Semiparametrically Efficient Off-Policy Evaluation in Linear Markov Decision Processes

Chuhan Xie [1] **Wenhao Yang** [2] **Zhihua Zhang** [1]

Abstract

We study semiparametrically efficient estimation in off-policy evaluation (OPE) where the underlying Markov decision process (MDP) is linear with a known feature map. We characterize the variance lower bound for regular estimators in the linear MDP setting and propose an efficient estimator whose variance achieves that lower bound. Consistency and asymptotic normality of our estimator are established under mild conditions, which merely requires the only infinite-dimensional nuisance parameter to be estimated at a $n^{-1/4}$ convergence rate. We also construct an asymptotically valid confidence interval for statistical inference and conduct simulation studies to validate our results. To our knowledge, this is the first work that concerns efficient estimation in the presence of a known structure of MDPs in the OPE literature.

1. Introduction

Off-policy evaluation (OPE) is one of the major tasks in offline reinforcement learning (Precup, 2000; Mahmood et al., 2014; Jiang & Li, 2016; Munos et al., 2016; Thomas & Brunskill, 2016; Xie et al., 2019; Uehara et al., 2022). In contrast to traditional online reinforcement learning problems, OPE focuses on estimating the expected long-term rewards of a policy using logged data that is generated in advance by a potentially different policy. OPE enjoys a wide range of applications especially when online long-term experimentation may be costly or unethical, such as in healthcare (Murphy, 2003; Chakraborty & Moodie, 2013; Luckett et al., 2019), education (Mandel et al., 2014), and recommendation systems (Chen et al., 2019).

[1]School of Mathematical Sciences, Peking University, Beijing, China [2]Academy of Advanced Interdisciplinary Studies, Peking University, Beijing, China. Correspondence to: Zhihua Zhang <zhzhang@math.pku.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

The OPE task in reality is challenging, mainly because of two reasons: 1). the behavior policy, which is used to generate logged data, is usually different from the target policy that is of our interest, and the gap between them tends to increase instability of estimation; 2). the OPE target is an accumulated long-term reward, while naive OPE approaches suffer from heavy computation as the time horizon increases. The latter is also called the "curse of horizon" (Liu et al., 2018).

In face of these challenges, there is a major line of works that aim to improve the sample efficiency of OPE estimators (Jiang & Li, 2016; Liu et al., 2018; Xie et al., 2019; Kallus & Uehara, 2020; 2022). Probably one of the most remarkable works is (Kallus & Uehara, 2022), which leverages the semiparametric theory to propose a *double reinforcement learning* (DRL) value estimator that is both doubly-robust against model misspecification of nuisance parameters and efficient (i.e., having the minimal variance) among a wide class of estimators. The proposed DRL estimator is a successful application of statistical theories to OPE problems due to its statistical nature. Recently, similar approaches have been applied to cases with more complex data generating mechanisms, such as OPE problems in Partially Observable MDPs (POMDPs) (Bennett & Kallus, 2021) and Confounded MDPs (CMDPs) (Shi et al., 2022).

Other than specific data generating mechanisms, all mentioned works implicitly assume no information on the structure of MDPs. This is not always the case, because in certain problems we may possess information that the underlying MDP, for example, is linear with a known feature map (Sutton & Barto, 2018). Therefore, a natural question arises:

> *Is there any efficiency gain from such linear structure of MDPs? If so, how to construct a "best" value estimator in the OPE problem with linear MDPs?*

In this paper we conduct a comprehensive analysis on this question, and finally give a positive answer to it. As far as we know, our work is for the first time to study efficient OPE estimation in the presence of a known structure of MDPs.

The main contributions of this paper are as follows.

