Efficient Evaluation of Natural Stochastic Policies in Offline Reinforcement Learning

Nathan Kallus
Cornell University
New York, NY 10044, USA

Masatoshi Uehara*
Harvard University
Cambridge, MA 02138, USA

Abstract

We study the efficient off-policy evaluation of natural stochastic policies, which are defined in terms of deviations from the behavior policy. This is a departure from the literature on off-policy evaluation where most work 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 very lax conditions. These also enjoy a (partial) double robustness property.

1 Introduction

In many emerging application domains for reinforcement learning (RL), exploration is highly limited and simulation unreliable, such as in healthcare (Gottesman et al., 2019). In these domains, we must use offline RL, where we evaluate and learn new sequential decision policies from existing observational data (Murphy, 2003; Zhang et al., 2013; Bibaut et al., 2019; Kallus and Uehara, 2019c). A key task in offline RL is that of off-policy evaluation (OPE), in which we evaluate a new policy from data logged by another behavior policy. Recent work in OPE (Kallus and Uehara, 2019a,b) has shown how efficiently leveraging problem structure, such as Markovianness and time-homogeneity, can significantly improve OPE and address issues such as the curse of horizon (Liu et al., 2018).

In most of the literature on OPE, 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 is observed in the data without intervention (Muñoz and Van Der Laan, 2012; Shpitser and Pearl, 2012; Haneuse and Rotnitzky, 2013; Young et al., 2014, 2019; Richardson and Robins, 2013; Díaz and van der Laan, 2018). Specifically, we consider policies defined as deviations from the behavior policy that generated the observed data.

There are two primary advantages to natural stochastic policies. A first advantage is implementability. People 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 $a$ (i.e., the natural value) would be a realistic intervention for small to
We setup the problem and notation, and summarize literature on OPE and natural stochastic policies.

### 2 Setup and Background

We setup the problem and notation, and summarize literature on OPE and natural stochastic policies.

#### 2.1 Problem Setup and Definitions

Consider an $H$-long time-varying Markov decision process (TMDP), with states $s_t \in S_t$, actions $a_t \in A_t$, rewards $r_t \in \mathbb{R}$, 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 $t = 1, \ldots, H$. A policy $(\pi_t(a_t \mid s_t))_{t \leq H}$ induces a distribution over trajectories $\mathcal{T} = (s_1, a_1, r_1, \ldots, s_T, a_H, r_H)$:

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

(1)

Given an evaluation policy $\pi^e$, which is typically pre-specified but which we will consider unknown in this paper, we are interested in its value, $J = \mathbb{E}_{p_{\pi^e}} \left[ \sum_{t=1}^H r_t \right]$, where the expectation is taken with respect to (w.r.t.) the density induced by evaluation policy, $p_{\pi^e}$. In the off-policy setting, our data consists of trajectory observations from some fixed policy, $\pi^b$, known as the behavior policy:

$$\mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(n)} \sim p_{\pi^b}, \quad \mathcal{T}^{(i)} = (s_1^{(i)}, A_1^{(i)}, R_1^{(i)}, \ldots, s_H^{(i)}, A_H^{(i)}, R_H^{(i)}).$$

(Off-policy data)
In observational studies, as we consider herein, $\pi^b$ is unknown, and the observed action $A^{(i)}_j$ 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 $\{T^{(i)}_n\}_{i=1}^n$. 

We define the following variables. Let $q_t = \mathbb{E}_{p_{\pi}}[\sum_{k=t}^H r_k | s_t, a_t]$, $v_t = \mathbb{E}_{p_{\pi}}[\sum_{k=t}^H r_k | s_t]$ be the $q$- and $v$-functions for $\pi^c$. Further, let $\eta_t = \frac{\pi_t(a_t | s_t)}{\pi^c(a_t | s_t)}$, $\lambda_t = \prod_{k=1}^t \eta_k$, $w_t = \frac{p^c(s_{t+1})}{p_\pi(s_{t+1})}$, $\mu_t = \eta_t w_t$, where $p_\pi(s_t)$ is a marginal density at $s_t$ of the trajectory density induced by the TDMP and $\pi$. Here, $\lambda_t$ is a cumulative importance ratio and $\mu_t$ is the marginal importance ratio. Given trajectory data, $T^{(1)}, \ldots, T^{(n)}$, we define the empirical expectation as $\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^n f(T^{(i)})$. Unless otherwise noted, all expectations and probabilities are w.r.t. $p^b_\pi$. Define the $L_2$ norm by $\|f\|_2 = (\mathbb{E}[f^2(T)])^{1/2}$.

**Boundedness.** We assume throughout the paper that $0 \leq r_t \leq R_{\max}$, $\eta_t \leq C$, $w_t \leq C'$, $\forall t \leq H$.

### 2.2 Natural Stochastic Policies

In OPE, $\pi^c$ is often pre-specified. Our focus is instead the case where $\pi^c$ depends on the natural value of the treatment in an observational study. Importantly, in this setting, both $\pi^c$ and $\pi^b$ are unknown. Natural stochastic policies are widely studied in the non-sequential (bandit) setting ($T = 1$). However, it has not been extensively studied in the longitudinal (RL) setting. A few exceptions are 

- Kennedy (2019); Young et al. (2014). Kennedy (2019) considers OPE with binary actions under a tilting policy in an NMDP (Non-Markov Decision Process). In comparison, we allow treatment to be continuous and focus on the Markovian setting that is central to RL. Young et al. (2014) considers OPE under a modified treatment policy in an NMDP using a parametric approach. In comparison, our methods are nonparametric and globally efficient, and we focus on the Markovian setting common in RL.

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 bandit and NMDP settings (Díaz and Hejazi 2020; Muñoz and Van Der Laan 2012; Haneuse and Rotnitzky 2013).

**Definition 1** (Tilting policy). A tilting policy is given by $u_t : \mathcal{A}_t \to \mathbb{R}$ and defined as

$$\pi^c_t(a_t | s_t) = u_t(a_t) \pi^b_t(a_t | s_t) / \int u_t(\tilde{a}_t) \pi^b_t(\tilde{a}_t | s_t) d\tilde{a}_t.$$  

(2)

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

$$\pi^c_t(a_t | s_t) = \mathbb{I}(a_t = 1) \frac{\delta \pi^b_t(1 | s_t)}{1 + (\delta - 1) \pi^b_t(1 | s_t)} + \mathbb{I}(a_t = 0) \frac{\delta^{-1} \pi^b_t(0 | s_t)}{1 + (\delta - 1) \pi^b_t(0 | s_t)},$$  

(3)

as considered by Kennedy (2019) in the binary-action NMDP setting. For $\delta = 1$ we get $\pi^c = \pi^b$; as $\delta$ shrinks, we tilt toward action 0; and, as $\delta$ grows, we tilt toward action 1. The parameter $\delta$ directly controls the amount of overlap; specifically $\pi^c_t(a_t | s_t) / \pi^b_t(a_t | s_t) \leq \max(\delta, \delta^{-1})$. For the general case in Definition 1, we have that $\pi^c_t(a_t | s_t) / \pi^b_t(a_t | s_t) \leq \max_{\delta_t} u_t(\tilde{a}_t) / \min_{\delta_t} u_t(\tilde{a}_t)$ so that the variation in $u_t$ can directly control the overlap. Tilting policies ensure that $\pi^c_t(\cdot | s_t)$ is absolutely continuous w.r.t. $\pi^b_t(\cdot | s_t)$ so that the density ratio always exists. In contrast, if $\pi^c_t$ is pre-specified and $\pi^b_t$ is unknown, we cannot always ensure that the density ratio exists, let alone is bounded.

**Definition 2** (Modified treatment policy). A modified treatment policy is given by the maps $\tau_t : \mathcal{S}_t \times \mathcal{A}_t \to \mathcal{A}_t$ and assigns the action $\tau_t(s_t, a_t)$ in state $s_t$ when the natural action value is $a_t$.

---

1 Although we do not explicitly use a counterfactual notation this is the same as the counterfactual value of following $\pi^c$ instead of $\pi^b$ if we had used potential outcomes and assumed the usual sequential ignorability and consistency assumption.

---

Ertefaie and Strawderman (2018). In our setting, the TDMP implicitly enforces these assumption.
Notice that the modified treatment policy has the same value as the evaluation policy defined by letting $\pi^b(\cdot | s_t)$ be the distribution of $a_t = \tau_1(s_t, a_t)$ for $\hat{a}_t \sim \pi^b(\cdot | s_t)$. For the purpose of OPE, we can therefore equivalently define the modified treatment policy as this transformation of $\pi^b$. For example, if for each $s_t$, $\tau_1(s_t, \cdot)$ has a differentiable inverse $\tilde{\tau}_1(s_t, \cdot)$, then $\pi^b(a_t | s_t) = \pi^b(\tilde{\tau}_1(s_t, a_t) | s_t)\pi^b(s_t, a_t)$, where $'$ denotes a differentiation w.r.t $a_t$. The simplest example of a modified treatment policy is $\tau_1(s_t, a_t) = a_t + b_t(s_t)$ for some function $b_t(s_t)$, for which $\pi^b(a_t | s_t) = \pi^b(a_t - b_t(s_t) | s_t)$. The function $b_t(s_t)$ quantifies the deviation from the natural value. Keeping $b_t(s_t)$ small ensures implementability.

### 2.3 Off-Policy Evaluation

Step-wise importance sampling (IS; Precup et al. 2000) and direct estimation of $q$-functions (DM; Munos and Szepesvári 2008) are two common approaches for OPE. 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). However, the asymptotic MSE of these can still grow exponentially in the horizon. Kallus and Uehara (2019a) show that the efficiency bound in the TMDP case is actually polynomial in $H, O(C' R_{\max}^2 H^2 / n)$, and give an estimator achieving it by combining marginalized IS (Xie et al. 2019) and $q$-modeling using cross-fold estimation.

All of the above methods focus on the case where $\pi^e$ is given explicitly. If the behavior policy is known, then natural stochastic policies can be regarded as given explicitly and these still apply. When $\pi^b$ is unknown, as in observational studies, we can still operationalize these methods for evaluating natural stochastic policies by first estimating $\pi^b$ from the data, plugging this into $\pi^e$, and then treating $\pi^e$ as specified by this estimate. This, however, will fail to be efficient, as in Section 2.3. In fact, the efficiency bounds for evaluating natural stochastic policies are different than in the pre-specified case.

### 3 Efficiency Bounds

In this section we calculate the efficiency bounds for evaluating natural stochastic policies in RL. We first briefly explain what the efficiency bound is (see van Der Laan and Robins 2003 for more detail).

A fundamental question is what is the smallest possible error we can hope to achieve in estimating $J$. In parametric models, the Cramér-Rao bound lower bounds the asymptotic MSE of all (regular) estimators. Our model, however, is nonparametric and consists of all TMDP distributions, i.e., any choice for $p_t(r_t | s_t, a_t), p_{t+1}(s_{t+1} | s_t, a_t)$, and $\pi_t(a_t | s_t)$. Semiparametric theory gives an answer to this question by extending the notion of a Cramér-Rao lower bound to nonparametric models. We first informally state the key property of the efficient influence function (EIF) from semiparametric theory in our setting, i.e., the estimand is $J$ and the model is all TMDP distributions. **Theorem 1** (Informal description of Theorem 25.20 of van der Vaart (1998)). The EIF $\phi(T)$ is the gradient of $J$ of smallest $L_2$ norm and it satisfies that, for any regular estimator $\hat{J}$ of $J$, $\text{AMSE}[\hat{J}] \geq \text{var}[\phi(T)]$, where $\text{AMSE}[\hat{J}]$ is the second moment of the limiting distribution of $\sqrt{n}(\hat{J} - J)$.

A regular estimator is any whose limiting distribution is insensitive to small changes of order $O(1/\sqrt{n})$ to the data-generating process $p_{\pi^b}$ that keep it a TMDP distribution. Here, $\text{var}[\phi]$ is called the efficiency bound as it is a lower bound on the asymptotic MSE of all regular estimators, which is a very general class. This class is so general in fact that this also implies local minimax bound for all estimators (see van der Vaart 1998 Theorem 25.21).

### Tilting Policies

In the next result we calculate the EIF and efficiency bound for tilting policies. **Theorem 2.** Let $\pi^e$ be as in Definition 2. Then the EIF and efficiency bound of $J$ are, respectively,

$$-J + \sum_{t=1}^H (\mu_t(r_t - v_t) + \mu_{t-1}v_t), \quad \Upsilon_{\text{TH1}} = \sum_{t=0}^H \mathbb{E}[\text{var}[\mu_t(r_t + v_{t+1}) | s_t]],$$
where \( \mu_0 = 1, v_0 = r_0 = 0 \). Moreover, \( T_{T11} \) is upper bounded by \( CC'R_{max}^2H^2 \).

Note that the function \( u_t \) that specifies the tilting policy is implicit in the variables \( \mu_t, v_t \) above, which depend on \( \pi_t^\tau \). The order of the efficiency bound, \( CC'R_{max}^2H^2 \), is the same as the case of a pre-specified evaluation policy (Kallus and Uehara 2019a) though there is some inflation.

**Remark 1** (Comparison to pre-specified evaluation policy). In a pre-specified evaluation policy case (Kallus and Uehara 2019a) show that the EIF and efficiency bound are, respectively,

\[
-J + \sum_{t=1}^{H} (\mu_t(q_t - q) + \mu_{t-1}v_t), \quad \sum_{t=0}^{H} \mathbb{E}[\mu_t^2\text{var}[r_t + v_{t+1} | s_t, a_t]].
\]

(4)

Specifically, if we let \( \pi^e \) be as in Definition 1 and assume that \( \pi^b \) is known then this is the efficiency bound. Compared with this quantity, \( T_{T11} \) is larger by \( \sum_{t=0}^{H} \mathbb{E}[\text{var}[\mu_tq_t | s_t]] \).

**Remark 2** (Non-Markovian Decision Processes). Kennedy (2019) provides the EIF for the binary-action tilting policy Eq. (3) under an NMDP. In comparison, our Theorem 2 handles the Markovian case as in Definition 1; vice versa by including the whole state-action history up to time \( t \) in the state variable \( s_t \). Therefore, using this transformation, Theorem 2 recovers the result of Kennedy (2019) as a special case. For more discussion of Kennedy (2019), refer to Appendix D.

**Remark 3** (Bandit case). When \( T = 1 \), the EIF is \( \eta_1(r_1 - v_1(s_1)) + v_1(s_1) \).

### Modified Treatment Policies

We next handle the case of modified treatment policies.

**Theorem 3.** Let \( \pi^e \) be as in Definition 2. Then the EIF and efficiency bound of \( J \) are, respectively,

\[
-J + \sum_{t=1}^{H} \mu_t(r_t - q_t) + \mu_{t-1}q_t, \quad \sum_{t=0}^{H} \mathbb{E}[\mu_t^2\text{var}[r_t + q_{t+1} | s_t, a_t]]
\]

(5)

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

**Remark 4** (Comparison between the case of pre-specified evaluation policies). Compared with the efficiency bound for a pre-specified evaluation policy, \( T_{MO1} \) is larger by \( \sum_{t=0}^{H} \mathbb{E}[\text{var}[\mu_tq_t | s_t]] \).

