^\star(s)$ is similarly performed. We next prove $\hat{J}_{\mathrm{TI2}}$ is efficient and partially doubly robust, mirroring the time-varying Markov decision process case.

THEOREM 8 (EFFICIENCY). *Suppose* $\forall k \le K$, $\|\hat{\pi}^b(a|s) - \pi^b(a|s)\|_2 \le \alpha_1$, $\|\hat{w}^{*(k)}(s) - w^*(s)\|_2 \le \alpha_2$, $\|\hat{q}^{(k)}(s,a) - q(s,a)\|_2 \le \beta$, *where* $\alpha_2 = \mathcal{O}_p(n^{-1/4})$, $\alpha_1 = \mathrm{o}_p(n^{-1/4})$, $\beta = \mathcal{O}_p(n^{-1/4})$, $\alpha_2\beta = \mathrm{o}_p(n^{-1/2})$. *Then,* $\sqrt{n}(\hat{J}_{\mathrm{TI2}} - J) \xrightarrow{d} \mathcal{N}(0, \Upsilon_{\mathrm{TI2}})$.

THEOREM 9 (PARTIAL DOUBLE ROBUSTNESS). *Assume* $\forall k \le K$, *for some* $w^{*\dagger}(s), q^\dagger(s,a)$, $\|\hat{w}^{*(k)}(s) - w^{*\dagger}(s)\|_2 = \mathrm{o}_p(1), \|\hat{\pi}^b(a|s) - \pi^b(a|s)\|_2 = \mathrm{o}_p(1), \|\hat{q}^{(k)}(s,a) - q^\dagger(s,a)\|_2 = \mathrm{o}_p(1)$. *Then, as long as* $w^{*\dagger}(s) = w^*(s)$ *or* $q^\dagger(s,a) = q(s,a)$, $\hat{J}_{\mathrm{TI2}} \xrightarrow{p} J$.

The result again essentially follows by showing that $|\hat{J}_{\mathrm{TI1}} - J - \mathbb{P}_n\{\phi(s,a,r,s')\}| \le \alpha_2\alpha_1 + \alpha_2\beta + \alpha_1\beta + \alpha_1^2 + o_p(n^{-1/2})$, where $\phi(s,a,r,s')$ is the efficient influence function. Under the above rate assumptions, the right-hand side is $\mathrm{o}_p(n^{-1/2})$ and the result is concluded from central limit theorem. Similar to the time-varying Markov decision process case, the estimator can be efficient even if α_2, β are slower than $\mathrm{o}_p(n^{-1/4})$, but α_1 must be $\mathrm{o}_p(n^{-1/4})$. This mirrors the partial double robustness in Section 5.

6.3. Modified Treatment Policies

THEOREM 10. *Let* π^{e} *be as in Definition 2. The efficient influence function and efficiency bound of* $J(\gamma)$ *with respect to* $\mathscr{M}_{\mathrm{MDP}}$ *are, respectively,*

$$\mu^*(s,a)\left\{r + \gamma q^\tau(s',a') - q(s,a)\right\}, \quad \Upsilon_{\mathrm{MO2}} = \mathrm{E}\left[\mu^{*2}(s,a)\mathrm{var}\left\{r + \gamma q^\tau(s',a') \mid s,a\right\}\right].$$

The derived efficient influence function differs significantly from the pre-specified case in Kallus and Uehara (2019a), which in particular does not at all involve a'. We compare the efficiency bound in Section 6.4.

We construct an efficient estimator as follows. With additional data $(s^{(j)},a^{(j)}) \sim p_e^{(1)}(s)\pi^{\mathrm{b}}(a \mid s), j = 1,\dots,m$ where $m = \Omega(n)$, and nuisance estimators $\hat{w}^{*(k)}, \hat{\pi}^{b,(k)}, \hat{q}^{(k)}$, we propose the estimator $\hat{J}_{\mathrm{MO2}}$ by taking Algorithm 1 and replacing $\hat{J}_k$ with