- **Efficiency theory.** We use semiparametric efficiency theory to obtain the variance lower bound for regular estimators in the linear MDP setting (Theorem 3.4), which is shown to be smaller than that in the general MDP setting (Corollary 3.6). In other words, there does exist some efficiency gain brought by the linear structure, and it is possible to construct an OPE estimator with a smaller variance for linear MDPs.
- **Efficient estimation and inference.** Based on the derived theory, we construct an OPE estimator that has the minimal variance among all regular estimators in our setting. We further characterize sufficient conditions for the proposed estimator to be valid, which merely requires the only infinite-dimensional nuisance parameter to be estimated at a $n^{-1/4}$ convergence rate (Theorem 4.1 and Corollary 4.3). We also provide an asymptotically valid confidence interval for statistical inference on the target value, based on an asymptotic normality argument for our proposed estimator.
- **Numerical illustration.** In simulation studies, we demonstrate that our estimator outperforms the DRL estimator (proposed in (Kallus & Uehara, 2022)) in the linear MDP setting and that our proposed confidence interval achieves its nominal coverage rate asymptotically.

1.1. Additional Related Work

Semiparametric statistics. Semiparametric theory enjoys a long history in statistics. It considers estimation in situations when we only have partial knowledge of data, and tend to assume nothing on other data features that we do not know (Bickel et al., 1993; Tsiatis, 2006). The theory is widely applied in causal inference and missing data, e.g., (Robins et al., 1995; Robins & Rotnitzky, 1995; Bang & Robins, 2005; Schwartz et al., 2011; Ray & van der Vaart, 2020; Cui et al., 2020). Concise reviews concerning its basic concepts, techniques and applications in causal inference are given in (Kennedy, 2016; 2022).

Efficient OPE. In the OPE literature, estimation methods can be roughly categorized into three types: 1). the *direct method* (DM), which requires directly estimating a Q-function and then averaging to obtain a value estimate (Bertsekas, 2012); 2). *importance sampling* (IS), which estimates a density ratio and computes the weighted average of the reward as a value estimate (Liu, 2001; Liu et al., 2018; Xie et al., 2019); and 3). the *doubly-robust* method (DR), which combines DM and IS by adding an estimated Q-function as a control variate to improve stability and efficiency of the estimator (Jiang & Li, 2016; Kallus & Uehara, 2020; 2022). There is yet another line of works on estimation of the mentioned Q-function and density ratio, including FQI (Ernst et al., 2005; Le et al., 2019), RBM (Antos et al., 2008), DICE (Nachum et al., 2019; Zhang et al., 2020), MWL/MQL (Uehara et al., 2020) and so on.

2. Preliminaries

Infinite-horizon MDPs. We consider an infinite-horizon MDP, represented by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, R, r)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and γ is a discount factor. $P\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ represents the transition probability kernel, i.e., $P(s' \mid s, a)$ is the probability of transiting to state s' from a given state-action pair (s, a). $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ represents the expected reward, i.e., $R(s, a)$ is the expectation of the immediate reward when action a is taken in state s, and $r(s, a) \in [0, R_{\max}]$ represents the corresponding random reward. For an MDP $\mathcal{M}$, a policy $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ gives a distribution over the action space $\mathcal{A}$ for any state $s \in \mathcal{S}$. Given a policy π, the value function and the Q-function are defined as follows:

$$V_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r^{(t)} \mid s^{(0)} = s \right],$$

$$Q_\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r^{(t)} \mid s^{(0)} = s, a^{(0)} = a \right],$$

where $r^{(t)} = r(s^{(t)}, a^{(t)})$ is the immediate reward at time t, and the expectation $\mathbb{E}_\pi(\cdot)$ is taken over all randomness of a trajectory $\tau = (s^{(0)}, a^{(0)}, r^{(0)}, s^{(1)}, a^{(1)}, r^{(1)}, \dots)$ generated from the MDP by iteratively applying the policy π. The value function and the Q-function measure the expected cumulative reward of a policy.