**Remark 5** (Bandit case). When \( T = 1 \), the EIF is \( \eta_1(r_1 - q_1(s_1, a_1)) + q_1(s_1, a_1) \). This matches the results in Díaz and van der Laan (2013), Díaz and van der Laan (2018).

### 4 Efficient and (Partially) Doubly Robust Estimation Methods

We propose efficient estimators for evaluating natural stochastic policies based on the obtained EIFs. Since both EIFs have second-order bias w.r.t. nuisances, we can obtain efficient estimators by estimating nuisances under nonparametric rate conditions and plugging these into the EIFs with a sample splitting technique (Zheng and van Der Laan 2011; Chernozhukov et al. 2018).

**Tilting Policies** We propose an estimator \( \hat{J}_{T11} \) for tilting polices in Algorithm 1. This is a meta-algorithm given estimation procedures for the nuisances \( w_t, \pi_t^\tau, q_t \), which we discuss how to estimate in Remark 3. We next prove \( \hat{J}_{T11} \) is efficient under nonparametric rate conditions on nuisance estimators, which crucially can be slower than \( O_p(n^{-1/2}) \) and do not require metric entropy conditions.

**Theorem 4 (Efficiency).** Suppose \( \forall k \leq K, \forall j \leq H, \|\tilde{w}_j^{(k)}(s_j) - w_j(s_j)\|_2 \leq \alpha_1, \|\hat{\pi}_j^{b(k)}(a_j | s_j) - \pi_j^b(a_j | s_j)\|_2 \leq \alpha_2, \|\hat{q}_j^{(k)}(s_j, a_j) - q_j(s_j, a_j)\|_2 \leq \beta \), where \( \alpha_1 = O_p(n^{-1/4}), \alpha_2 = O_p(n^{-1/4}), \beta = O_p(n^{-1/4}), \alpha_1\beta = o_p(n^{-1/2}) \). Then, \( \sqrt{n}(\hat{J}_{T11} - J) \overset{d}{\rightarrow} N(0, \Sigma_{T11}) \).
Efficient Off-Policy Evaluation for Natural Stochastic Policies

Theorem 5 (Partial double robustness). Suppose \( \forall k \leq K, \forall j \leq H, \) for some \( w_j, q_j \), \( \| \hat{w}_j^{(k)}(s_j) - w_j^{(k)}(s_j) \|_2 = o_p(1), \| \hat{\pi}_j^{(k)}(a_j | s_j) - \pi_j^{(k)}(a_j | s_j) \|_2 = o_p(1), \| \hat{q}_j^{(k)}(s_j, a_j) - q_j^{(k)}(s_j, a_j) \|_2 = o_p(1). \) As long as either \( q_j = q \) or \( w_j = w \), we have \( J_{T_{11}} \xrightarrow{p} J. \)

Remark 6 (Comparison to Kallus and Uehara (2019a)). Since we have to estimate \( \pi^b \) (and hence \( \pi^* \)) consistently anyway for our estimator to work, a careful reader might wonder whether we might as well plug in the estimated \( \pi^* \) into estimators that are efficient for the pre-specified case such as Kallus and Uehara (2019a). Specifically, we could replace Eq. (6) in Algorithm 1

\[
\phi^{(k)}(T) = \sum_{t=1}^{H} \hat{w}_t^{(k)}(s_t) \hat{\pi}_t^{(k)}(s_t, a_t)(r_t - \hat{q}_t^{(k)}(s_t, a_t)) + \hat{w}_{t-1}^{(k)}(s_{t-1}) \hat{\pi}_{t-1}^{(k)}(s_{t-1}, a_{t-1}) \hat{\pi}_t^{(k)}(s_t),
\]

which corresponds to plugging our estimated nuisances into the EIF derived in Kallus and Uehara (2019a). However, this can fail to achieve a \( \sqrt{n} \)-convergence rate, let alone fail to achieve efficiency. Specifically, in Theorem 4 we used the fact that Eq. (6) has a second-order bias structure w.r.t. \( w_t, \pi_t, q_t \) to ensure the \( \sqrt{n} \)-consistency and efficiency. In contrast, the above does not have this structure. It only has such a structure when \( \hat{v}_t^{(k)} \) is the integral of \( \hat{q}_t^{(k)} \) with respect to the true \( \pi_t^\gamma \), which Kallus and Uehara (2019a) use to achieve efficiency, but that is not the case here.

Remark 7 (Estimation of v-functions). Although \( \hat{q}_t^{(k)} \) does not explicitly appear in Eq. (6), we do need to estimate \( \hat{q}_t^{(k)} \) first and then compute \( \hat{v}_t^{(k)} \) based on it as in Algorithm 1 instead of directly estimating \( v_t \). The reason is that we cannot generally say that Eq. (6) has a second-order bias structure w.r.t. \( w_t, \pi_t, v_t \). Therefore, the efficiency result would not be guaranteed. To achieve the efficiency, it is crucial to use the specific construction of \( \hat{v}_t^{(k)} \) in Algorithm 1 which ensures a certain compatibility between the nuisance estimators, as they all use the same estimated behavior policy.
**Modified Treatment Policies** We similarly define the estimator \( \hat{J}_{\text{MO1}} \) for the case of modified treatment policies by taking Algorithm 2 and (a) replacing \( \hat{\pi}_t^e(k)(a_t | s_t) \) by \( \hat{\pi}_t^s(k)(a_t | s_t) = \hat{\pi}_t^s(k)(s_t, a_t) | s_t) \) and (b) replacing (6) by \( \phi(k)(T) = \sum_{t=1}^T \hat{w}_t^s(k)(s_t, a_t)(r_t - \hat{q}_t^s(s_t, a_t)) + \hat{w}_{t-1}^s(s_{t-1}, a_{t-1})\hat{q}_{t-1}^s(s_{t-1}, a_{t-1})\hat{q}_t^s(s_t, \tau_t(s_t, a_t)) \).

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

**Theorem 6** (Efficiency). Suppose \( \forall k, \forall j \leq H, \|w_j^s(k)(s_j) - w_j^s(s_j)\|_2 \leq \alpha_1, \|p_j^b(k)(a_j | s_j) - p_j^b(a_j | s_j)\|_2 \leq \alpha_2, \|q_j^s(s_j, a_j) - q_j^s(s_j, a_j)\|_2 \leq \beta \) where \( \alpha_1 + \alpha_2 + \beta = \alpha_p(1) \). Then, \( \sqrt{n}(\hat{J}_{\text{MO1}} - J) \overset{d}{\rightarrow} N(0, \Sigma_{\text{MO1}}) \).

**Theorem 7** (Double robustness). Assume \( \forall k, \forall j \leq H, \) for some \( \pi_j^b, q_j^s, w_j^s \), \( \|w_j^s(k)(s_j) - w_j^s(s_j)\|_2 = \alpha_p(1), \|\hat{\pi}_t^s(a_j | s_j) - \pi_j^b(a_j | s_j)\|_2 = \alpha_p(1), \|q_j^s(s_j, a_j) - q_j^s(s_j, a_j)\|_2 = \alpha_p(1) \). Then as long as either \( q_j^s = q_j \) or \( \pi_j^b = \pi_j, w_j^s = w_j \), we have \( \hat{J}_{\text{MO1}} \overset{p}{\rightarrow} J \).

These theorems arise from the bias structure \( |\hat{J}_{\text{MO1}} - J| - \mathcal{P}_n[\phi(T)]| \leq (\alpha_1 + \alpha_2 + \beta + \alpha_p(n^{-1/2})) \).

**Remark 8** (Nuisance Estimation). Our estimators require that we estimate \( \pi_t^b, w_t, q_t \) at some slow rate. Estimating \( \pi_t^b \) amounts to fitting a probabilistic classification model in the case of finitely many actions and to conditional density estimation in the case of continuous actions. Once we fit \( \pi_t^b \), we also immediately have an estimate of \( \pi_t^s \). We can then use standard methods for estimating \( w_t \) and \( q_t \) that assume \( \pi_t^e \) is given by plugging in our estimate for \( \pi_t^e \). For estimating \( q_t \) with known \( \pi_t^e \), there are many standard methods such as fitted q-iteration [Antos et al., 2008]. For estimating \( w_t \) with known \( \pi_t^e \), we refer the reader to the extensive discussion in [Kalus and Uehara, 2019a; Xie et al., 2019c; Yin and Wang, 2020]. Refer to Appendix B for more detail. Generally speaking, if the estimate for \( q_t \) or \( w_t \) would have had no convergence rate \( r_n \), if \( \pi_t^e \) were given exactly, then the estimate does not deteriorate as long as the plugged-in estimate for \( \pi_t^e \) also has rate \( r_n \). For all of our nuisances, we do not require any metric entropy conditions. And, since we only require slow nonparametric rates, this enables the use of flexible machine learning methods as blackbox nuisance estimators — our analysis does not depend on the particular method used, which is in contrast to IS and direct-method estimates.

### 5 Extension to Time-Homogeneous Markov Decision Process

We next extend the results to time-homogeneous MDPs where transitions, rewards, and policies do not depend on \( t \), i.e., \( p_t(r | s, a) = p(r | s, a), p_t(s' | s, a) = p(s' | s, a), \pi_t^b = \pi^b, \tau_t = \tau \). Here, the estimator we consider is an average discounted reward, \( J(\gamma) = (1 - \gamma) \lim_{T \to \infty} \mathbb{E}_{\pi^b, r^1} \sum_{t=1}^T \gamma^{t-1} r_t \) when the initial state distribution is \( p^{(1)}_b(s) \). Though we can still apply methods for TMDP here, if we correctly leverage the time-homogeneity, we should do better in that the rate of MSE should be \( \mathcal{O}(1/N) \), not \( \mathcal{O}(1/N) \), when we observe \( N \) trajectories of length \( H \). Per Liu et al. (2018),

\[
J(\gamma) = \mathbb{E}_{(s, a, r) \sim p^{(\infty)}_b} [r p^{(\infty)}_b(s) \pi^e(a | s) / (p^{(\infty)}_b(s) \pi^b(a | s))],
\]

where \( p^{(\infty)}_b \) is the stationary distribution associated with the MDP, \( \pi^b \), and the initial state distribution \( p^{(1)}_b(s) \), and \( p^{(\infty)}_b \) is the \( \gamma \)-discounted average visitation distribution associated with the MDP, \( \pi^e \),
and the (potentially different) initial state distribution \( p^{(1)}_i(s) \). When \( \pi^e \) is pre-specified, Kallus and Uehara (2019b) derived the efficiency bound of \( J(\gamma) \) under a nonparametric model and proposed an efficient estimator. Here, we present corresponding results for the efficient evaluation of natural stochastic policies in time-homogeneous infinite-horizon RL. We consider the observed data to be \( n \) i.i.d. draws from the stationary behavior distribution: for \( i = 1, \ldots, n \),

\[
(s^{(i)}), a^{(i)}, s'^{(i)}, a'^{(i)}) \sim p^{(\infty)}_b(s, a, r, s', a') = p^{(\infty)}_b(s)\pi^b(a \mid s)p(s' \mid s, a)p(r \mid s, a)\pi^b(a' \mid s').
\]

We consider a fully nonparametric model in which we make no restrictions on the above distributions. In this section, we define \( q(s, a) = \mathbb{E}_{p^{(1)}}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s, a_1 = a], q^*(s, a) = q(s, \tau(s, a)), v(s) = \mathbb{E}_{p^{(1)}}[q(s, a) \mid s], w^*(s) = p^{(\infty)}_e(s)/p^{(\infty)}_b(s), \) and \( \mu^*(s, a) = w^*(s)q(s, a) \).

**Tilting Policies**

First we consider the case of tilting policies.

**Theorem 8.** Let \( \pi^e \) be as in Definition \( \square \). Then the EIF and efficiency bound of \( J(\gamma) \) are, respectively,

\[-J(\gamma) + \mu^*(s, a)(r + \gamma v(s') - v(s)), \quad \Upsilon_{T2}^e = \mathbb{E}[\text{var} \{\mu^*(s, a)(r + \gamma v(s')) \mid s]}.\]

With additional data \( s^{(j)}_i \sim p^{(1)}_c(s), j = 1, \ldots, m \) where \( m = \Omega(n) \) (or, if \( p^{(1)}_c \) is known), we propose the estimator \( \hat{J}_T^e \) for \( J(\gamma) \) by taking Algorithm \( \square \) and replacing \( J_k \) with

\[
\hat{J}_k = \frac{1 - \gamma}{m} \sum_{j=1}^{m} \hat{v}^{(k)}_j(s^{(j)}) + \frac{1}{1 - \gamma} \sum_{i \in I_k} \hat{w}^e(s^{(i)})(s^{(i)}, a^{(i)})(\mu^{(i)} + \gamma \hat{v}^{(k)}(s^{(i)})) - \hat{v}^{(k)}(s^{(i)})),
\]

\[
\hat{\mu}^e(k)(a \mid s) = u(a)\pi^e(h(k)(a \mid s) / \int u(\hat{a})\pi^e(h(k)(\hat{a} \mid s)da, \quad \hat{v}^{(k)}(s, a) = \hat{\mu}^e(k)(a \mid s)/\hat{\pi}^e(k)(a \mid s), \quad \hat{\pi}^e(k)(a \mid s), \quad \hat{w}^e(s^{(i)}), \quad \hat{w}^e(s^{(i)}),
\]

given nuisance estimators \( \hat{\pi}^e(k), \hat{\pi}^e(k), \hat{w}^e(k) \). We prove \( \hat{J}_{T2}^e \) is efficient and doubly robust in Appendix \( \square \). To estimate \( \pi^e \), we can follow Remark \( \square \). To estimate \( w^*, q \), we can solve the following moment equations using some set of test functions (cf. Liu et al., 2018; Kallus and Uehara, 2019b):

\[
0 = (1 - \gamma)\mathbb{E}_{s \sim p^{(1)}}[f(s)] + \mathbb{E}_{(s, a, s') \sim p^{(1)}}[\gamma w^*(s)q(s, a)f(s') - f(s)] \tag{10} \quad \forall f(s),
\]

\[
0 = \mathbb{E}_{(s, a, r, s') \sim p^{(1)}}[g(s, a)(r + \gamma v(s') - q(s, a))] \tag{11} \quad \forall g(s, a),
\]

**Remark 9** (Comparison of \( \hat{J}_{T2}^e \) with Kallus and Uehara (2019b) and Tang et al. (2020)). When the evaluation policy is pre-specified, Kallus and Uehara (2019b) proposed an estimator that is similar but uses \( \hat{v}^{(k)} \) in place of the last \( \hat{v}^{(k)} \) in Eq. (8). Then, under similar rate conditions to Theorem 10, they prove it achieves the asymptotic MSE \( \Upsilon_{PR}^e = \mathbb{E}[\mu^2(s, a)\text{var} \{r + \gamma v(s') \mid s, a \}], which is the efficiency bound when the evaluation policy is pre-specified so the estimator is efficient. Notice that \( \Upsilon_{PR}^e \) is smaller than \( \Upsilon_{T2}^e \) by \( \mathbb{E}[w^{*2}(s)\text{var} \{q(s, a)(r + \gamma v(s')) \mid s] \}. As in Remark \( \square \), not purely plugging in an estimated \( \pi^e \) into the EIF for the pre-specified case can fail. Also in the pre-specified evaluation policy and known behavior policy case, Tang et al. (2020) propose an estimator with a form similar to Eq. (8) (without sample splitting) although they did not calculate the asymptotic MSE, and the estimation way of \( \hat{v}^{(k)} \) is different. However, even if we used oracle values for all nuisances, the estimator of Tang et al. (2020) is inefficient in this pre-specified policy case since its variance would be \( \Upsilon_{T2}^e \), which is larger than \( \Upsilon_{PR}^e \). The similarity to Eq. (8) appears to be coincidental. For additional detail see Appendix \( \square \).
Modified Treatment Policies

First we consider the case of modified treatment policies.