$$\hat{J}_k = \tfrac{1}{|I_k|}\textstyle\sum_{i\in I_k} \hat{w}^{*(k)}(s)\hat{\eta}^{(k)}(s^{(i)},a^{(i)})\left\{r^{(i)} + \gamma\hat{q}^{(k)\tau}(s'^{(i)},a'^{(i)}) - \hat{q}^{(k)}(s^{(i)},a^{(i)})\right\}$$
$$+ \tfrac{(1-\gamma)}{m}\textstyle\sum_{j=1}^m \hat{q}^{(k)\tau}(s^{(j)},a^{(j)}), \quad \hat{\pi}^{e,(k)}(a_t \mid s_t) = \hat{\pi}^{b,(k)}(\tilde{\tau}(s_t,a_t) \mid s_t)\tilde{\tau}'(s_t,a_t),$$

and $\hat{\eta}^{(k)}$, $\hat{v}^{(k)}$ are as in Eq. (9). To estimate w^* we can use Eq. (10) and to estimate q we can use:

$$0 = \mathrm{E}_{(s,a,r,s',a')\sim p_b}[g(s,a)\left\{r - q(s,a) + \gamma q^\tau(s',a')\right\}] \quad \forall g \in \mathscr{G}. \quad (12)$$

We next prove $\hat{J}_{\mathrm{MO2}}$ is efficient and doubly robust, mirroring the time-varying Markov decision process case.

THEOREM 11 (EFFICIENCY). *Assume* $\forall k \leq K$, $\|\hat{w}^{*(k)}(s) - w^*(s)\|_2 \leq \alpha_2, \|\hat{\pi}^{b,(k)}(a|s) - \pi^b(a|s)\|_2 \leq \alpha_1, \|\hat{q}^{(k)}(s,a) - q(s,a)\|_2 \leq \beta$, *where* $(\alpha_2 + \alpha_1)\beta = \mathrm{o}_p(n^{-1/4}), \max\{\alpha_2, \alpha_1, \beta\} = \mathrm{o}_p(1)$. *Then,* $\sqrt{n}(\hat{J}_{\mathrm{MO2}} - J) \xrightarrow{d} \mathcal{N}(0, \Upsilon_{\mathrm{MO2}})$.

THEOREM 12 (DOUBLE ROBUSTNESS). *Assume* $\forall k \leq K$, *for some* $w^{*\dagger}, \pi^{b\dagger}, q^\dagger$, $\|\hat{w}^{*(k)}(s) - w^{*\dagger}(s)\|_2 = \mathrm{o}_p(1), \|\hat{\pi}^{b,(k)}(a|s) - \pi^{b\dagger}(a|s)\|_2 = \mathrm{o}_p(1), \|\hat{q}^{(k)}(s,a) - q^\dagger(s,a)\|_2 = \mathrm{o}_p(1)$. *Then, as long as* $w^{*\dagger} = w^*, \pi^{b\dagger} = \pi^{\mathrm{b}}$ *or* $q^\dagger = q$, $\hat{J}_{\mathrm{MO2}} \xrightarrow{p} J$.

The result again essentially follows by showing that $|\hat{J}_{\mathrm{TI1}} - J - \mathbb{P}_n\{\phi(s,a,r,s')\}| \leq \alpha_1\beta + \alpha_2\beta + o_p(n^{-1/2})$, where $\phi(s,a,r,s')$ is the efficient influence function. Under the above rate assumptions, the right-hand side is $\mathrm{o}_p(n^{-1/2})$ and the result is concluded from central limit theorem.

6.4. Comparison with the case of pre-specified evaluation policy

Comparison with Kallus and Uehara (2019a): When the evaluation policy is pre-specified, Kallus and Uehara (2019a) proposed an estimator that is similar but uses $\hat{q}^{(k)}(s^{(i)}, a^{(i)})$ in place of the last $\hat{v}^{(k)}(s^{(i)})$ in Eq. (8). Under similar rate conditions to Theorem 8, they prove it is efficient when the evaluation policy is pre-specified, achieving the efficiency bound $\Upsilon_{\mathrm{PR}} = \mathrm{E}\left[\mu^{*2}(s,a)\mathrm{var}\left\{r + \gamma v(s') \mid s, a\right\}\right]$. Notice that Υ_{PR} is smaller than Υ_{TI2} by $\mathrm{E}\left(w^{*2}(s)\mathrm{var}\left[\eta(s,a)q(s,a) \mid s\right]\right)$. Hence the more q-functions vary over actions, the greater the difference. As in Remark 5, naïvely plugging in an estimated π^{e} can fail to be efficient or even $\sqrt{n}$-consistent for evaluating natural stochastic policies.