Off-policy evaluation (OPE). We assume access to an offline dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$, consisting of n independent and identically distributed draws of state-action-reward-state quadruplets, with each s_i following some state distribution $p^{(0)}_{\pi^b}(s)$ and each a_i following a behavior policy $\pi^b(a \mid s)$[1]. Based on the offline dataset $\mathcal{D}$, the goal of off-policy evaluation is to estimate the value function of a known target/evaluation policy $\pi^e(a \mid s)$[2] (which may be different from the behavior policy $\pi^b(a \mid s)$), averaged over a user-specified initial state distribution $p^{(0)}_{\pi^e}(s)$, i.e.,

$$v_{\pi^e} = \mathbb{E}_{p^{(0)}_{\pi^e}}[V_{\pi^e}(s)] = \mathbb{E}_{\pi^e}\left[\sum_{t=0}^{\infty} \gamma^t r^{(t)} \right]. \tag{1}$$

Here, the subscripts π^b, π^e in $p^{(0)}_{\pi^b}, p^{(0)}_{\pi^e}$ are merely to distinguish the two initial distributions, which do not necessarily depend on the corresponding policies.

[1]The superscript b is an abbreviation of *behavior*.

[2]The superscript e is an abbreviation of *evaluation*.

Semiparametric theory. Here we introduce a little bit of semiparametric theory, while a detailed background is deferred to Appendix A.1. Consider an estimation problem, where we hope to estimate a one-dimensional functional $\beta(F)$ of the data distribution F. Suppose it is known that the data distribution F belongs to a model $\mathcal{F} = \{F_\theta \colon \theta \in \Theta\}$ (θ is not necessarily parametric). Now, given observed data and under some regularity conditions, an estimator $\widehat{\beta}_n$ for $\beta(F)$ is called *efficient* if $\sqrt{n}(\widehat{\beta}_n - \beta(F))$ is asymptotically normal with the minimal asymptotic variance among all regular estimators. The corresponding minimal asymptotic variance (denoted $\mathcal{V}_{\mathcal{F}}(\beta, F)$) is called the *semiparametric efficiency bound*. The efficiency bound can be regarded as a semiparametric extension of the Cramer-Rao lower bound (Shao, 2003) for unbiased estimators in parametric models. In addition, if the estimator $\widehat{\beta}_n$ is efficient, then it must be *asymptotically linear* in the sense that

$$\sqrt{n}(\widehat{\beta}_n - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\text{eff}}(O_i) + o_p(1),$$

for some function $\psi_{\text{eff}}(O)$ satisfying $\mathbb{E}[\psi_{\text{eff}}(O)^2] < \infty$ and $\mathbb{E}[\psi_{\text{eff}}(O)] = 0$, where $\{O_i\}_{i=1}^n$ is the observed data (Van der Vaart, 2000). The function $\psi_{\text{eff}}(O)$ is then called the *efficient influence function*. Note that the efficiency bound and the efficient influence function are related by $\mathcal{V}_{\mathcal{F}}(\beta, F) = \mathbb{E}[\psi_{\text{eff}}(O)^2]$.

Double Reinforcement Learning (DRL). Kallus & Uehara (2022) proposed an estimator under the infinite-horizon setting based on *double reinforcement learning* (Kallus & Uehara, 2020), which is known to achieve the efficiency bound with respect to a fully nonparametric model that will be defined later. To state their results, we now define some additional notation. Given a policy π, we further define the γ-*discounted average visitation frequency* as

$$p_{\pi,\gamma}^{(\infty)}(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t p_\pi^{(t)}(s),$$
$$p_{\pi,\gamma}^{(\infty)}(s,a) = p_{\pi,\gamma}^{(\infty)}(s) \pi(a \mid s),$$

where $p_\pi^{(t)}(s)$ is the marginal distribution of $s^{(t)}$ under policy π, starting from some initial state distribution $s^{(0)} \sim p_\pi^{(0)}$. We also define the instantaneous, state, and state-action density ratios as

$$\eta(s,a) = \frac{\pi^e(a \mid s)}{\pi^b(a \mid s)}, \quad w(s) = \frac{p_{\pi^e,\gamma}^{(\infty)}(s)}{p_{\pi^b}^{(0)}(s)},$$
$$w(s,a) = w(s)\eta(s,a) = \frac{p_{\pi^e,\gamma}^{(\infty)}(s,a)}{p_{\pi^b}(s,a)},$$

where we denote $p_{\pi^b}(s,a) = p_{\pi^b}^{(0)}(s)\pi^b(a \mid s)$ for notational simplicity. Now we are ready to state the previous results on DRL.