**Theorem 9.** Let \( \pi' \) be as in Definition 2. Then the EIF and efficiency bound of \( J(\gamma) \) are, respectively,
\[
-J(\gamma) + \mu^*(s, a)(r + \gamma q^*(s', a') - q(s, a)), \quad \Upsilon_{MO2} = E[\mu'^2(s, a) \text{var}[r + \gamma q^*(s', a') \mid s, a]].
\]

With additional data \((s^{(j)}, a^{(j)})_{j=1}^m \sim \pi^b(a \mid s)\) from the initial state-action distribution, where \( m = \Omega(n) \), we propose the estimator \( \hat{J}_{MO2} \) by taking Algorithm 1 and replacing \( \hat{J}_k \) with
\[
\hat{J}_k = \frac{1}{|I_k|} \sum_{i \in I_k} \hat{w}^b(k)(s^{(i)}) \hat{q}(s^{(i)}, a^{(i)})(r^{(i)} + \gamma \hat{q}^{(i)} \tau(s^{(i)}, a^{(i)}) - \hat{q}^b(s^{(i)}, a^{(i)})) + \frac{(1-\gamma)}{m} \sum_{j=1}^m \hat{q}^b(s^{(j)}, a^{(j)}), \quad \hat{\pi}^b(k)(a_t \mid s_t) = \hat{\pi}^b(k)(s_t, a_t) \mid s_t) \tilde{\tau}(s_t, a_t),
\]
and \( \hat{q}^{(i)}, \tilde{\tau}(k) \) as in Eq. (9), given nuisance estimators \( \hat{w}^b(k), \hat{\pi}^b(k), \hat{q}^b(k) \). We prove \( \hat{J}_{MO2} \) is efficient and doubly robust in Appendix C. To estimate \( w^* \) we can use Eq. (10) and to estimate \( \hat{q} \) we can use:
\[
0 = E_{(s, a, r, s', a')} \hat{w}^b(k)[g(s, a) - q(s, a) + \gamma q^*(s', a')] \quad \forall g(s, a).
\]

6 Empirical Study

In this section we examine the performance of different OPE estimators in an infinite-horizon setting. We use the Taxi environment, which is a commonly used tabular environment for OPE, which has \( S = \{1, \ldots, 2000\}, A = \{1, \ldots, 6\} \) (Dietterich 2000); we refer the reader to Liu et al. 2018 Section 5 for more details), and consider the observation of a single \((N = 1)\) trajectory of varying length \( H \) \((1, 2, 5, 10) \times 10^4\). For each \( H \) we take 60 replications of the experiment. We compare the stationary marginal IS estimator \( \hat{J}_{MIS} \) (Liu et al. 2018), the direct method \( \hat{J}_{DM} \), and our proposed estimators \( \hat{J}_{TI2}, \hat{J}_{MO2} \). We do not compare to step-wise IS (Precup et al. 2000) and DR (Jiang and Li 2016) since these estimators do not work for single-trajectory data as shown in Kallus and Uehara 2019b 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 MDP and define it as \( \pi^b \). We consider evaluating either a tilting policy with \( u(a) = [a/2] \) or a modified treatment policy with \( \tau(s, a) = (s + a) \mod 6 \). We set \( \gamma \) as 0.98. We estimate \( \pi^b \) as \( \tilde{\pi}^b(a \mid s) = \sum_{i=1}^n \mathbb{I}[a^{(i)} = a, s^{(i)} = s] / \sum_{i=1}^n \mathbb{I}[s^{(i)} = s] \) 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 \( \mathcal{N}(30, 1.0) \) to either the \( q\)- or \( w^*\)-function estimates to simulate misspecification. In Figs. 1 to 4 we report the MSE of each estimator over the 60 replications with 95% confidence intervals.

We find the performance of \( \hat{J}_{TI2} \) or of \( \hat{J}_{MO2} \) 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. This is expected because both estimators are efficient in the tabular (hence parametric) setting when the evaluation policy is pre-specified (Uehara et al. 2020; Kallus and Uehara 2019b); proving the analogous result for natural stochastic policies is future work. Still, because having either parametric misspecification or nonparametric rates for \( w^* \) and \( q \) is unavoidable in practice for continuous state-action spaces, \( \hat{J}_{TI2} \) and \( \hat{J}_{MO2} \) would be superior to \( \hat{J}_{DM} \) and \( \hat{J}_{MIS} \).

7 Conclusions

We considered the evaluation of natural stochastic policies in RL, both in finite and infinite horizon. We derived the efficiency bounds and proposed estimators that achieved them under lax conditions. An important next question is learning natural stochastic policies. This can perhaps be done using an off-policy policy gradient approach by extending Kallus and Uehara 2020 to natural stochastic policies, where we take gradients in \( u \) or \( \tau \) that specify how and where we deviate from \( \pi^b \).
References

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.

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.

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. Cham: Springer International Publishing.

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

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

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.

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). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. arXiv preprint arXiv:1908.08526.

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

Kallus, N. and M. Uehara (2019c). 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 (2020). Statistically efficient off-policy policy gradients. arXiv preprint arXiv: 2002.04014.

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

Lagoudakis, M. and R. Parr (2004). Least-squares policy iteration. Journal of Machine Learning Research 4 (6), 1107–1149.

Le, H., C. Voloshin, and Y. Yue (2019). Batch policy learning under constraints. In Proceedings of the 36th International Conference on Machine Learning, pp. 3703–3712.

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.

Munos, R. and C. Szepesvári (2008). Finite-time bounds for fitted value iteration. Journal of Machine Learning Research 9, 815–857.

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

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

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.

Richardson, T. S. and J. M. Robins (2013). Single world intervention graphs (swigs): a unification of the counterfactual and graphical approaches to causality. CSSS Working Paper 128.

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).

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

Uehara, M., J. Huang, and N. Jiang (2020). Minimax weight and q-function learning for off-policy evaluation. arxiv preprint arxiv:1910.12809.

Ueno, T., M. Kawanabe, T. Mori, S.-I. Maeda, and S. Ishii (2011). Generalized td learning. Journal of Machine Learning Research 12, 1977–2020.
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.

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

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

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. Hernán, 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.
A Notations

Table 2: Notation

| Symbol | Description |
|--------|-------------|
| $H, K$ | Number of horizon, number of splitting folds |
| $r_t, s_t, a_t$ | Reward, state, action at $t$ |
| $p_t(r_t | s_t, a_t), p_{t+1}(s_{t+1} | s_t, a_t)$ | Reward density, transition density |
| $J_T, J_s, J_a$ | History up to time $r_t, s_t, a_t$, including reward variables |
| $H_{a_t}, H_{a_t}$ | History up to time $s_t, a_t$, excluding reward variables |
| $\pi_t(a_t | r_t), \pi_t(a_t | s_t), \pi(a_t | s)$ | Policy in NMDP, TMDP, MDP case, respectively |
| $\pi^*_t, \pi_t^*$ | Target and behavior policies at $t$, respectively |
| $J$ | Empirical expectation (based on sample from a behavior policy) |
| $J(\gamma)$ | Asymptotic variance, variance |
| $v_t = v_t(H_{s_t}), v_t(s_t), v(s)$ | Empirical process |
| $q_t = q_t(H_{a_t}), q_t(s_t, a_t), q(s, a)$ | Tangent space |
| $q'(s, a)$ | Marginal state density ratio in TMDP, MDP case, respectively |
| $\lambda_t$ | Marginal state action density ratio in TMDP, MDP case, respectively |
| $w_t(s), w^*(s)$ | Instantaneous density ratio $\pi_t^*/\pi_t^*$ |
| $\mu_t(s, a), \mu^*(s, a)$ | Tangent space |
| $\eta_t$ | Upper bound of density ratio, marginal ratio, and reward, respectively |
| $\Lambda$ | Projection of $\Lambda$ onto $B$ |
| $C, C', R_{\max}$ | Direct sum |
| $\Pi(A|B)$ | $L^p$-norm $E[|f|^p]^{1/p}$ |
| $\| \cdot \|_p$ | Inequality up to an universal constant |
| $\leq$ | Expectation with respect to a sample from a policy $\pi$ |
| $E_\pi[], P_\pi$ | Same as the above for $\pi = \pi^b$ |
| $E[\cdot], \mathbb{P}$ | Empirical expectation (based on sample from a behavior policy) |
| $P_n[\cdot], \mathbb{P}_n$ | The size of $\mathcal{U}_t$ |
| $\mathcal{U}_j$ | Empirical expectation on $\mathcal{U}_j$ |
| $\mathcal{U}_{a,j}$ | Empirical process $\sqrt{n}(\mathbb{P}_n - \mathbb{P})$ |
| $\mathcal{G}_n$ | Asymptotic variance, variance |
| $\text{AMSE}[], \text{var}[]$ | Normal distribution with mean $a$ and variance $b$ |
| $\mathcal{N}(a, b)$ | Deterministic policy |
| $\tau(s, a)$ | Blip function in a tilting policy $\pi^b(a | s; \pi^b)$ |
| $u(a; \delta), u(a)$ | Normalizing constant of a tilting policy |
| $\sigma(s)$ | Inverse function of $a \rightarrow \tau(s, a)$ |
| $\tilde{\tau}(s, \cdot)$ | Derivative w.r.t. $a \rightarrow \tilde{\tau}(s, a)$ |
| $\tilde{\tau}'(s, \cdot)$ | Estimators for tilting policies, modified treatment policies |
| $J_{TT}, J_{MO}$ | Stationary distribution generated by an MDP and a behavior policy |
| $p^{(c)}_b(s, a, r, s')$ | Initial (evaluation) distribution |
| $p^{(1)}_b(s)$ | Average visitation distribution with $\gamma$ generated by an MDP and an evaluation policy |
| $f(n) = \mathcal{O}(n^a)$ | $f(n)$ is bounded above by $n^a$ asymptotically |
| $f(n) = \Omega(n^a)$ | $f(n)$ is bounded below by $n^a$ asymptotically |
| $f(n) = \mathcal{O}_p(n^a)$ | $f(n)/n^a$ is bounded in probability |
| $f(n) = \omega_p(n^a)$ | $f(n)/n^a$ converges to 0 in probability |
| $\Upsilon$ | Each efficiency bound |
| $\mathcal{S}_d$ | Convergence in distribution |
| $\rightarrow$ | Convergence in probability |
B  Nuisance estimation

Our algorithm allows any estimators for \( q \)-functions and marginal density ratios to be used. Here we discuss some standard ways to estimate these nuisance functions. We focus on the case of tilting policies for brevity.

B.1  Estimation of \( q \)-functions

In the tabular case, a model-based approach is the most common way to estimate \( q \)-functions from off-policy data. In the non-tabular case, we have to rely on some function approximation. The key equation to derive these methods is the Bellman equation:

\[
q_t(s_t, a_t) = \mathbb{E}[r_t + q_{t+1}(s_{t+1}, \pi^e) \mid s_t, a_t],
\]

where \( q_t(s_t, \pi) = \int q_t(s_t, a_t)\pi(a_t \mid s_t)da_t \). One of the most common ways to operationalize this is using fitted \( q \)-iteration \cite{Antos2008, Le2019} expressed here using an estimated evaluation policy, \( \hat{\pi}^e \):

- Set \( \hat{q}_{H+1} = 0 \).
- For \( t = H, \ldots, 1 \):
  - Estimate \( \hat{q}_t \) by regressing \( r_t + \hat{q}_{t+1}(s_{t+1}, \hat{\pi}^e) \) onto \( s_t, a_t \).

Note that the regressions can also be replaced with \( Z \)-estimation \cite{Ueno2011} based on the moment equations:

\[
\mathbb{E}[g_t(s_t, a_t) \{r_t + q_{t+1}(s_{t+1}, \pi^e) - q_t(s_t, a_t)\}] = 0, \quad \forall g_t, t \leq H.
\]

When we let \( q_t \) and \( g_t \) be the class of linear functions, this leads to LSTD \cite{Lagoudakis2004}.

B.2  Estimation of Marginal Density Ratios

In the tabular case, a model-based approach \cite{Yin2020} would be a competitive way to estimate marginal density ratios:

\[
\hat{w}_t(s_t) = \frac{1}{\hat{p}_{t+1}(s_t)} \int \hat{p}_t(s_t \mid s_{t-1}, a_{t-1}) \prod_{k=0}^{t-1} \left( \pi^e_k(a_k \mid s_k)\hat{p}_k(s_k \mid s_{k-1}, a_{k-1}) \right) d\mathcal{H}_{a_{t-1}},
\]

where \( \hat{p}_t, \hat{p}_{t+1} \) are each an empirical frequency (histogram) estimator. When the behavior policy is known, \cite{Xie2019} also proposed a similar method to estimate \( w_t \). In the non-tabular case, we have to rely on some function approximation methods. We have \( w_t = \mathbb{E}[\lambda_{t-1} \mid s_t] \). Thus, for example, \( w_t \) is estimated by regressing \( \lambda_{t-1} \) onto \( s_t \) \cite{Kallus2019}. The problem of this approach is it would incur an error with an exponential horizon dependence. How to circumvent this problem is a future work.

C  Theoretical Results in Section 5

**Tilting Policies** We prove the proposed estimator \( \hat{J}_{\text{TT2}} \) is efficient under nonparametric rate conditions for nuisance estimators. Partially double robustness similarly holds.

**Theorem 10** (Efficiency). Suppose \( \forall k \leq K, \|\hat{w}^{(k)}(s) - w^*(s)\|_2 = \alpha_1, \|\hat{q}^{(k)}(s, a) - q(s, a)\|_2 = \beta, \) where \( \alpha_1 = \mathcal{O}(n^{-1/4}), \) \( \alpha_2 = \mathcal{O}(n^{-1/4}), \) \( \beta = \mathcal{O}(n^{-1/4}), \) \( \alpha_1\beta = o_p(n^{-1/2}) \). Then, \( \sqrt{n}(\hat{J}_{\text{TT2}} - J) \overset{d}{\to} \mathcal{N}(0, \Sigma_{\text{TT2}}). \)
Theorem 11 (Partially Double Robustness). Assume $\forall k \leq K$, for some $w^*(s), q^*(s, a)$, $\|\hat{w}^{(k)}(s) - w^*(s)\|_2 = o_p(1), \|\hat{q}^{(k)}(s, a) - q^*(s, a)\|_2 = o_p(1)$. Then, as long as $w^*(s) = w(s)$ or $q^*(s, a) = q(s, a)$, $\hat{J}_{\text{MO2}} \to J$.

Modified Treatment Policies We prove that the proposed estimator $\hat{J}_{\text{MO2}}$ is efficient under nonparametric rate conditions for nuisance estimators. Double robustness similarly holds.