Comparison with Tang et al. (2020): In the case of a pre-specified evaluation policy and known behavior policy, Tang et al. (2020) propose an estimator with a form similar to Eq. (8) without sample splitting and where $\hat{v}$ is directly estimated (rather than computed as a function of other nuisance estimates). The similarity to Eq. (8) appears to be coincidental and superficial. In the case of pre-specified evaluation policy, even if we used oracle values for all nuisances, the estimator of Tang et al. (2020) is inefficient since its variance would be equal to Υ_{TI2}, which is larger than Υ_{PR}. Tang et al. (2020) do not study the asymptotic properties of their estimator, but we can show that using cross-fitting of nuisance estimates, the asymptotic mean squared error of the estimator is also equal to Υ_{TI2}. See Appendix A. Again, this is inefficient in the case of a pre-specified evaluation policy. And, in the case of a natural stochastic policy, $\hat{v}$ must be computed in a fashion compatible with $\hat{q}$ and $\hat{\pi}^b$, that is, $\hat{v}(s) = \int \hat{q}(s,a)\hat{\pi}^e(a \mid s)\mathrm{d}(a)$ in order to ensure the partially doubly robust structure and hence efficiency.

7. EMPIRICAL STUDY

7.1. Taxi Environment

In this section, we first confirm the doubly robust property of the proposed estimators, then examine the performance of different off-policy evaluation estimators in a time-invariant infinite-horizon Markov decision process setting. We use two standard environments also used in Liu et al. (2018). In Appendix F, we examine the performance of proposed estimators in a time-variant finite-horizon Markov decision process setting.

The first environment is the Taxi environment is a commonly used tabular Markov decision process environment for off-policy evaluation, which has $\mathcal{S} = \{1, \ldots, 2000\}$, $\mathcal{A} = \{1, \ldots, 6\}$ (Dietterich, 2000; we also refer the reader to Liu et al., 2018, Section 5 for more details), and we consider separate experiments for the case of tilting and modified treatment policies. We consider our data coming from observing a single trajectory of varying length $n \in \{1, 2.5, 5, 10\} \times 10^4$. For each n we run 60 replications of the experiment. We compare the marginal IS estimator $\hat{J}_{\mathrm{MIS}}$ (Liu et al., 2018), the direct method $\hat{J}_{\mathrm{DM}}$, and one of our proposed estimators $\hat{J}_{\mathrm{TI2}}, \hat{J}_{\mathrm{MO2}}$,

depending on whether we are considering tilting or modified treatment policies. Note $\hat{J}_{\mathrm{MIS}}$ and $\hat{J}_{\mathrm{DM}}$ are given by setting $\hat{v}^{(k)} = 0$ and $\hat{w}^{(k)} = 0$, respectively, in $\hat{J}_{\mathrm{TI2}}, \hat{J}_{\mathrm{MO2}}$ (that is, the target policies are unknown in both $\hat{J}_{\mathrm{MIS}}$ and $\hat{J}_{\mathrm{DM}}$ and we use a plug-in estimate). We do not compare to step-wise IS (Precup et al., 2000) and DR (Jiang and Li, 2016) as these estimators do not converge when given single-trajectory data (as shown in Kallus and Uehara, 2019a, Section 7). Behavior and evaluation policies are set as follows. We run 150 iterations of q-learning to learn a near-optimal policy for the Markov decision process and define this to be π^b. We consider evaluating either a tilting policy with $u(a) = \lceil a/2 \rceil$ or a modified treatment policy with $\tau(s,a) = (s+a) \bmod 6$. We set $\gamma = 0.98$. We estimate π^{b} as $\hat{\pi}^b(a \mid s) = \sum_{i=1}^n \mathbb{I}\left[a^{(i)} = a, s^{(i)} = s\right] / \sum_{i=1}^n \mathbb{I}\left[s^{(i)} = s\right]$ and w^*- and q-functions by solving Eqs. (10) to (12) using $\{\mathbb{I}[s = i] : i = 1, \ldots, 2000\}$ and $\{\mathbb{I}[s = i, a = j] : i = 1, \ldots, 2000, j = 1, \ldots, 6\}$ as test functions, respectively. We use these nuisance estimates to construct all estimators. To validate double robustness, we also add Gaussian noise $\mathcal{N}(3.0, 1.0)$ to either the q- or w^*-function estimates to simulate misspecification. In Fig. 1–6, we report the mean squared error of each estimator over the 60 replications ($1/60 \sum_{\ell=1}^{60} |\hat{J}_\ell - J|^2$) with 95% confidence intervals. The mean squared error results are virtually the same with cross-fitting, Algorithm 1, or without cross-fitting, Algorithm 2.