Theorem 2.1 (Kallus & Uehara, 2022)[3]**.** *Consider the fully nonparametric model for the data distribution $\mathcal{F}_{np} = \{p(s,a,r,s') \colon p(s,a,r,s') = p_s(s)p_{a|s}(a \mid s)p_{r|s,a}(r \mid s,a)p_{s'|s,a}(s' \mid s,a)\}$. The efficient influence function and the efficiency bound with respect to the model $\mathcal{F}_{np}$ for estimating v_{π^e} are given by*

$$\psi_{\mathit{eff,np}}(s,a,r,s') = \frac{1}{1-\gamma} w(s,a) \mathsf{br}(s,a,r,s'),$$
$$\mathcal{V}_{\mathit{np}}(v_{\pi^e}) = \frac{1}{(1-\gamma)^2} \mathbb{E}\left[w(s,a)^2 \mathsf{br}(s,a,r,s')^2\right],$$

where $\mathsf{br}(s,a,r,s') = r + \gamma V_{\pi^e}(s') - Q_{\pi^e}(s,a)$ is the Bellman residual. In addition, the following estimator is efficient with respect to $\mathcal{F}_{np}$:

$$\widehat{v}_{\mathit{DR}} = \mathbb{E}_{p_{\pi^e}^{(0)}}[\widehat{V}_{\pi^e}(s)] + \frac{1}{n(1-\gamma)} \sum_{i=1}^{n} \widehat{w}(s_i,a_i) \widehat{\mathsf{br}}_i,$$

where $\widehat{\mathsf{br}}_i = r_i + \gamma \widehat{V}_{\pi^e}(s_i') - \widehat{Q}_{\pi^e}(s_i,a_i)$ is a Bellman residual estimate, $\widehat{w}(s,a)$ and $\widehat{Q}_{\pi^e}(s,a)$ are some estimates of $w(s,a)$ and $Q_{\pi^e}(s,a)$ satisfying certain conditions, and $\widehat{V}_{\pi^e}(s)$ is defined in terms of $\widehat{Q}_{\pi^e}(s,a)$ by taking expectation over $a \sim \pi^e(\cdot \mid s)$.

A rigorous statement of Theorem 2.1 including specific conditions for $\widehat{w}(s,a)$ and $\widehat{Q}_{\pi^e}(s,a)$ is displayed in Appendix D.1. Note that the model $\mathcal{F}_{\text{np}}$ encodes no restriction on the data distribution other than the Markov property of MDPs. Theorem 2.1 gives a lower bound on the variance of any regular estimator and proposes the DRL estimator $\widehat{v}_{\text{DR}}$ that achieves this lower bound.

3. Semiparametric Efficiency in Linear MDPs

We first analyze the efficiency bound in the case of linear MDPs. Assume the transition probability P and the expected reward R possess linear structures, i.e.,

$$P(s' \mid s,a) = \phi(s,a)^\top \nu_0(s'), \quad R(s,a) = \phi(s,a)^\top \omega_0, \tag{2}$$

where $\phi = (\phi_1, \ldots, \phi_d)^\top \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a known feature map, $\nu_0 \colon \mathcal{S} \to \mathbb{R}^d$ is a vector-valued function, and $\omega_0 \in \mathbb{R}^d$ is a vector. Although the transition probability and the expected reward are set to be linear, which seems like a parametric assumption, there are still nonparametric components in the model such as $\nu_0(s')$ and the distribution of random rewards. Therefore, the class of linear MDPs is indeed a semiparametric model as discussed before.

In the MDP literature, linear structures are often assumed to lower the complexity of the model class, especially when the

[3]Note that our parameter of interest v_{π^e} defined in (1) differs from that of (Kallus & Uehara, 2022) by a $1/(1-\gamma)$ scale, so relevant quantities in this theorem are also properly scaled.

state and action spaces are too large for analysis and computation (Sutton & Barto, 2018). We assume the feature map and the true parameters satisfy the following assumptions.