Theorem 12 (Efficiency). Assume $\forall k \leq K$, $\|\hat{w}^{(k)}(s) - w^*(s)\|_2 = \alpha_1, \|\hat{q}^{(k)}(s, a) - q^*(s, a)\|_2 = \alpha_2$, where $(\alpha_1 + \alpha_2)\beta = o_p(n^{-1/4}), \max\{\alpha_1, \alpha_2, \beta\} = o_p(1)$. Then, $\sqrt{n}(\hat{J}_{\text{MO2}} - J) \xrightarrow{d} N(0, \Sigma_{\text{MO2}})$.

Theorem 13 (Double Robustness). Assume $\forall k \leq K$, for some $w^*, \pi^*, q^*$, $\|\hat{w}^{(k)}(s) - w^*(s)\|_2 = o_p(1), \|\hat{q}^{(k)}(s, a) - q^*(s, a)\|_2 = o_p(1)$. Then, as long as $w^* = w, \pi^* = \pi^b$ or $q^* = q$, $\hat{J}_{\text{MO2}} \to J$.

D Results for NMDPs and Relations with \cite{Kennedy2019}

TMDP, NMDP So far, we have so focused on the TMDP setting where the trajectory distribution $p_\pi$ is given by Eq. (1). For completeness, we also consider relaxing the Markov assumption, yielding a non-Markov decision process (NMDP), where the trajectory distribution $p_\pi(T)$ is

$$p_\pi(s_1) \prod_{t=1}^H \pi_t(a_t \mid h_s, r_t) p_t(r_t \mid h_s, a, p_{t+1}(s_{t+1} \mid h_s),$$

where $h_s$ is $(s, a, \cdots, a)$ and $h_s$ is $(s, a, \cdots, s)$. From now on, we write $\mathcal{H}_{s_t}$ as $\{S_1, A_1, \cdots, A_t\}$ and similarly for $\mathcal{H}_{a_t}$ and $\mathcal{H}_{r_t}$.

D.1 Efficiency Bounds and Estimators under NMDP

We can extend Theorem 2 to the NMDP case as follows.

Theorem 14. Under NMDP, the EIF and the efficiency bound are

$$\sum_{t=1}^H \lambda_t \{y_t - v_t(h_s)\} + \lambda_{t-1} v_t(h_s),$$

$$\sum_{t=0}^H \mathbb{E}[\lambda_t^2 (\mathcal{H}_{S_{t-1}}) \text{var}[\eta_t(\mathcal{H}_A)\{R_t + v_{t+1}(\mathcal{H}_{S_{t+1}})\} \mid \mathcal{H}_{S_t}]].$$

The intuitive proof is as follows. Consider a transformation of NMDP to a TMDP. In a transformed NMDP, the marginal density ratio $\mu_t$ is a cumulative importance ratio in an original NMDP. The efficiency bound in this TMDP is equal to the efficiency bound in Theorem 14.

Similarly, we can extend Theorem 3 to the NMDP case as follows.

Theorem 15. Under NMDP, the EIF and efficiency bound are

$$\sum_{t=1}^H \lambda_t \{r_t - q_t(h_a)\} + \lambda_{t-1} q_t(h_s, \tau_t(h_a)),$$

$$\sum_{t=0}^H \mathbb{E}[\lambda_t^2 \text{var}[r_t + q_{t+1}(\mathcal{H}_{S_{t+1}}, \tau(\mathcal{H}_{A_{t+1}})) \mid \mathcal{H}_{A_t}]].$$

D.2 Relation with Kennedy (2019)

We explain how our result Theorem [14] reduces to Theorem 2 (Kennedy (2019)). Note that in Section 8.2 of Appendix in Kennedy (2019), they derived the EIF for general natural stochastic policies under an NMDP. However, specific forms regarding specific policies such as Theorem [14] and Theorem [15] were not obtained in their paper. In addition, our results under the TMDP and MDP are totally novel compared with theirs.

Lemma 1. Consider the policy \(\bar{\pi}\) and suppose the reward is observed only at \(H\). Then, the EIF in Theorem [14] is equal to that in Kennedy (2019):

\[
\sum_{t=1}^{H} a_t \frac{1 - \pi_t(h_s)}{\delta/(1 - \delta)} - (1 - a_t) \delta \pi_t(h_s_t) \left[ \frac{\delta \pi_t(h_s_t) q_t(h_s_t, 1) + (1 - \pi_t(h_s_t)) q_t(h_s_t, 0)}{\delta \pi_t(h_s_t) + (1 - \pi_t(h_s_t))} \right] \\
\times \left( \prod_{s=1}^{t} \frac{\delta a_s + (1 - a_s)}{\delta \pi_s(h_s) + 1 - \pi_s(h_s)} \right) + \sum_{s=1}^{H} \frac{\delta a_s + (1 - a_s) r_H}{\delta \pi_s(h_s) + 1 - \pi_s(h_s)},
\]

where \(\pi(h_s_t) = \pi^b(1|h_s_t)\).

Proof. We explain the above is equal to

\[
\lambda_H R_H + \sum_{t=1}^{H} \{-\lambda_t v_t + \lambda_{t-1} v_t\}
\]

in Theorem [14]. What we need is proving

\[
-\lambda_t + \lambda_{t-1} = \frac{a_t \{1 - \pi_t(h_s_t)\} - (1 - a_t) \delta \pi_t(h_s_t)}{\delta/(1 - \delta)} \lambda_t.
\]

This is proved by

\[
\left\{ 1 + \frac{a_t \{1 - \pi_t(h_s_t)\} - (1 - a_t) \delta \pi_t(h_s_t)}{\delta/(1 - \delta)} \right\} \lambda_t = \left\{ 1 + \frac{a_t \{1 - \pi_t(h_s_t)\} - (1 - a_t) \delta \pi_t(h_s_t)}{\delta/(1 - \delta)} \right\} \frac{\delta a_t + (1 - a_t)}{\delta \pi_t(h_s_t) + 1 - \pi_t(h_s_t)} \lambda_{t-1} = \lambda_{t-1}.
\]

The efficient estimator is constructed by plug the nuisance estimators \(\hat{\pi}_t^b, \hat{q}_t\) into the EIF as in Algorithm 1. Kennedy (2019) proposed an estimation method with nuisance estimators \(\hat{\pi}_t^b, \hat{q}_t\) under an NMDP by plugging them into the EIF in Lemma 1. They also proved that the estimator is efficient as long as \(\|\pi_t^b - \pi_t^\ast\| = o_p(n^{-1/4}), \|\hat{q}_t - q_t\| = o_p(n^{-1/4})\). This theorem is easily extended to a continuous tilting policy case. They mentioned the doubly robustness property regarding \(\pi_t^b, q_t\) does not hold in the sense that it always requires the correct specification of \(\pi_t^b\). Unlike the NMDP case, in the TMDP or MDP, we can see a partially doubly robustness result such as Theorem [10].

E Comparison with Tang et al. (2020)

We mainly discuss theoretical properties of the estimator \(\hat{J}_{V2}\) proposed in Tang et al. (2020). We also highlight the difference between \(\hat{J}_{V2}\) and \(\hat{J}_{MO2}\), and analyze the difference between \(\hat{J}_{V2}\) and the efficient estimation in the pre-specified evaluation policy case (Kallus and Uehara 2019b).

When the evaluation policy is pre-specified and the behavior policy is known, Tang et al. (2020) proposed an estimator \(\hat{J}_k\) by replacing (6) in Algorithm 1 with

\[
\hat{J}_k = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left[ \hat{\mu}^*(k)(S^{(i)}) \eta(S^{(i)}, A^{(i)}) \{ R^{(i)} + \gamma \hat{v}^*(k)(S^{(i)}) - \hat{v}^*(k)(S^{(i)}) \} \right]
\]

\[4\text{In their paper, they did not use a sample splitting. This is a version with a sample splitting.}\]
Theorem 18. Suppose \( \hat{\alpha} \) is estimation of \( \alpha \) in practice, is doubly robust even if the behavior policy is unknown. More specifically, we have the following

\[
\hat{J}_{\alpha} = \frac{1}{|I_k|} \sum_{i \in I_k} \sum_{t=1}^{H} \left[ \hat{w}_t^{(k)}(S_t^{(i)}) \eta_h(S_t^{(i)}, A_t^{(i)}) \{ R_t^{(i)} + \hat{\pi}_t^{(k)}(S_{t+1}^{(i)}) - \hat{\pi}_t^{(k)}(S_t^{(i)}) \} \right] + \mathbb{E}_{\hat{p}_k^{(1)}(s_1)}[\hat{v}_0^{(k)}(s_1)],
\]

given nuisance estimators \( \hat{v}_t^{(k)}(s), \hat{\pi}_t^{(k)}(s) \). We denote this estimator as \( \hat{J}_{V2} \). In the finite horizon case, the corresponding estimator \( \hat{J}_{V1} \) is constructed by

\[
\hat{J}_k = \frac{1}{|I_k|} \sum_{i \in I_k} \sum_{t=1}^{H} \left[ \hat{w}_t^{(k)}(S_t^{(i)}) \eta_h(S_t^{(i)}, A_t^{(i)}) \{ R_t^{(i)} + \hat{v}_t^{(k)}(S_{t+1}^{(i)}) - \hat{v}_t^{(k)}(S_t^{(i)}) \} \right] + \mathbb{E}_{\hat{p}_k^{(1)}(s_1)}[\hat{v}_0^{(k)}(s_1)],
\]

given nuisance estimators \( \hat{v}_t^{(k)}(s), \hat{\pi}_t^{(k)}(s) \). Importantly, these estimators allow direct estimation of state value functions. In this sense, it is still different from \( \hat{J}_{T11}, \hat{J}_{T12} \) since the construction of \( \hat{v}(s) \) is restricted to \( \hat{v}(s) = \int u(a)\hat{\pi}(a|s)\hat{q}(s,a)da \). The estimator \( \hat{J}_{V2} \) is doubly robust if the behavior policy is known in the sense that it is consistent as long as the either model of \( w(s) \) and \( q(s,a) \) is correct.

**Theorem 16 (Double Robustness).** Suppose \( \forall k \leq K, \) for some \( w^*, v^* \), \( \| \hat{w}^*(k)(s) - w^*(s) \|_2 = o_p(1), \| \hat{v}^*(k)(s) - v^*(s) \|_2 = o_p(1) \). Then, as long as \( w^* = w^* \) or \( v^* = v \), \( \hat{J}_{V2} \xrightarrow{p} J \).

In addition, the asymptotic MSE is calculated under nonparametric rate conditions if the behavior policy is known.

**Theorem 17 (Asymptotic Results).** Suppose \( \forall k \leq K, \) \( \| \hat{w}^*(k)(s) - w^*(s) \|_2 = o_p(1), \| \hat{v}^*(k)(s) - v^*(s) \|_2 = \beta \), where \( \alpha_1 \beta = o_p(n^{-1/2}) \). Then, \( \sqrt{n}(\hat{J}_{V2} - J) \xrightarrow{d} N(0, \Sigma_{T12}) \).

Here, we use a fact that the bias term \( \hat{J}_{V2} - J = \alpha_1 \beta \). Note that these theoretical properties are new results though the estimator itself was proposed in Tang et al. (2020).

The asymptotic MSE of \( \hat{J}_{V2} \) in Theorem 17 is larger than the efficiency bound in the pre-specified evaluation policy case:

\[
\mathbb{E}[\mu^2(s,a)\var\{r + \gamma v(s')\}|s, a].
\]

The difference is

\[
\mathbb{E}[w^2(s,a)\var\eta(s,a)\{r + \gamma v(s')\}|s].
\]

Therefore, \( \hat{J}_{V2} \) is not efficient in the pre-specified case. In addition, the doubly robustness property does not hold if the behavior policy is unknown. On the other hand, the estimator \( \hat{J}_{Q2} \) in Kallus and Uehara (2019b), which is defined on the basis of \( \hat{J}_k \):

\[
\hat{J}_k = \frac{1}{|I_k|} \sum_{i \in I_k} \left[ \hat{w}^*(k)(S^{(i)}) \hat{\pi}^*(k)(S^{(i)}, A^{(i)}) \{ R^{(i)} + \gamma \hat{\pi}^*(k)(S^{(i)}, A^{(i)}) - \hat{\pi}^*(k)(S^{(i)}, A^{(i)}) \} \right] + (1 - \gamma)\mathbb{E}_{\hat{p}_k^{(1)}(s_1)}[\hat{v}^{(k)}(s_1)]
\]

is doubly robust even if the behavior policy is unknown. More specifically, we have the following theorem. Refer to Kallus and Uehara (2019b, Theorem 13).

**Theorem 18.** Suppose \( \forall k \leq K, \) for some \( w^{*1}, \pi^{*1}, v^1 \), \( \| \hat{w}^*(k)(s) - w^{*1}(s) \|_2 = o_p(1), \| \hat{\pi}^*(k)(s) - \pi^{*1}(s) \|_2 = o_p(1), \| \hat{v}^*(k)(s) - v^1(s) \| = o_p(1) \). Then, as long as \( w^{*1} = w, \pi^{*1} = \pi^*, v^1 = v \), \( \hat{J}_{Q2} \xrightarrow{p} J \).

The same theorem does not hold for \( \hat{J}_{V2} \). In theory, \( \hat{J}_{Q2} \) appears to be superior to \( \hat{J}_{V2} \). However, in practice, \( \hat{J}_{V2} \) might be superior especially when the action is high-dimensional. This is because estimation of \( \hat{v}(s) \) is easier than estimation of \( \hat{q}(s,a) \).
F Proofs

Throughout this paper, we define

\[ c(s) = \frac{1}{\int u(a)\pi^b(a|s)da}. \]

Then, \( \pi^c \) is equal to \( u(a)\pi^b(a|s)c(s) \).

F.1 Warm-up

As a warm-up, we prove the results for the bandit case and NMDP case. In a bandit setting, we drop an index \( t \).