We find the performance of $\hat{J}_{\mathrm{TI2}}, \hat{J}_{\mathrm{MO2}}$ is consistently good, with or without of model specification due to double robustness. While MIS and DM fail when their respective model is misspecified, they do well when well-specified. Since either parametric misspecification or nonparametric rates for w^* and q is unavoidable in practice for large or continuous state-action spaces, $\hat{J}_{\mathrm{TI2}}$ and $\hat{J}_{\mathrm{MO2}}$ are seen to be superior to $\hat{J}_{\mathrm{DM}}$ and $\hat{J}_{\mathrm{MIS}}$.

7.2. *CartPole Environment*

We next consider the CartPole environment where the state space is continuous and four-dimensional and the action space is binary (Brockman et al., 2016). We set the target and behavior policy in the following way. First, we run Deep Q-Network (DQN) in an online manner to learn q^*, following OpenAI's default implementation. Then, we define the behavior policy as $\pi^{\mathrm{b}}(a \mid s) \propto \exp(q^*(s,a))$, and we consider a tilting evaluation policy with $u(a) = \exp(2.0a)$. The training data is generated by executing the behavior policy with a fixed horizon length $H = 1000$. In other words, if the agent visits the terminal absorbing state before 1000 steps, the rest of the trajectory will consist of repeating the last state. We consider observing $N \in [50, 100, 200, 400]$ trajectories, *i.e.*, $n = N \times H \in \{5, 10, 20, 40\} \times 10^4$ transitions.

We compare to $\hat{J}_{\mathrm{MIS}}$, $\hat{J}_{\mathrm{DM}}$, and $\hat{J}_{\mathrm{DRL}}$, the latter being the the DRL estimator of Kallus and Uehara (2019a) with naïvely plugged-in target evaluation policy as discussed in Section 6.4, and our proposed estimator $\hat{J}_{\mathrm{TI2}}$ using π_b-, w- and q-estimators as explained in Appendix E. We also compare these to DualDICE (Nachum et al., 2019), which is a variant of $\hat{J}_{\mathrm{MIS}}$ where w is estimated in a minimax fashion using two neural networks. We choose hyperparameters to be the same as in the implementation of Kallus and Uehara (2019a).

Results and Discussion: We run 60 replications of the experiment for each N. To enhance interpretability, we consider the mean squared error of each algorithm relative to $(J - J_{\pi^{\mathrm{b}}})^2$, where $J_{\pi^{\mathrm{b}}}$ is the value of the behavior policy π^{b}. To estimate the latter normalizer, we estimate each of $J, J_{\pi^{\mathrm{b}}}$ as a simple sample average using 1000 on-policy trajectories following Kallus and Uehara (2019a).

In Table 1, we report the log mean squared error for varying N along with standard errors. We observe that $\hat{J}_{\mathrm{TI2}}$ clearly outperforms the other estimators. MIS always performs worse than $\hat{J}_{\mathrm{TI2}}$ and $\hat{J}_{\mathrm{TI2}}$. This suggests that w-estimation is more challenging than q-estimation in this