Assumption 3.1. For any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, the $\phi_i(s,a) \geq 0$ and $\sum_{i=1}^d \phi_i(s,a) = 1$. In addition, $\{\phi(s,a)\}_{s\in\mathcal{S},a\in\mathcal{A}}$ spans the whole space $\mathbb{R}^d$.

Assumption 3.2. The true parameters ω_0 and ν_0 satisfy

$$\|\omega_0\|_2 \leq 1, \quad \int \nu_0(s)\mathrm{d}s = \mathbf{1}_d,$$
$$\sup_{a\in\mathcal{A}} \left\| \int \nu_0(s)\phi(s,a)^\top \mathrm{d}s \right\|_2 \leq 1.$$

Assumption 3.3. The matrices $\mathbb{E}[\phi(s,a)\phi(s,a)^\top]$ and $\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]$ are invertible, where $\Omega(s,a) = \mathrm{var}(r \mid s,a)$ is the conditional variance of the reward.

We denote $\mathcal{F}_{\text{lin}}$ as the class of linear MDPs satisfying Assumptions 3.1-3.3. Note that $\mathcal{F}_{\text{lin}} \subset \mathcal{F}_{\text{np}}$, and the true data distribution is obtained by taking $p_s(s) = p^{(0)}_{\pi^b}(s)$, $p_{a|s}(a \mid s) = \pi^b(a \mid s)$, and setting $\omega = \omega_0$, $\nu(s') = \nu_0(s')$ as in (2). For notational simplicity, we denote $\Phi_{\pi^e,\gamma} = \frac{1}{1-\gamma}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}[\phi(s,a)]$ and $p_{\pi^b}(s,a) = p^{(0)}_{\pi^b}(s)\pi^b(a \mid s)$. The following theorem gives the efficient influence function and the efficiency bound for estimating v_{π^e}, as defined in (1), with respect to the model $\mathcal{F}_{\text{lin}}$.

Theorem 3.4. *Define*

$$\psi_{\textit{eff},1}(s,a,\varepsilon) = \Phi_{\pi^e,\gamma}^\top \left\{\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1} \cdot \phi(s,a)\Omega(s,a)^{-1}\varepsilon, \tag{3}$$

$$\psi_{\textit{eff},2}(s,a,s') = \gamma\Phi_{\pi^e,\gamma}^\top \left\{V_{\pi^e}(s')I - P_\Delta(V_{\pi^e})P_\Delta(1)^{-1}\right\} \cdot \frac{\Delta(s')^{-1}\phi(s,a)}{\phi(s,a)^\top\nu_0(s')}, \tag{4}$$

where $\varepsilon = r - \phi(s,a)^\top\omega_0$, $\Omega(s,a) = \mathrm{var}(\varepsilon \mid s,a)$,

$$\Delta(s') = \int p_{\pi^b}(s,a)\frac{\phi(s,a)\phi(s,a)^\top}{\phi(s,a)^\top\nu_0(s')}\mathrm{d}s\mathrm{d}a, \tag{5}$$

$$P_\Delta(f) = \int \Delta(s')^{-1}f(s')\mathrm{d}s' \quad \forall f\colon \mathcal{S}\to\mathbb{R}. \tag{6}$$