Proof of Remark 3. The entire regular parametric submodel is

\[ \{ p_\theta(s)p_\theta(a | s)p_\theta(r|s,a) \}, \]

where it matches with a true pdf at \( \theta = 0 \). The score functions of the nonparametric model is decomposed as

\[ c(s) = \int u(a) \pi^b(a | s) c(s) da. \]

We calculate the gradient of the target functional \( J(\pi^c) \) w.r.t. the nonparametric model. Since

\[ J(\pi^c) = \int r p_\theta(r | a, s) \pi^c(a | s; \pi^b(s)) p_\theta(s) d(a, s, r), \]

we have

\[ \nabla J(\pi^c) = E \left[ \pi^c(A | S) \pi^b(A | S) \{ R - q(S, A) \} g(J) + v(S) g(J) + R \nabla_\theta \pi^c(A | S; \pi^b) \right]. \]

Especially, the third term is

\[ E \left[ R \nabla_\theta \pi^c(A | S; \pi^b) \right] \]

\[ = E \left[ \left\{ E[R | S, A] \frac{u(A)}{c(S)} - E \left[ \frac{u(A)}{c(S)} \right] \frac{u(A)}{c(S)} \right\} g(A) \right] \]

\[ = E \left[ \left\{ E[R | S, A] \frac{u(A)}{c(S)} - E \left[ \frac{R u(A)}{c(S)} \right] \right\} g(J) \right] \]

\[ = E \left[ \frac{\pi^c(A | S)}{\pi^b(A | S)} \{ q(S, A) - v(S) \} \right]. \]

Then, we have

\[ \nabla J(\pi^c) = E \left[ \left\{ \frac{\pi^c(A | S)}{\pi^c(A | S)} (R - v(S)) + v(S) \right\} g(J) \right]. \]

Since the gradient is unique for the current case, this concludes that the following is the EIF:

\[ \frac{\pi^c(a | s)}{\pi^c(a | s)} (r - v(s)) + v(s). \]
We calculate the gradient of the target functional
where it matches with the true pdf at
we have
Then, we have
where it matches with the true pdf at
This is proved by
submodel corresponding an NMDP is
Calculation of derivatives under the nonparamtic NMDP.
Proof of Theorem 14.
Since the gradient is unique for the current case, this concludes that the following is the EIF:
Proof of Remark 5.
The proof is also mentioned in Muñoz and Van Der Laan (2012). We add the proof here since this would improve the reader’s understanding in the RL case. The entire regular parametric submodel is
where it matches with the true pdf at \( \theta = 0 \). The score functions of the nonparametric model is decomposed as
We calculate the gradient of the target functional \( J(\pi^e) \) w.r.t. the nonparametric model. Since
we have
\[
\nabla_{\theta} J(\pi^e) \equiv \mathbb{E} \left[ \left\{ \pi^e(\tau(S,A)|S) \right\} \{ R - q(S,A) \} + \mathbb{E}[R | S, \tau(S,A)] \right] g(J) \cdot
\]
Since the gradient is unique for the current case, this concludes that the following is the EIF:
\[
\frac{\pi^e(a|s)}{\pi^e(a|s)} \{ r - q(s,a) \} + q(s, \tau(s,a)).
\]

Proof of Theorem 14
Calculation of derivatives under the nonparametric NMDP. The entire regular parametric submodel corresponding an NMDP is
where it matches with the true pdf at \( \theta = 0 \). The score function of the model is decomposed as
\[
g(J_{rH}) = \nabla \log p_\theta(s_k|h_{a_{k-1}}) + \sum_{k=1}^{H} \nabla \log p_\theta(a_k|h_{s_k}) + \sum_{k=1}^{H} \nabla \log p_\theta(r_k|h_{a_k})
\]
Then, we have
\[
\nabla_{\theta} \mathbb{E}_{\pi^e} \left[ \sum_{t=1}^{H} r_t \right] = \mathbb{E}[\{ -J + \sum_{c=1}^{H} \lambda_c \{ r_c - v_c(\mathcal{H}_{s_c}) \} \} g(J_{rH})].
\]
This is proved by
\[
\nabla_{\theta} \mathbb{E}_{\pi^e} \left[ \sum_{t=1}^{H} r_t \right] = \nabla \theta \left[ \int \sum_{t=1}^{H} r_t \left( \prod_{k=1}^{H} p_\theta(s_{k}|h_{a_{k-1}}) \pi^e(a_{k}|h_{s_{k}}; \theta) p_\theta(r_{k}|h_{a_{k}}) \right) d\mu(h_{rH}) \right].
\]
Note that except for the third term, the proof is the same as Kallus and Uehara (2019a). The third term is calculated by

\[
\sum_{m=1}^{H} \sum_{t=1}^{H} \left\{ r_t \prod_{k=1, k \neq m}^{t} p_0(s_k|h_{a_{k-1}}) \pi^*(a_k|h_{a_k}; \theta) p_0(r_k|h_{a_k}) \right\} \times 
\pi^* \left\{ g_{A_m|h_{s_m}} - \int g_{A_m|h_{s_m}} \pi^*(a_m | h_{s_m}) d(a_m) \right\} d\mu(j_{r_H}) 
= \sum_{m=1}^{H} \mathbb{E}_{\pi^*} \left\{ \sum_{t=1}^{H} r_t g_{A_m|h_{s_m}} - \sum_{t=1}^{H} r_t \mathbb{E}_{\pi^*}[g_{A_m|h_{s_m}} | h_{s_m}] \right\} 
= \sum_{m=1}^{H} \mathbb{E}_{\pi^*} \left[ \sum_{t=1}^{H} r_t | H_{a_m} \right] - \mathbb{E}_{\pi^*} \left[ \sum_{t=1}^{H} r_t | H_{a_c} \right] g_{A_c|h_{c_c}}). 
\]

Then, \( \nabla_\theta \mathbb{E}_{\pi^*} \left[ \sum_{t=1}^{H} r_t \right] \) is equal to

\[
\sum_{c=1}^{H} \mathbb{E}_{\pi^*} \left[ \left\{ \mathbb{E}_{\pi^*}(r_c | s_1) - \mathbb{E}_{\pi^*}(r_c) \right\} g(J_{s_H}) \right] + \mathbb{E}_{\pi^*} \left[ \left\{ r_c - \mathbb{E}_{\pi^*}(r_c | H_{a_c}) \right\} g(J_{r_H}) \right] 
+ \mathbb{E}_{\pi^*} \left[ \left( \mathbb{E}_{\pi^*} \left[ \sum_{t=c+1}^{H} r_t | H_{a_{c+1}} \right] - \mathbb{E}_{\pi^*} \left[ \sum_{t=c+1}^{H} r_t | H_{a_c} \right] \right) g(J_{r_H}) \right] 
+ \mathbb{E}[\lambda_c \left\{ \mathbb{E}_{\pi^*} \left[ \sum_{t=c}^{H} r_t | H_{a_c} \right] - \mathbb{E}_{\pi^*} \left[ \sum_{t=c}^{H} r_t | H_{a_c} \right] g(J_{r_H}) \right\}] 
= \mathbb{E} \left[ -J + \sum_{c=1}^{H} \left\{ \lambda_c r_c - \lambda_c \mathbb{E}_{\pi^*}(r_c | H_{a_c}) + \lambda_{c-1} \mathbb{E}_{\pi^*}(r_c | H_{a_{c-1}}) \right\} g(J_{r_H}) \right] + 
+ \mathbb{E} \left[ \sum_{c=1}^{H} \left\{ \lambda_c \mathbb{E}_{\pi^*}(r_c | H_{a_c}) - \lambda_c \mathbb{E}_{\pi^*}(r_c | H_{a_c}) \right\} g(J_{r_H}) \right] 
= \mathbb{E} \left[ -J + \sum_{c=1}^{H} \lambda_c \left\{ r_c - v_c(H_{s_c}) \right\} + \lambda_{c-1} v_c(H_{s_{c-1}}) \right] g(J_{r_H})].
\]

This concludes that the following is a derivative:

\[
-J + \sum_{c=1}^{H} \lambda_c \left\{ r_c - v_c(H_{s_c}) \right\} + \lambda_{c-1} v_c(H_{s_{c-1}}). \quad (13)
\]

**Projection onto the tangent space** Then, based on the argument [Tsiatis, 2006], we need to project \( \phi \) onto the tangent space spanned by the nonparametric model derived by an NMDP. Writing
the gradient \( \phi \) as \( \phi \), the projection of \( \phi \) onto the tangent space is calculated as
\[
\sum_j \mathbb{E}[\phi|R_j, \mathcal{H}_{A_j}] - \mathbb{E}[\phi|\mathcal{H}_{A_j}] + \mathbb{E}[\phi|A_j, \mathcal{H}_{S_j}] - \mathbb{E}[\phi|\mathcal{H}_{S_j}] + \mathbb{E}[\phi|S_j, \mathcal{H}_{A_{j-1}}] - \mathbb{E}[\phi|\mathcal{H}_{A_{j-1}}] \\
= \{ \sum_j \mathbb{E}[\phi|R_j, \mathcal{H}_{A_j}] - \mathbb{E}[\phi|\mathcal{H}_{A_j}] \} + \mathbb{E}[\phi \mid \mathcal{H}_{A_H}] - \mathbb{E}[\phi] \\
= \{ \sum_c \lambda_c (R_c - \mathbb{E}[r_c \mid \mathcal{H}_{A_c}]) \} + \left\{ -\rho^+ + \sum_{c=1}^H \lambda_c \mathbb{E}[r_c \mid \mathcal{H}_{A_c}] - \lambda_c v_c(\mathcal{H}_{S_c}) + \lambda_{c-1} v_c(\mathcal{H}_{S_c}) \right\} + \{ 0 \} \\
= \phi.
\]
This concludes that \( \phi \) is actually the EIF.

**Calculation of the efficiency bound**  The efficiency bound is
\[
\text{var} \left[ \sum_{c=1}^H \lambda_c \{ R_c - v_c(\mathcal{H}_{S_c}) \} + \lambda_{c-1} v_c(\mathcal{H}_{S_c}) \right] \\
= \sum_{t=0}^H \mathbb{E}[\text{var}[\mathbb{E} \sum_{c=1}^H \lambda_c \{ R_c + v_{c+1}(\mathcal{H}_{S_{c+1}}) - v_c(\mathcal{H}_{S_c}) \} | \mathcal{J}_{S_{t+1}} | \mathcal{J}_{S_t}]] \\
= \sum_{t=0}^H \mathbb{E}[\lambda_t \{ R_t + v_{t+1}(\mathcal{H}_{S_{t+1}}) \} | \mathcal{J}_{S_t}]] = \sum_{t=0}^H \mathbb{E}[\lambda_t^2 \text{var}[R_t + v_{t+1}(\mathcal{H}_{S_{t+1}}) | \mathcal{H}_{S_t}]].
\]

**Proof of Theorem 15** Almost the same as the proof of Theorem 3.

**F.2 Proof of Results Related to a TMDP**

**Proof of Theorem 3**

**Calculation of derivatives under a nonparametric TMDP**  The entire regular parametric submodel is
\[
\{ p_\theta (s_1)p_\theta (a_1|s_1)p_\theta (r_1|s_1,a_1)p_\theta (s_1|s_1,a_1)p_\theta (a_1|s_1)p_\theta (r_1|s_1,a_1) \cdots p_\theta (r_H|s_H,a_H) \}.
\]
The score function of the parametric submodel is
\[
g(j_{RH}) = \sum_{k=1}^H \nabla_\theta \log p_\theta (s_k \mid s_{k-1},a_{k-1}) + \nabla_\theta \log p_\theta (a_k \mid s_k) + \nabla_\theta \log p_\theta (r_k \mid s_k,a_k) \\
= \sum_{k=1}^H g_{S_k|S_{k-1},A_{k-1}} + \sum_{k=1}^H g_{A_k|S_k} + \sum_{k=1}^H g_{R_k|S_k,A_k}.
\]
We have
\[
\nabla_\theta \mathbb{E}_{\pi^*} \sum_{t=1}^H r_t = \nabla_\theta \int \sum_{t=1}^H r_t \left\{ \prod_{k=1}^t p_\theta (s_k|a_{k-1},s_{k-1})p_\pi^*(a_k|s_k) \right\} d\mu(j_{RH}) \\
= \sum_{c=1}^H \{ \mathbb{E}_{\pi^*} [(E_{\pi^*}[r_c|s_1] - E_{\pi^*}[r_c])g_{s_1}] + \mathbb{E}_{\pi^*} [(r_c - E_{\pi^*}[r_c|s_c,a_c])g_{R_c|S_c,A_c}] \}
\]
This concludes that the third term has been calculated as the proof of Theorem 14. Then, the calculation is the same as in Theorem 2 (Kallus and Uehara, 2019a). The third term has been calculated as the proof of Theorem 14. Then, \( \nabla \theta E_{\pi^t} [\sum_{t=1}^H r_t] \) is equal to

\[
\sum_{c=1}^H \{E_{\pi^t} [(E[r_{c} | s_1] - E_{\pi^t} [r_{c}])g] + E_{\pi^t} [\mu_c(s_c, a_c) (r_c - E[r_{c} | s_c])g] \\
+ E_{\pi^t} [\mu_c(s_c, a_c) (E_1 \sum_{t=1}^H r_t | s_c, a_c) - E_2 \sum_{t=1}^H r_t | s_c] g(J_{R_{c}}) \} \]

In the end, we can conclude that the following is a derivative:

\[
-\rho^n + \sum_{c=1}^H \mu_c(r_c - v_c(s_c)) + \mu_{c-1}v_c(s_c). 
\tag{14}
\]

**Projection onto the tangent space** Then, based on the argument in Appendix B (Kallus and Uehara, 2020), we need to project it onto the tangent space spanned by the nonparametric model deduced by an MDP. Writing the gradient (14) as \( \phi \), the projection of \( \phi \) onto the tangent space is calculated as follows:

\[
\sum_{j=1}^H \mathbb{E}[\phi|R_j, S_j, A_j] - \mathbb{E}[\phi|S_j, A_j] + \mathbb{E}[\phi|A_j, S_j] - \mathbb{E}[\phi|S_j] + \mathbb{E}[\phi|S_j, A_{j-1}, S_{j-1}] - \mathbb{E}[\phi|A_{j-1}, S_{j-1}] \\
= \{ \sum_{j=1}^H \mu_j(R_j - \mathbb{E}[R_j | S_j, A_j]) \} \\
+ \mathbb{E}\left[ \sum_{c=j}^H \mu_c \{R_c - v_c(S_c) + v_{c+1}(S_{c+1}) \} | A_j, S_j \right] - \mathbb{E}\left[ \sum_{c=j}^H \mu_c \{R_c - v_c(S_c) + v_{c+1}(S_{c+1}) \} | S_j \right] \\
+ \mathbb{E}\left[ \sum_{c=j}^H \mu_c \{R_c - v_c(S_c) + v_{c+1}(S_{c+1}) \} | S_{j-1} \right] - \mathbb{E}\left[ \sum_{c=j}^H \mu_c \{R_c - v_c(S_c) + v_{c+1}(S_{c+1}) \} | S_{j-1}, A_{j-1} \right] \\
+ \mu_{j-1}v_j(S_j) - \mathbb{E}[\mu_{j-1}v_j(S_j) | S_{j-1}, A_{j-1}] \\
= \{ \sum_{j=1}^H \mu_j(R_j - \mathbb{E}[R_j | S_j, A_j]) \} + \mu_j \mathbb{E}[R_j | S_j, A_j] - v_j(S_j) + \mathbb{E}[v_{j+1}(S_{j+1}) | S_j, A_j] \\
+ \mu_{j-1}v_j(S_j) - \mathbb{E}[\mu_{j-1}v_j(S_j) | S_{j-1}, A_{j-1}] \\
= -\rho^n + v_1(s_1) + \sum_{j=1}^H \mu_j(S_j, A_j) \{R_j - v_j(S_j) + v_{j+1}(S_{j+1}) \} = \phi.
\]

This concludes that \( \phi \) is actually the EIF.
Efficiency bound  The efficiency bound is
\[
\text{var}\left[ \sum_{c=1}^{H} \mu_c \{ R_c - v_c(S_c) \} + \mu_{c-1} v_c(S_c) \right] = \sum_{t=0}^{H} \mathbb{E}[\text{var}[\mathbb{E}\{ \sum_{c=1}^{H} \mu_c \{ R_c + v_{c+1}(S_{c+1}) - v_c(S_c) \} | J_{S_{t+1}} ] | J_{S_t}] = \sum_{t=0}^{H} \mathbb{E}[\mu_t \{ R_t + v_{t+1}(S_{t+1}) \} | J_{S_t}] = \sum_{t=0}^{H} \mathbb{E}[u_t^2(S_t) \text{var}[\eta_t(S_t, A_t)\{ R_t + v_{t+1}(S_{t+1}) \} | S_t]].
\]