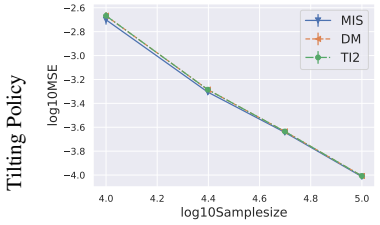

Fig. 1: Well-specified q, w^*

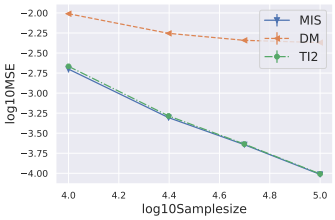

Fig. 2: Misspecified q

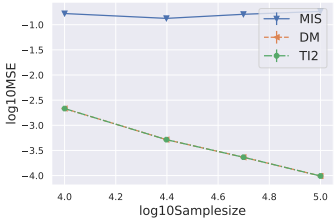

Fig. 3: Misspecified w^*

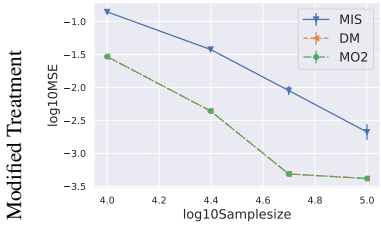

Fig. 4: Well-specified q, w^*

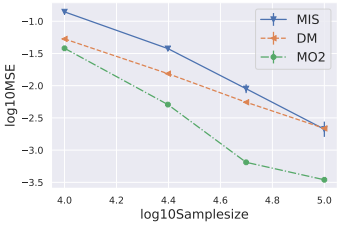

Fig. 5: Misspecified q

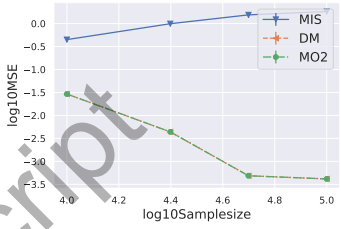

Fig. 6: Misspecified w^*

Table 1: Comparison of estimators in Cartpole with varying $N \in [50, 100, 200, 400]$. We report log relative mean squared errors and their estimated standard errors. In Fig. 1,3,4,6, the orange line overlaps with the green line.

	50	100	200	400
DM	$-1.05(-0.03)$	$-1.46(-0.02)$	$-1.73(-0.02)$	$-1.82(-0.03)$
MIS	$-0.52(-0.10)$	$-0.84(-0.23)$	$-0.87(-0.13)$	$-0.92(-0.09)$
DualDICE	$-0.32(-0.07)$	$-0.34(-0.09)$	$-0.31(-0.07)$	$-0.30(-0.09)$
Naïve DRL	$-1.13(-0.04)$	$-1.45(-0.03)$	$-1.75(-0.02)$	$-1.81(-0.02)$
TI2 (proposed)	$-1.21(-0.03)$	$-1.53(-0.03)$	$-1.79(-0.02)$	$-1.92(-0.02)$

environment.We observe DualDICE has the worst performance, which can be attributed to the instability of the minimax optimization of the two neural networks in its w-estimation. The mean squared error results are virtually the same with cross-fitting, Algorithm 1, or without cross-fitting, Algorithm 2.

8. Conclusions

We considered the evaluation of natural stochastic policies in reinforcement learning, both in finite and infinite horizons. We derived the efficiency bounds and proposed estimators that achieved them under lax conditions on nuisance estimators that permit flexible machine learning methods. An important next question is learning natural stochastic policies. More specifically, we can apply a policy-based learning approach in the sense that we optimize our policy-value estimates among the class of natural stochastic policies. As mentioned in Kennedy (2019), this direction can be important in making natural stochastic policies more prescriptive. Another important next direction is considering what other natural stochastic policies beyond tilting and modified treatment policies admit doubly or partially doubly robust evaluation.

REFERENCES

Agarwal, A., N. Jiang, S. M. Kakade, and W. Sun (2019). Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*.

Agarwal, A., S. M. Kakade, J. D. Lee, and G. Mahajan (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research 22*(98), 1–76.

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

Bibaut, A., I. Malenica, N. Vlassis, and M. Van Der Laan (2019). More efficient off-policy evaluation through regularized targeted learning. In *Proceedings of the 36th International Conference on Machine Learning*, Volume 97, pp. 654–663.

Boruvka, A., D. Almirall, K. Witkiewitz, and S. A. Murphy (2018). Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association 113*(523), 1112 – 1121.

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

Chen, Q., V. Syrgkanis, and M. Austern (2022). Debiased machine learning without sample-splitting for stable estimators. *arXiv preprint arXiv:2206.01825*.

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

Chernozhukov, V., W. Newey, and V. Semenova (2019). Inference on average welfare with high-dimensional state space.

Díaz, I. and M. J. van der Laan (2013). Assessing the causal effect of policies: an example using stochastic interventions. *The international journal of biostatistics 9*, 161–174.

Díaz, I. and M. J. van der Laan (2018). *Stochastic Treatment Regimes*, pp. 219–232. Springer International Publishing.

Díaz, I., N. Williams, K. L. Hoffman, and E. J. Schenck (2020). Non-parametric causal effects based on longitudinal modified treatment policies. *arXiv preprint arXiv:2006.01366*.

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of Artificial Intelligence Research 13*, 227–303.