Then, the efficient influence function for estimating v_{π^e} w.r.t. the model $\mathcal{F}_{lin}$ is given by $\psi_{\textit{eff}}(s,a,\varepsilon,s') = \psi_{\textit{eff},1}(s,a,\varepsilon) + \psi_{\textit{eff},2}(s,a,s')$, and the efficiency bound is given by $\mathcal{V}(v_{\pi^e}) = \mathrm{var}\{\psi_{\textit{eff},1}(s,a,\varepsilon)\} + \mathrm{var}\{\psi_{\textit{eff},2}(s,a,s')\}$ where

$$\mathrm{var}\{\psi_{\textit{eff},1}(s,a,\varepsilon)\} = \Phi_{\pi^e,\gamma}^\top \left\{\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\Phi_{\pi^e,\gamma}, \tag{7}$$

$$\mathrm{var}\{\psi_{\textit{eff},2}(s,a,s')\} = \gamma^2\Phi_{\pi^e,\gamma}^\top \left\{P_\Delta(V_{\pi^e}^2) - P_\Delta(V_{\pi^e})P_\Delta(1)^{-1}P_\Delta(V_{\pi^e})\right\}\Phi_{\pi^e,\gamma}. \tag{8}$$

The proof of Theorem 3.4 is deferred to Appendix A.2. Basically, the efficient influence function can be decomposed into two separated parts. The first part, $\psi_{\text{eff},1}(s,a,\varepsilon)$, results from the linear structure of the expected reward, and its form highly resembles that of restricted moment models (see e.g., Tsiatis, 2006), which also take into account the variance of noise in estimation. The weighting with the inverse variance $\Omega(s,a)^{-1}$ on the noise ε in (3) also shares the same spirit with the variance-aware idea in other OPE works (Min et al., 2021). The second part $\psi_{\text{eff},2}(s,a,s')$ results from the linear structure of the transition probability. The matrix $\Delta(s')^{-1}$ can be regarded as representing a "subspace" formed by distorting the transition probability $P(s' \mid s,a) = \phi(s,a)^\top\nu_0(s,a)$ with the feature map, and the term in the brace in (4) can be regarded as a "projection" of $V_{\pi^e}(s')$ onto such "subspace". Since the mechanisms of generating r and s' are independent given the current state-action pair (s,a), the two parts $\psi_{\text{eff},1}(s,a,\varepsilon)$ and $\psi_{\text{eff},2}(s,a,s')$ are uncorrelated conditional on (s,a).

Remark 3.5. In the tabular case where $\mathcal{S}$ and $\mathcal{A}$ are both finite sets and $d = |\mathcal{S}||\mathcal{A}|$, two models $\mathcal{F}_{\text{np}}$ and $\mathcal{F}_{\text{lin}}$ become equivalent. In Appendix A.3, we show that the efficient influence function $\psi_{\text{eff}}(s,a,\varepsilon,s')$ given in Theorem 3.4 degenerates into $\psi_{\text{eff,np}}(s,a,r,s')$ in Theorem 2.1 and so as the efficiency bounds. In this sense, our result is consistent with the previous work.

Since the nonparametric model $\mathcal{F}_{\text{np}}$ is larger than our linear model $\mathcal{F}_{\text{lin}}$, the efficiency bound of the former is larger than that of the latter for linear MDPs.

Corollary 3.6. *For linear MDPs, $\mathcal{V}(v_{\pi^e}) \leq \mathcal{V}_{np}(v_{\pi^e})$, where $\mathcal{V}(v_{\pi^e})$ is defined in Theorem 3.4 and $\mathcal{V}_{np}(v_{\pi^e})$ is defined in Theorem 2.1.*

In Appendix A.4, we provide a finer analysis on the gap between the two efficiency bounds. Specifically, the efficiency bound $\mathcal{V}_{\text{np}}(v_{\pi^e})$ can also be decomposed into reward and transition parts just like (7)-(8), and each part is larger than its correspondence in the decomposition of $\mathcal{V}(v_{\pi^e})$. For the reward part, the difference between the efficiency bounds is characterized by the minimum eigenvalue of a p.s.d. matrix; and for the transition part, our simulation suggests that unbalancedness of feature maps $\phi(s,a)$ distributed in $\mathbb{R}^d$ may increase the corresponding difference.

To better characterize the scale of the efficiency bound, below we provide an upper bound for $\mathcal{V}(v_{\pi^e})$.