Order of the efficiency bounds  Since we have
\[
\sum_{c=1}^{H} r_c = \sum_{c=1}^{H} \{ r_c + v_c(s_{c+1}) - v_c(s_c) \},
\]
we have
\[
\text{var}_s[\sum_{c=1}^{H} r_c] = \text{var}_s[\sum_{c=1}^{H} \{ r_c + v_c(s_{c+1}) - v_c(s_c) \}] = \sum_{t=0}^{H} \text{var}_s[\mu_t \{ R_t + v_t(s_{t+1}) | s_t \}].
\]
Then, we can conclude that
\[
\sum_{t=0}^{H} \text{var}_s[\mu_t \{ R_t + v_t(s_{t+1}) | s_t \}] = (R_{\text{max}} H)^2.
\]

Finally, from importance sampling,
\[
\sum_{t=0}^{H} \text{var}_s[w_t^2 \text{var}_s[\lambda_t \{ r_t + v_t(s_{t+1}) \} | s_t]] = \sum_{t=0}^{H} \text{var}_s[w_t^2 \text{var}_s[\lambda_t^2 \{ r_t + v_t(s_{t+1}) - v_t(s_t) \}^2 | s_t]] \leq CC' \sum_{t=0}^{H} \text{var}_s[\{ r_t + v_t(s_{t+1}) - v_t(s_t) \}^2] = CC' \sum_{t=0}^{H} \text{var}_s[\{ \sum_{t=1}^{H} r_t - J \}^2] \leq CC'(R_{\text{max}} H)^2.
\]

Remark 10. Noting
\[
\mathbb{E}[\text{var}[f(Z)|X]] = \mathbb{E}[\text{var}[f(Z)|X,Y]|X] + \mathbb{E}[\text{var}[f(Z)|X,Y]|X],
\]
for random variables \(X,Y,Z\), the difference of this efficiency bound and the efficiency bound of the pre-specified evaluation policies is
\[
\sum_{t=0}^{H} \mathbb{E}[u_t^2(S_t) \text{var}[\eta_t(S_t, A_t)\eta_t(S_t, A_t) | S_t]].
\]

Proof of Theorem \(3\)

EIF and efficiency bound under a TMDP  The entire regular parametric submodel is
\[
\{ p_0(s_1)p_0(a_1|s_1)p_0(r_1|s_1, a_1)p_0(s_1|s_1, a_1)p_0(a_1|s_1)p_0(r_1|s_1, a_1) \cdots p_0(r_H|s_H, a_H) \},
\]
where it matches with the true pdf at \(\theta = 0\). The score function of the parametric submodel is
\[
g(j_{rh}) = \sum_{k=1}^{H} \nabla_\theta \log p_0(s_k | s_{k-1}, a_{k-1}) + \nabla_\theta \log p_0(a_k | s_{k-1}) + \nabla_\theta \log p_0(r_k | s_k, a_k)
\]
The target functional is

\[
\int \sum_{k=1}^{H} r_k \left\{ \prod_{k=1}^{t} p_{\theta}(s_k | \tau(a_k-1, s_{k-1}), s_{k-1} \pi^b_k(a_k | s_k ; \theta) p_{\theta}(r_k | \tau(a_k, s_k), s_k) \right\} d\mu(j_{TH})
\]

\[
= \sum_{c=1}^{H} \left[ \mathbb{E}_{\pi^c} \left[ \sum_{t=c+1}^{H} r_t | s_c, \tau(s_c, a_c) \right] - \mathbb{E}_{\pi^c} \left[ \sum_{t=c+1}^{H} r_t | a_c, s_c \right] \right] g_{A_c | S_c}
\]

where \( \pi^c_k(a_k | s_k) \) is \( \pi^b_k(\tilde{\tau}(s, a) \mid s) \tilde{\tau}'(s, a) \), \( \tilde{\tau}_k(\cdot, s) \) is the inverse function of \( \tau_k(\cdot, s) \). Here, we use a change of variables: \( \tau(a_k-1, s_{k-1}) = u_{k-1} \), and write \( u_k \) as \( a_k \). We have

\[
\nabla_{\theta} \mathbb{E}_{\pi^c} \left[ \sum_{t=1}^{H} r_t | s, \tau(s, a) \right] = \nabla_{\theta} \int \sum_{k=1}^{H} r_k \left\{ \prod_{k=1}^{t} p_{\theta}(s_k | \tau(a_k-1, s_{k-1}), s_{k-1} \pi^b_k(a_k | s_k ; \theta) p_{\theta}(r_k | a_k, s_k) \right\} d\mu(j_{TH})
\]

Except for the third line, the proof is almost the same as that of Kallus and Uehara (2019a, Theorem 2). The third line is proved by

\[
\sum_{c=1}^{H} \int \sum_{t=1}^{H} r_k \left\{ \prod_{k \neq c}^{t} p(a_k | a_{k-1}, s_{k-1}) p_{\pi^b_c}(a_k | s_k) p(r_k | a_k, s_k) \right\} p(s_c | \tau(a_c-1, s_{c-1}), s_{c-1})
\]

\[
\times \nabla_{\theta^b}(a_c | s_c ; \theta) p(r_c | \tau(a_c, s_c), s_c) d\mu(j_{TH})
\]

\[
= \sum_{c=1}^{H} \int \sum_{t=1}^{H} r_k \left\{ \prod_{k \neq c}^{t} p(a_k | a_{k-1}, s_{k-1}) p_{\pi^b_c}(a_k | s_k) p(r_k | a_k, s_k) \right\} p(s_c | \tau(a_c-1, s_{c-1}), s_{c-1})
\]

\[
\times \pi^b(c | s_c) g_{A_c | S_c} d\mu(j_{TH})
\]

\[
= \mathbb{E}_{\pi^b} \left[ \prod_{k=1}^{t-1} q_k \mathbb{E}_{\pi^c} \left[ \sum_{c=t}^{H} r_t | s_c, \tau(s_c, a_c) \right] g_{A_c | S_c} \right]
\]

\[
= \mathbb{E}_{\pi^b} \left[ \prod_{k=1}^{t-1} q_k \left( \mathbb{E}_{\pi^c} \left[ \sum_{c=t+1}^{H} r_t | s_c, \tau(s_c, a_c) \right] - \mathbb{E}_{\pi^c} \left[ \sum_{c=t}^{H} r_t | a_c, s_c \right] \right) g_{A_c | S_c} \right] .
\]

Then,

\[
\nabla_{\theta} \mathbb{E}_{\pi^c} \left[ \sum_{t=1}^{H} r_t \right] = \sum_{c=1}^{H} \left\{ \mathbb{E}_{\pi^b} \left[ (\mathbb{E}_{\pi^c}[r_c | s_1] - \mathbb{E}_{\pi^c}[r_c])g \right] + \mathbb{E}_{\pi^b} \left[ \mu_c(s_c, a_c)(r_c - \mathbb{E}[r_c | s_c, a_c])g \right] \right\}
\]

\[
+ \mathbb{E}_{\pi^b} \left[ \mu_c(\mathbb{E}_{\pi^c}[\sum_{t=1}^{H} r_t | s_{c+1}] - \mathbb{E}_{\pi^c}[\sum_{t=c+1}^{H} r_t | s_c, a_c])g \right]
\]
Then, as in the proof of Theorem 2 via the projection onto the tangent space, the EIF becomes

\[ \mu_{c-1} \left( \mathbb{E}_{\pi^v} \left[ \sum_{t=1}^H r_t|s_c, \tau(s_c, a_c) \right] - \mathbb{E}_{\pi^v} \left[ \sum_{t=1}^H r_t|s_c \right] \right) g \].

Here, note that we define \( \mathbb{E}_{\pi^v}[\prod_{k=1}^t \eta_k | s_t, a_t] = \mu_t \). Then, \( \nabla \theta \mathbb{E}_{\pi^v}[\sum_{t=1}^H r_t] \) is equal to

\[ \mathbb{E} \left[ -J + \sum_{c=1}^H \left\{ \mu_c \left( r_c - \mathbb{E}_{\pi_e} \left[ \sum_{t=1}^H r_t|s_c, a_c \right] \right) + \mu_{c-1} \sum_{t=1}^H \mathbb{E}_{\pi_e}[r_t|s_c] \right\} g(J_H) \right] \]

\[ + \mathbb{E} \left[ \mu_{c-1} \left( \mathbb{E}_{\pi^v}[\sum_{t=1}^H r_t|s_c, \tau(s_c, a_c)] - \mathbb{E}_{\pi^v}[\sum_{t=1}^H r_t|s_c] \right) g \] \]

\[ = \mathbb{E} \left[ -J + \sum_{c=1}^H \left\{ \mu_c \left( r_c - q_c(s_c, a_c) \right) + \mu_{c-1} q_c(s_c, \tau(a_c, s_c)) \right\} g(J_H) \right] \]

Then, as in the proof of Theorem 2 via the projection onto the tangent space, the EIF becomes

\[ -J(\pi^v) + \sum_{c=1}^H \mu_c \left( r_c - q_c(s_c, a_c) \right) + \mu_{c-1} q_c(s_c, \tau(a_c, s_c)) \].

In addition, as in the proof of Theorem 2 the efficiency bound is

\[ \sum_{c=0}^H \mathbb{E}[\mu_c^2(S_c, A_c) \var[var[R_c + q_{c+1}(s_{c+1}, \tau(A_{c+1}, S_{c+1})) | S_c, A_c]] \].

**Order of the efficiency bound** First, we observe

\[ \var[var[R_c + q_{c+1}(s_{c+1}, \tau(A_{c+1}, S_{c+1})) | S_c = s_c, A_c = a_c] = \var[var[r_c + q_{c+1}(s_{c+1}, a_{c+1}) | s_c, a_c] \].

This is proved by

\[ \var[var[R_c + q_{c+1}(s_{c+1}, \tau(A_{c+1}, S_{c+1})) | S_c = s_c, A_c = a_c] \]

\[ = \int \{ r_c + q_{c+1}(s_{c+1}, \tau(a_{c+1}, s_{c+1})) - q_c(s_c, a_c) \}^2 p(r_c | s_c, a_c) p(s_{c+1} | s_c, a_c) \pi^h(a_{c+1} | s_{c+1}) d(r_c, s_{c+1}, a_{c+1}) \]

\[ = \int \{ r_c + q_{c+1}(s_{c+1}, u_{c+1}) \}^2 p(r_c | s_c, a_c) p(s_{c+1} | s_c, a_c) \pi^h(u_{c+1} | s_{c+1}) d(r_c, s_{c+1}, u_{c+1}) \]

\[ \pi^h(u_{c+1}, s_{c+1}) | s_{c+1}) \pi^h(u_{c+1}, s_{c+1}) d(r_c, s_{c+1}, u_{c+1}) \]

\[ = \var[var[r_c + q_{c+1}(s_{c+1}, a_{c+1})] | s_c, a_c] \].

Then, we have

\[ \sum_{c=0}^H \mathbb{E}[\mu_c^2(S_c, A_c) \var[var[R_c + q_{c+1}(s_{c+1}, \tau(A_{c+1}, S_{c+1})) | S_c, A_c] \]

\[ \leq CC' \sum_{c=0}^H \mathbb{E}[\var[var[r_c + q_{c+1}(s_{c+1}, \tau(a_{c+1}, s_{c+1})) | s_c, a_c] \]

\[ = CC' \sum_{c=0}^H \mathbb{E}[\var[var[r_c + q_{c+1}(a_{c+1}, a_{c+1})] | s_c, a_c] \]

\[ = CC' \sum_{c=0}^H \mathbb{E}[\{ r_c + q_{c+1}(s_{c+1}, a_{c+1}) - q_c(s_c, a_c) \}^2] \]

25
\begin{align*}
&= CC' \mathbb{E}_{\pi_\tau}[(\sum_{c=0}^{H} r_c + q_{c+1}(s_{c+1}, a_{c+1}) - q_c(s_c, a_c))^2] = CC' \mathbb{E}_{\pi_\tau}[(\sum_{c=1}^{H} r_c)^2] = CC'R_{\max}^2H^2.
\end{align*}

From the first line to the second line, we use $\mu_c \leq CC'$. From the second line to the third line, we use (15). From the fourth line to the fifth line, we use a fact that the cross-term is equal to 0. \hfill \Box

**Remark 11.** Noting
\[
\mathbb{E}[\text{var}[f(Z)|X]] = \text{var}[\mathbb{E}[f(Z)|X,Y]|X] + \mathbb{E}[\text{var}[f(Z)|X,Y]|X],
\]
for random variables $X, Y, Z$, the difference regarding the efficiency bound between the above and that of the pre-specified evaluation policy is
\[
\sum_{t=1}^{H} \mathbb{E}[\mu_t^2(S_t, A_t) \text{var} [R_t + q_{t+1}(S_{t+1}, \tau(S_{t+1}, A_{t+1})) | S_{t+1}, R_t, S_t, A_t]]
\]
\[
= \sum_{t=1}^{H} \mathbb{E}[\mu_t^2(S_t, A_t) \text{var} [q_{t+1}(S_{t+1}, \tau(S_{t+1}, A_{t+1})) | S_{t+1}]].
\]

**Proof of Theorem 4** First, we prove
\[
\mathbb{P}_{U_t}[\phi(\hat{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)})]\mathcal{L}_1] - \mathbb{P}_{\tilde{U}_t}[\phi(w, \pi^b, q)\mathcal{L}_1]
\]
\[
= \mathcal{G}_{U_t}[\phi(\hat{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)})] - \phi(w, \pi^b, q)]
\]
\[
+ \mathbb{E}[\phi(\hat{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)})]\mathcal{L}_1] - \mathbb{E}[\phi(w, \pi^b, q)\mathcal{L}_1]
\]
\[
= o_p(n^{-1/2}).
\]

From now on, we drop an index (1) for simplicity.