Duan, Y. and M. Wang (2020). Minimax-optimal off-policy evaluation with linear function approximation. *ICML 2020 (To appear)*.

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

Díaz, I. and N. S. Hejazi (2020). Causal mediation analysis for stochastic interventions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

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

Ertefaie, A. and R. L. Strawderman (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika 105*, 963–977.

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

Gustafson, D. H., F. M. McTavish, M.-Y. Chih, A. K. Atwood, R. A. Johnson, M. G. Boyle, M. S. Levy, H. Driscoll, S. M. Chisholm, L. Dillenburg, et al. (2014). A smartphone application to support recovery from alcoholism: a randomized clinical trial. *JAMA psychiatry 71*(5), 566–572.

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

Haneuse, S. and A. Rotnitzky (2013). Estimation of the effect of interventions that modify the received treatment. *Statistics in Medicine 32*, 5260–5277.

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

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

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

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

Kallus, N. and M. Uehara (2020b). Statistically efficient off-policy policy gradients. *ICML 2020 (To appear)*.

Kennedy, E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association 114*, 645–656.

Kosorok, M. R. and E. E. Moodie (2015). *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. USA: Society for Industrial and Applied Mathematics.

Lawless, J., J. Kalbfleisch, and C. Wild (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology 61*(2), 413–438.

Liao, P., K. Greenewald, P. Klasnja, and S. Murphy (2020). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4*(1), 1–22.

Liao, P., P. Klasnja, and S. Murphy (2020). Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association (To appear)*.

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

Luckett, D. J., E. B. Laber, A. R. Kahkoska, D. M. Maahs, E. Mayer-Davis, and M. R. Kosorok (2018). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 1–34.

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

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

Muñoz, I. D. and M. Van Der Laan (2012). Population intervention causal effects based on stochastic interventions. *Biometrics 68*, 541–549.

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

Narita, Y., S. Yasui, and K. Yata (2019). Efficient counterfactual learning from bandit feedback. *AAAI*.

Precup, D., R. S. Sutton, and S. P. Singh (2000). Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766.

Riley, W., J. Obermayer, and J. Jean-Mary (2008). Internet and mobile phone text messaging intervention for college smokers. *Journal of American College Health 57*(2), 245–248.

Robins, J. (2004). Optimal structural nested models for optimal sequentialdecisions. *In Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*.

Robins, J. M., A. Rotnitzky, and D. O. Scharfstein (1999). *Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models*, Volume 116 of *Statistical Models in Epidemiology: The Environment and Clinical Trials*. NY: Springer-Verlag.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*, 846–866.

Rosenbaum, P. R. (1983). The central role of the propensity score in observational studies for causal effects. *70*, 41–55.

Schulman, J., S. Levine, P. Abbeel, M. Jordan, and P. Moritz (2015). Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR.

Shi, C., S. Zhang, W. Lu, and R. Song (2020). Statistical inference of the value function for reinforcement learning in infinite horizon settings. *arXiv preprint arXiv:2001.04515*.

Shpitser, I. and J. Pearl (2012). Effects of treatment on the treated: Identification and generalization. *Proceedings of the 25th Conference On Uncertainty in Artificial Intelligence*.

Tang, Z., Y. Feng, L. Li, D. Zhou, and Q. Liu (2020). Harnessing infinite-horizon off-policy evaluation: Double robustness via duality. *ICLR 2020 (To appear)*.

Tchetgen Tchetgen, E. J. and S. Vansteelandt (2013). Alternative identification and inference for the effect of treatment on the treated with an instrumental variable.

Thomas, J. G. and D. S. Bond (2015). Behavioral response to a just-in-time adaptive intervention (jitai) to reduce sedentary behavior in obese adults: Implications for jitai optimization. *Health Psychology 34*(S), 1261.

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

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

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

Wang, R., D. P. Foster, and S. M. Kakade (2020). What are the statistical limits of offline rl with linear function approximation?. *arXiv preprint arXiv:2010.11895*.

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

Young, J. G., M. A. Herńan, and J. M. Robins (2014). Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiologic methods 3*, 1–19.

Young, J. G., R. W. Logan, J. M. Robins, and M. A. Hernán (2019). Inverse probability weighted estimation of risk under representative interventions in observational studies. *Journal of the American Statistical Association 114*, 938–947.

Zhang, B., A. Tsiatis, E. Laber, and M. Davidian (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 681.

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