**Part 1:** (17) is $o_p(1/\sqrt{n})$ First, we consider controlling the following term:
\[
\mathbb{E}\left[\sum_{t=1}^{H} \hat{w}_t(S_t)\hat{\eta}_t(S_t, A_t)\{R_t - \hat{q}_t(S_t, A_t)\} + \hat{w}_{t-1}(S_{t-1})\hat{\eta}_{t-1}(S_{t-1}, A_{t-1})\hat{v}_t(S_t)\mathcal{L}_1\right] - J.
\]

By some algebra, we have
\[
\mathbb{E}\left[\sum_{t=1}^{H} \hat{w}_t(S_t)\hat{\eta}_t(S_t, A_t)\{R_t - \hat{q}_t(S_t, A_t)\} + \hat{w}_{t-1}(S_{t-1})\hat{\eta}_{t-1}(S_{t-1}, A_{t-1})\hat{v}_t(S_t)\mathcal{L}_1\right] - J
\]
\[
= \mathbb{E}\left[\sum_{t=1}^{H} \hat{w}_t(S_t)\hat{\eta}_t(S_t, A_t)\{R_t - \hat{q}_t(S_t, A_t)\}\right]
\]
\[
+ \mathbb{E}_{\pi_\tau}[\hat{q}_t(s_t)|A_t]|\mathcal{L}_1] - J + \mathbb{E}\left[\sum_{t=1}^{H} \hat{w}_{t-1}(S_{t-1})\hat{\eta}_{t-1}(S_{t-1}, A_{t-1})\{\hat{v}_t(S_t) - \mathbb{E}_{\pi_\tau}[\hat{q}_t(s_t)|A_t]\}|\mathcal{L}_1\right]
\]
\[
= \mathbb{E}\left[\sum_{t=1}^{H} \{\hat{w}_t(S_t)\hat{\eta}_t(S_t, A_t) - w_t(S_t)\eta_t(S_t, A_t)\}\{\hat{q}_t(S_t, A_t) - q_t(S_t, A_t)\}\right] + \mathbb{E}\left[\sum_{t=1}^{H} \{\hat{w}_{t-1}(S_{t-1})\hat{\eta}_{t-1}(S_{t-1}, A_{t-1}) - w_{t-1}(S_{t-1})\eta_{t-1}(S_{t-1}, A_{t-1})\}\{-\hat{v}_t(S_t) + v_t(S_t)\}|\mathcal{L}_1\right]
\]
\[
+ \mathbb{E}\left[\sum_{t=1}^{H} \mu_t(S_t, A_t)\{\hat{q}_t(S_t, A_t) + q_t(S_t, A_t)\} + \mu_{t-1}(S_{t-1}, A_{t-1})\{\hat{v}_t(S_t) - v_t(S_t)\}|\mathcal{L}_1\right]
\]
Here, the terms from (18) to (19) are the same as the decomposition in Kallus and Uehara (2019a). Since (19) and (20) are equal to 0, the above is equal to

\[
\mathbb{E}\left[\sum_{t=1}^{H} \{\hat{\mu}_t(S_t, A_t) - \mu_t(S_t, A_t)\} \{R_t - \hat{q}_t(S_t, A_t) + v_{t+1}(S_{t+1})\} \mid \mathcal{L}_1\right]
\]

(19)

\[
+ \mathbb{E}\left[\sum_{t=1}^{H} \hat{w}_{t-1}(S_{t-1})\hat{\pi}_{t-1}(S_{t-1}, A_{t-1}) \int \{\hat{\pi}_t^e(a_t|S_t) - \pi_t^e(a_t|S_t)\} \hat{q}_t(a_t, S_t) da_t \mid \mathcal{L}_1\right].
\]

(20)

Therefore, noting what we want to control \(\mathbb{E}\left[\sum_{t=1}^{H} \hat{w}_t(S_t)\hat{\pi}_t(S_t, A_t)\{R_t - \hat{v}_t(S_t, A_t)\} + \hat{v}_t(S_t) \mid \mathcal{L}_1\right] - J\), the following holds:

\[
\mathbb{E}\left[\sum_{t=1}^{H} \hat{w}_t(S_t)\hat{\pi}_t(S_t, A_t)\{R_t - \hat{v}_t(S_t, A_t)\} + \hat{v}_t(S_t) \mid \mathcal{L}_1\right] - J = \mathbb{E}\left[\sum_{t=1}^{H} \hat{w}_t(S_t)\hat{\pi}_t(S_t, A_t)\{-\hat{v}_t(S_t) + \hat{q}_t(S_t, A_t)\} + w_t(S_t) \int \{\hat{\pi}_t^e(a_t|S_t) - \pi_t^e(a_t|S_t)\} \hat{q}_t(a_t, S_t) da_t \mid \mathcal{L}_1\right]
\]

+ \alpha_1 \beta + \alpha_2 \beta + (\alpha_1 + \alpha_2) \alpha_2 + \mathbb{E}\left[\sum_{t=1}^{H} w_t(S_t) \int \{\hat{\pi}_t^e(a_t|S_t) - \pi_t^e(a_t|S_t)\} \hat{q}_t(a_t, S_t) da_t \mid \mathcal{L}_1\right].

Here, we use \(\|\hat{\pi}_t - v_t\|_2 \leq C\|\hat{q}_t - q_t\|_2 = \beta\). Next, we also use a fact \(\|\hat{\pi}_t(S_t) - c_t(S_t)\|_2^2 = \alpha_2^2\) since

\[
\|\hat{\pi}_t(S_t) - c_t(S_t)\|_2^2 = \int \left\{w(a_t)\hat{\pi}_t^b(a_t|s_t) - \pi_t^b(a_t|s_t)u(a_t)\right\}^2 \pi_t^b(a_t|s_t) p_{\pi_t^b}(s_t) d\mu(a_t, s_t)
\]

\[
\lesssim \|\hat{\pi}_t^b(A_t|S_t) - \pi_t^b(A_t|S_t)\|_2^2.
\]

Then, the main term in the above is further expanded as follows.

\[
\mathbb{E}\left[\sum_{t=1}^{H} w_t(S_t)\hat{\pi}_t(S_t, A_t)\{-\hat{v}_t(S_t) + \hat{q}_t(S_t, A_t)\} \mid \mathcal{L}_1\right] + \mathbb{E}\left[\sum_{t=1}^{H} w_t(S_t) \int \{\hat{\pi}_t^e(a_t|S_t) - \pi_t^e(a_t|S_t)\} \hat{q}_t(a_t, S_t) da_t \mid \mathcal{L}_1\right]
\]

\[
= \mathbb{E}\left[\sum_{t=1}^{H} w_t(S_t)\hat{\pi}_t(S_t, A_t)\{-\hat{v}_t(S_t) + \hat{q}_t(S_t, A_t)\} \mid \mathcal{L}_1\right] + \mathbb{E}\left[\sum_{t=1}^{H} w_t(S_t) \int \{\hat{\pi}_t^e(a_t|S_t) - \pi_t^e(a_t|S_t)\} \hat{q}_t(a_t, S_t) da_t \mid \mathcal{L}_1\right]
\]

+ \mathbb{E}\left[\sum_{t=1}^{H} w_t(S_t) \int u_t(a_t)\hat{\pi}_t^b(a_t|S_t) - c_t(S_t) \pi_t^b(a_t|S_t) \mid \hat{q}_t(a_t, S_t) da_t \mid \mathcal{L}_1\right].

27
This concludes that (17) is proved as in the Part 1 using the same decomposition. Then, CLT concludes the proof.

Noting we have $1/c_i(S_t) = \int u_i(g_t)\pi^b_i(g_t|S_t)dg_t$, the following holds:

$$E \left[ \sum_{t=1}^{H} w_t(S_t) \int \hat{c}_t(S_t)u_t(g_t)\pi^b_t(g_t|S_t)\{ -\hat{c}_t(S_t) \int u_t(a_i)\hat{q}_t(a_i, S_t)\hat{\pi}_t(a_i|S_t)d\alpha_i + \hat{q}_t(g_t, S_t) \}dg_t|\mathcal{L}_1 \right] +$$

$$+ E \left[ \sum_{t=1}^{H} w_t(S_t) \int u_t(a_i)\{ -\hat{c}_t^2(S_t)/c_t(S_t) - \hat{c}_t(S_t) - c_t(S_t) \} \int u_t(a_i)\hat{\pi}_t^b(a_i|S_t)\hat{q}_t(a_i, S_t)d\alpha_i|\mathcal{L}_1 \right]$$

$$= E \left[ \sum_{t=1}^{H} w_t(S_t) \int \{ \hat{c}_t(S_t) - c_t(S_t) \}u_t(g_t)\{ \pi^b_t(g_t|S_t) - \hat{\pi}_t^b(g_t|S_t) \}\hat{q}_t(g_t, S_t)d\alpha_i|\mathcal{L}_1 \right] +$$

$$+ E \left[ \sum_{t=1}^{H} w_t(S_t) \{ -\hat{c}_t^2(S_t)/c_t(S_t) + \hat{c}_t(S_t) + \hat{c}_t(S_t) - c_t(S_t) \} \int u_t(a_i)\hat{\pi}_t^b(a_i|S_t)\hat{q}_t(a_i, S_t)d\alpha_i|\mathcal{L}_1 \right]$$

$$= E \left[ \sum_{t=1}^{H} w_t(S_t) \int \{ \hat{c}_t(S_t) - c_t(S_t) \}u_t(g_t)\{ \pi^b_t(g_t|S_t) - \hat{\pi}_t^b(g_t|S_t) \}\hat{q}_t(g_t, S_t)d\alpha_i|\mathcal{L}_1 \right] +$$

$$+ E \left[ \sum_{t=1}^{H} w_t(S_t) \{ -\hat{c}_t^2(S_t)/c_t(S_t) + \hat{c}_t(S_t) + \hat{c}_t(S_t) - c_t(S_t) \} \int u_t(a_i)\hat{\pi}_t^b(a_i|S_t)\hat{q}_t(a_i, S_t)d\alpha_i|\mathcal{L}_1 \right]$$

$$\leq \|\pi^b_i(A_i|S_t) - \hat{\pi}_t^b(A_i|S_t)\|_2 \leq \alpha_2^2.$$

This concludes that (17) is $\alpha_1\beta + \alpha_2\beta + \alpha_1\alpha_2 + \alpha_2^2$. Under the assumption for the convergence rates, this is equal to $o_p(n^{-1/2})$.

**Part 2:** Following Kallus and Uehara (2019b), this is proved if the following is proved:

$$E[\{\phi(\hat{w}^{(1)}, \hat{\pi}^{b(1)}), \hat{q}^{(1)}) - \phi(w, \pi^{b(1)}, q^{(1)})\}^2]\mathcal{L}_1] = o_p(1).$$

This is proved as in the Part 1 using the same decomposition.

**Final part**

$$\mathbb{P}_{\hat{\mathcal{U}}}[\phi(\hat{w}^{(1)}, \hat{\pi}^{b(1)}), \hat{q}^{(1)})]|\mathcal{L}_1] + E_{\hat{\mathcal{U}}}E[\phi(w, \pi^{b(1)}, q^{(1)})]|\mathcal{L}_2]$$

$$= \mathbb{P}[\phi(w, \pi^b, q)] + o_p(n^{-1/2}).$$

Then, CLT concludes the proof.

**Proof of Theorem** As we have ever seen,

$$\hat{J}_{T_1} - \mathbb{P}n[\phi(w^t, \pi^b, q^t)] = \alpha_1\beta + \alpha_2\beta + \alpha_1\alpha_2 + \alpha_2^2 + o_p(n^{-1/2}).$$

Under the assumption, the above is equal to $o_p(1)$. In addition, the mean of $\mathbb{P}n[\phi(w^t, \pi^b, q^t)]$ is $J$. Therefore, the statement holds from the law of large numbers.

\[\square\]
Proof of Theorem 6. First, we prove

\[
P_{\hat{\mu}_1}[\phi(\tilde{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)}) | \mathcal{L}_1] - P_{\mu_1}[\phi(w, \pi^b, q) | \mathcal{L}_1] \\
= G_{\hat{\mu}_1}[\phi(\tilde{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)})] - E[\phi(w, \pi^b, q)] \\
+ E[\phi(\tilde{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)}) | \mathcal{L}_1] - E[\phi(w, \pi^b, q) | \mathcal{L}_1] \\
= o_p(n^{-1/2}),
\]  

(21)

Part 1: (22) is \(o_p(n^{-1/2})\)

\[
\begin{align*}
= E[\phi(\tilde{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)}) | \mathcal{L}_1] - E[\phi(w, \pi^b, q) | \mathcal{L}_1] \\
= E \sum_{t=1}^H \{ \hat{\mu}_t(S_t, A_t) - \mu_t(S_t, A_t) \} \{ -\hat{q}_t(S_t, A_t) + q_t(S_t, A_t) \} \\
+ E \sum_{t=1}^H \{ \hat{\mu}_{t-1}(S_{t-1}, A_{t-1}) - \mu_{t-1}(S_{t-1}, A_{t-1}) \} \{ -\hat{q}_t(S_t, \tau(S_t, A_t)) + q_t(S_t, \tau(S_t, A_t)) \} | \mathcal{L}_1 \\
+ E \sum_{t=1}^H \mu_t(S_t, A_t) \{ -\hat{q}_t(S_t, A_t) + q_t(S_t, A_t) \} + \mu_{t-1}(S_{t-1}, A_{t-1}) \{ \hat{q}_t(S_t, \tau(S_t, A_t)) - q_t(S_t, \tau(S_t, A_t)) \} | \mathcal{L}_1 \\
+ E \sum_{t=1}^H \{ \hat{\mu}_t(S_t, A_t) - \mu_t(S_t, A_t) \} \{ R_t - q_t(S_t, A_t) + q_{t+1}(S_{t+1}, \tau(S_{t+1}, A_{t+1})) \} | \mathcal{L}_1 \\
= E \sum_{t=1}^H \{ \hat{\mu}_t(S_t, A_t) - \mu_t(S_t, A_t) \} \{ -\hat{q}_t(S_t, A_t) + q_t(S_t, A_t) \} | \mathcal{L}_1 \\
+ E \sum_{t=1}^H \{ \hat{\mu}_{t-1}(S_{t-1}, A_{t-1}) - \mu_{t-1}(S_{t-1}, A_{t-1}) \} \{ -\hat{q}_t(S_t, \tau(S_t, A_t)) + q_t(S_t, \tau(S_t, A_t)) \} | \mathcal{L}_1 \\
= \sum_{t=1}^H \| \hat{\mu}_t(S_t, A_t) - \mu_t(S_t, A_t) \|_2 - \| \hat{q}_t(S_t, A_t) + q_t(S_t, A_t) \|_2 \\
+ \| \hat{\mu}_{t-1}(S_{t-1}, A_{t-1}) - \mu_{t-1}(S_{t-1}, A_{t-1}) \|_2 - \| \hat{q}_t(S_t, \tau(S_t, A_t)) + q_t(S_t, \tau(S_t, A_t)) \|_2 \\
= (\alpha_1 + \alpha_2)\beta = o_p(n^{-1/2}).
\end{align*}
\]

Here, we use

\[
E[\mu_k(S_k, A_k)f(S_k, A_k)] = E[\prod_{t=1}^k \eta_t(S_t, A_t)f(S_k, A_k)] = E[\prod_{t=1}^{k-1} \eta_t(S_t, A_t)\eta_k(S_k, A_k)f(S_k, A_k)] \\
= E[\prod_{t=1}^{k-1} \eta_t(S_t, A_t)f(S_k, \tau_k(A_k, S_k))] = E[\mu_{k-1}(S_{k-1}, A_{k-1})f(S_k, \tau_k(A_k, S_k))],
\]

and

\[
\| \hat{q}(S, \tau(S, A)) - q(S, \tau(S, A)) \|^2 = \int \{ \hat{q}(s, \tau(s, a)) - q(s, \tau(s, a)) \}^2 \pi^b(a | s)p(s)d(s, a) \\
= \int \{ \hat{q}(s, a) - q(s, a) \}^2 \pi^c(a | s)p(s)d(s, a) \\
\leq C' \| \hat{q}(S, A) - q(S, A) \|^2.
\]

29
We calculate the gradient of the target functional where it matches with the true pdf at what we need is deriving the gradient. This is done as follows:

This is proved as in the Part 1 using the same decomposition.

Final part

\[
\mathbb{P}_{\mathcal{L}_1}[\phi(\hat{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)}) | \mathcal{L}_1] = \mathbb{P}_{\mathcal{L}_2}[\phi(w, \pi^{(1)}, q^{(1)}) | \mathcal{L}_2] = \mathbb{P}_n[\phi(w, \pi^{(1)}, q)] + o_p(n^{-1/2}).
\]

Then, CLT concludes the proof.

**Proof of Theorem 7** From the proof of Theorem [6] we have

\[
\hat{J}_{T11} = \mathbb{P}_n[\phi(w^t, \pi^{ht}, q^t)] + (\alpha_1 + \alpha_2)\beta + o_p(n^{-1/2}).
\]

From the law of large numbers, noting the expectation of \(\mathbb{P}_n[\phi(w^t, \pi^{ht}, q^t)]\) is \(J\) under the assumption, the statement is concluded.

**F.3 Proof of Results Related to an MDP**

**Proof of Theorem 8** Here, the entire regular parametric submodel is

\[
\{p_\theta(a | s)p_\theta(r | a, s)p_\theta(s' | s, a)p_\theta(r | s, a)\},
\]

where it matches with the true pdf at \(\theta = 0\). Note that the marginal distribution of \(s\) is determined by the above model because it has to be a stationary distribution. The score function of the nonparametric model is decomposed as

\[
g(j) = \log p_\theta(a | s) + \log p_\theta(s' | s, a) + \log p_\theta(r | s, a) = g_{A|S} + g_{R|S,A} + g_{S'|S,A}.
\]

We calculate the gradient of the target functional \(J(\pi^e)\) w.r.t. the nonparametric model. Since

\[
\begin{align*}
J(\pi^e) &= \int \gamma_c p_c^e(g, a) \cdot E_{\pi^e}(\gamma^t r_t p_\theta(r_t | s_t, a_t) \prod_{k=1}^t \pi^e(a_k | s_k; \theta)p_\theta(s_{k+1} | s_k, a_k) p_c^e(1)(s_1) d(h_{s_{k+1}}) \\
&= \lim_{T \to \infty} \int c_T(\gamma) \sum_{t=1}^H \gamma^t r_t p_\theta(r_t | s_t, a_t) \prod_{k=1}^t \pi^e(a_k | s_k; \theta)p_\theta(s_{k+1} | s_k, a_k) \delta^{(1)}(s_1) d(h_{s_{k+1}}) \\
&= \int c_T(\gamma) \sum_{t=1}^H \mathbb{E}_{\pi^e}[^t r_t],
\end{align*}
\]

what we need is deriving the gradient \(\phi\) satisfying

\[
\nabla J(\pi^e) = \mathbb{E}_{\pi^e_{\infty}}[\phi(s, a, r, s') g(s, a, r, s')].
\]

This is done as follows: \(\nabla J(\pi^e) =

\[
\lim_{T \to \infty} \int c_T(\gamma) \sum_{t=1}^H \gamma^t r_t \nabla_\theta p_\theta(r_t | s_t, a_t) \prod_{k=1}^t \pi^e(a_k | s_k; \theta)p_\theta(s_{k+1} | s_k, a_k) \delta^{(1)}(s_1) d(h_{s_{k+1}})
\]

Part 2: \([21]\) is \(o_p(n^{-1/2})\) Following [Kallus and Uehara (2019a)], this is proved if the following is proved:

\[
\mathbb{E}[(\phi(\hat{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)}) - \phi(w, \pi^{(1)}, q^{(1)}))^2 | \mathcal{L}_1] = o_p(1).
\]

This is proved as in the Part 1 using the same decomposition.
Therefore, the EIF is

\[
+ \lim_{T \to \infty} \int c_T(\gamma) \sum_{t=1}^{H} \mathbb{E} \mathbb{E} \left[ \gamma t r_t p_t(r_t | s_t, a_t) \left\{ \prod_{k=1}^{t} \pi^\varepsilon(a_k | s_k; \theta) \nabla p_a(s_{k+1} | s_k, a_k) \right\} p^{(1)}_t(s_t) d\eta(s_{t+1}) \right] \\
+ \lim_{T \to \infty} \int c_T(\gamma) \sum_{t=1}^{H} \mathbb{E} \mathbb{E} \left[ \gamma t r_t p_t(r_t | s_t, a_t) \left\{ \prod_{k=1}^{t} \nabla \pi^\varepsilon(a_k | s_k; \theta) p_a(s_{k+1} | s_k, a_k) \right\} p^{(1)}_t(s_t) d\eta(s_{t+1}) \right] 
\]

**First term** This is equal to

\[
\lim_{T \to \infty} \int c_T(\gamma) \sum_{t=1}^{H} \mathbb{E} \mathbb{E} \left[ \gamma t r_t g_R|S,A(r_t | s_t, a_t) \right] = \mathbb{E} \mathbb{E} \left[ r g_R|S,A(r | s, a) \right] \\
= \mathbb{E} \mathbb{E} \left[ r - \mathbb{E} \mathbb{E} \left[ r | s, a \right] g_R|S,A(r | s, a) \right] = \mathbb{E} \mathbb{E} \left[ r - \mathbb{E} \mathbb{E} \left[ r | s, a \right] g(s, a, r, s') \right].
\]

**Second term** This is equal to

\[
\lim_{T \to \infty} \gamma \int c_T(\gamma) \sum_{t=1}^{H} \mathbb{E} \mathbb{E} \left[ \gamma t r_t \gamma t-1 r_t g|S,A(s_{t+1} | s_t, a_t) \right] = \gamma \mathbb{E} \mathbb{E} \left[ r g|S,A(s | s', s) \right] \\
= \gamma \mathbb{E} \mathbb{E} \left[ r (s') g|S,A(s' | s, a) \right] = \gamma \mathbb{E} \mathbb{E} \left[ r (s') g \right] g(s, a, r, s') \right].
\]

**Third term** This is equal to

\[
\lim_{T \to \infty} \int c_T(\gamma) \sum_{t=1}^{H} \mathbb{E} \mathbb{E} \left[ \gamma t r_t g|S,A(s | s', s) \right] = \gamma \mathbb{E} \mathbb{E} \left[ r g|S,A(s | s', s) \right] \\
= \mathbb{E} \mathbb{E} \left[ r + \gamma v(s' - v(s)) \right] g(s, a, r, s') \right].
\]

In summary,

\[
\nabla J(\pi^\varepsilon) = \mathbb{E} \mathbb{E} \left[ \left\{ r - \mathbb{E} \mathbb{E} \left[ r | s, a \right] + \gamma v(s') - \gamma \mathbb{E} \mathbb{E} \left[ r | s', a \right] + q(s, a) - v(s) \right\} g(s, a, r, s') \right] \\
= \mathbb{E} \mathbb{E} \left[ r + \gamma v(s' - v(s)) g(s, a, r, s') \right] \\
= \mathbb{E} \mathbb{E} \left[ \mu^*(s, a) \left\{ r + \gamma v(s' - v(s)) - J \right\} g(s, a, r, s') \right].
\]

Therefore, the EIF is

\[
\mu^*(s, a) \left\{ r + \gamma v(s') - v(s) \right\} - J.
\]

The efficiency bound is

\[
\mathbb{E} [w^2(S) \text{var}[\eta(S, A) | R + \gamma v(S')] | S].
\]

**Proof of Theorem 9** This is similarly proved as in Theorem 8 by some calculation.

**Proof of Theorem 10** Here, we have

\[
P_{\mathcal{U}_1} \left[ \phi(w, a, q) \right] \mathcal{L}_1 = P_{\mathcal{U}_1} \left[ \phi(w, a, q) \right] \mathcal{L}_1 \\
= \mathbb{E} \mathbb{E} \left[ \phi(w, a, q) \right] \mathcal{L}_1
\]

(23)
Then, the proof is immediately concluded from CLT. For the rest of the proof, we prove (24) is $o_p(n^{-1/2})$. The part (23) is $o_p(n^{-1/2})$ is similarly proved from the same decomposition. We have

$$E[\hat{w}'(S)\hat{\eta}(S, A)\{R - \hat{q}(S, A) + \gamma\hat{v}(S')\}] + E_{p^{(1)}}[\hat{v}(S)] - J$$

$$= E[\hat{w}'(S)\hat{\eta}(S, A) - \hat{w}'(S)\eta(S, A)]\{R - \hat{q}(S, A) + \gamma\hat{v}(S')\} + (1 - \gamma)E_{p^{(1)}}[\hat{v}(s_1) - v(s_1)]$$

$$+ E[\hat{w}'(S)\hat{\eta}(S, A)\{q(S, A) - \hat{q}(S, A)\} + \hat{w}'(S)\gamma\hat{v}(S') + (1 - \gamma)E_{p^{(1)}}[\hat{v}(s_1)]$$

$$+ \|\hat{q}(S, A) - \hat{q}(S, A)\|_2\|\hat{\mu}(S, A) - \hat{\mu}(S, A)\|_2$$

$$\leq E[\hat{w}'(S)\eta(S, A)\{\hat{q}(S, A) + \gamma\hat{v}(S')\}] + (1 - \gamma)E_{p^{(1)}}[\hat{v}(s_1)] + \alpha_1\alpha_2 + \alpha_1\beta + \alpha_2\beta + \alpha_2^2.$$
\[ \|\hat{\pi}(A|S) - \pi(A|S)\|_2^2 = \alpha_2^2. \]

In summary,
\[
\mathbb{E}[\phi(\hat{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)})|L_1] - \mathbb{E}[\phi(w^*, \pi^b, q)|L_1] = \alpha_1 \alpha_2 + \alpha_1 \beta + \alpha_2 \beta + \alpha_3^2.
\]

Under the assumption regarding convergence rates, this is equal to \( o_p(n^{-1/2}) \).

**Proof of Theorem 17** This is immediately concluded from
\[
\mathbb{E}[\phi(\hat{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)})|L_1] - \mathbb{E}[\phi(w^*, \pi^b, q)|L_1] = \alpha_1 \alpha_2 + \alpha_1 \beta + \alpha_2 \beta + \alpha_3^2,
\]
which is obtained in the proof of Theorem 10.

**Proof of Theorem 12** Here, we have
\[
\mathbb{E}[\phi(\hat{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)})|L_1] - \mathbb{E}[\phi(w^*, \pi^b, q)|L_1] = o_p(n^{-1/2}).
\]

Then, the proof is immediately concluded from CLT. It remains to bound (26) is \( o_p(n^{-1/2}) \). The part that (27) is \( o_p(n^{-1/2}) \) is similarly proved from the same decomposition. We have
\[
\mathbb{E}[\hat{\nu}(S)\hat{\eta}(S, A)\{R - \hat{q}(S, A) + \gamma \hat{q}(S', \tau(A', S'))\}] + (1 - \gamma)\mathbb{E}_{\hat{p}_{(1)}}[\hat{v}(s_1)] - J
\]
\[
= \mathbb{E}[\hat{w}(S)\hat{\eta}(S, A) - w^*(S)\eta(S, A)]\{R - q(S, A) + \gamma q(S', \tau(A', S'))\}
\]
\[
+ \mathbb{E}[w^*(S)\eta(S, A)\{q(S, A) - \hat{q}(S, A)\} + w^*(S)\{\gamma \hat{q}(S, \tau(S, A)) - \gamma q(S, \tau(S, A))\}
\]
\[
+ (1 - \gamma)\mathbb{E}_{\hat{p}_{(1)}}[\hat{v}(s_1) - v(s_1)]
\]
\[
\leq \mathbb{E}[\|\hat{w}(S)\hat{\eta}(S, A) - w^*(S)\eta(S, A)\|^2] \leq \alpha_2^2 + \alpha_2 \beta + \alpha_3^2.
\]

**Proof of Theorem 7** This is immediately concluded from the proof of Theorem 12 as in Kallus and Uehara (2019a).

**F.4 Other Proofs**

**Proof of Theorem 16** Immediately, concluded from Theorem 17.

**Proof of Theorem 17** Here, we have
\[
\mathbb{E}[\phi(\hat{w}^{(1)}, \hat{\pi}^{(1)}, \hat{q}^{(1)})|L_1] - \mathbb{E}[\phi(w^*, \pi^b, q)|L_1] = o_p(n^{-1/2}).
\]

Then, the proof is immediately concluded from CLT. For the rest of the proof, we prove (28) is \( o_p(n^{-1/2}) \). The part (27) is \( o_p(n^{-1/2}) \) is similarly proved from the same decomposition. We have
\[
\mathbb{E}[\hat{w}(S)\hat{\eta}(S, A)\{R - \hat{v}(S) + \gamma \hat{v}(S')\}] + (1 - \gamma)\mathbb{E}_{\hat{p}_{(1)}}[\hat{v}(s_1)] - J
\]
We prove (30) is concluded from CLT. The proof is as follows. Here, we have

\[
E[\hat{w}^*(S)\eta(S, A) - w^*(S)\eta(S, A)] \{R - v(S) + \gamma v(S')\} + E[w^*(S)\eta(S, A)\{v(S) - \hat{v}(S)\} + w^*(S)\{\gamma \hat{v}(S) - \gamma v(S')\}] + (1 - \gamma)E_{P^s(v)}[\hat{v}(s_1) - v(s_1)] + E[\{\hat{w}^*(S)\eta(S, A) - w^*(S)\eta(S, A)\}\{v(S) - \hat{v}(S) + \gamma \hat{v}(S') - \gamma v(S')\}] \lesssim \|\hat{w}^*(S)\eta(S, A) - w^*(S)\eta(S, A)\|^2 \|\hat{v}(S) - v(S)\|^2 = \alpha_1 \beta = o_p(n^{-1/2}).
\]

Here, we use the following fact:

\[
E[\eta(S, A)\{R - v(S) + \gamma v(S')\}] = 0.
\]

**Remark 12.** In the finite horizon case, we also have a double robustness and efficiency. The argument is as follows. Here, we have

\[
P_{U_t}[\phi(\hat{w}^{(1)}, \hat{v}^{(1)})|L_1] - P_{U_t}[\phi(w, v)|L_1] = G_{U_t}[\phi(\hat{w}^{(1)}, \hat{v}^{(1)}) - \phi(w, v)] + E[\phi(\hat{w}^{(1)}, \hat{v}^{(1)})|L_1] - E[\phi(w, v)|L_1] = o_p(n^{-1/2}).
\]

We prove (30) is \(o_p(n^{-1/2})\). The fact (29) is \(o_p(n^{-1/2})\) is similarly proved. Then, the final statement is concluded from CLT. The proof is as follows:

\[
E[\sum_{t=1}^{H} \hat{w}_t(S_t)\eta(S_t, A_t)\{R_t - \hat{v}_t(S_t)\} + \hat{w}_{t-1}(S_{t-1})\eta(S_{t-1}, A_{t-1})\hat{v}_t(S_t)] = E[\sum_{t=1}^{H} \{\hat{w}_t(S_t) - w_t(S_t)\}\eta_t(S_t, A_t)\{\hat{v}_t(S_t) + v_t(S_t)\}] + E[\sum_{t=1}^{H} \{\hat{w}_{t-1}(S_{t-1}) - w_{t-1}(S_{t-1})\}\eta_{t-1}(S_{t-1}, A_{t-1})\{\hat{v}_t(S_t) + v_t(S_t)\}] + E[\sum_{t=1}^{H} w_t(S_t)\{\hat{v}_t(S_t) + v_t(S_t)\} + w_{t-1}(S_{t-1})\{\hat{v}_t(S_t) - v_t(S_t)\}] + E[\sum_{t=1}^{H} \{\hat{w}_t(S_t) - w_t(S_t)\}\eta_t(S_t, A_t)\{R_t - v_t(S_t) + v_{t+1}(S_{t+1})\}] \lesssim \sum_{t=1}^{H} \|\hat{v}_t(S_t) - v_t(S_t)\|_2 \|\hat{w}_t(S_t) - w_t(S_t)\|_2 = \alpha_1 \beta.
\]