Proposition 3.7. *Let $\underline{\sigma} = \lambda_{\min}\left(\mathbb{E}[\phi(s,a)\phi(s,a)^\top]\right)$. Then*

$$\mathcal{V}(v_{\pi^e}) \leq O\left(\frac{1}{\underline{\sigma}(1-\gamma)^2} + \frac{1}{\underline{\sigma}^2(1-\gamma)^4}\right).$$

When the offline dataset is of good coverage, the minimum eigenvalue $\underline{\sigma}$ defined in Proposition 3.7 is roughly of order

$1/d$, so the upper bound is roughly $O\left(\frac{d}{(1-\gamma)^2}+\frac{d^2}{(1-\gamma)^4}\right)$ for $\mathcal{V}(v_{\pi^e})$. For $\mathcal{V}_{\text{np}}(v_{\pi^e})$, Gheshlaghi Azar et al. (2013) provides evidence that its upper bound is $O\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\right)$ (see Appendix A.5 for details). Therefore, in the tabular case, knowledge of a specific linear structure of the MDP reduces the minimum possible variance of an estimator from $O(|\mathcal{S}||\mathcal{A}|)$ to $O(d^2)$.

4. Efficient Estimation in Linear MDPs

In this section we construct an estimator that attains the efficiency bound in Theorem 3.4. For notation simplicity, define $\phi_{\pi^e}(s)=\mathbb{E}_{\pi^e}[\phi(s,a)\mid s]$, $\Phi_{\pi^e}=\mathbb{E}_{p^{(0)}_{\pi^e}}[\phi_{\pi^e}(s)]$ and $A=\int \nu_0(s)\phi_{\pi^e}(s)^\top \mathrm{d}s$. Without much calculation, we can show that

$$
\begin{aligned}
&v_{\pi^e}=\mathbb{E}_{p^{(0)}_{\pi^e}}[V_{\pi^e}(s)]=\Phi_{\pi^e}^\top(I-\gamma A)^{-1}\omega_0,\\
&V_{\pi^e}(s')=\phi_{\pi^e}(s')^\top(I-\gamma A)^{-1}\omega_0,\\
&\mathbb{E}_{p^{(\infty)}_{\pi^e}}[\phi(s,a)^\top]=(1-\gamma)\Phi_{\pi^e}^\top(I-\gamma A)^{-1}.
\end{aligned}
$$

Detailed derivation is deferred to Appendix B.1.

Our estimator is based on the following three estimating functions:

$$
\psi_0(\widehat{\eta})=\Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}})^{-1}\widehat{\omega}, \tag{9}
$$

$$
\begin{aligned}
\psi_1(s,a,r;\widehat{\eta},\widetilde{\mathcal{D}})=\Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}})^{-1}\times\\
\left\{\mathbb{E}_{\widetilde{\mathcal{D}}}[\phi(s,a)\widehat{\Omega}(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\times\\
\phi(s,a)\widehat{\Omega}(s,a)^{-1}(r-\phi(s,a)^\top\widehat{\omega}),
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
\psi_2(s,a,s';\widehat{\eta})=\gamma\Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}})^{-1}\\
\times\left[V_{\pi^e;\widehat{\eta}}(s')I-P_{\widehat{\Delta}}(V_{\pi^e;\widehat{\eta}})P_{\widehat{\Delta}}(1)^{-1}\right]\\
\times\frac{\widehat{\Delta}(s')^{-1}\phi(s,a)}{\phi(s,a)^\top\widehat{\nu}(s')},
\end{aligned} \tag{11}
$$

where $\mathbb{E}_{\widetilde{\mathcal{D}}}$ means the sample average over some dataset $\widetilde{\mathcal{D}}\subseteq\mathcal{D}$, $\widehat{\eta}=(\widehat{\omega},\widehat{\nu},\widehat{\Omega},\widehat{\Delta})$ is some estimate of the true parameter $\eta=(\omega_0,\nu_0,\Omega,\Delta)$, and

$$
A_{\widehat{\nu}}=\int\widehat{\nu}(s)\phi_{\pi^e}(s)^\top\mathrm{d}s, \tag{12}
$$

$$
V_{\pi^e;\widehat{\eta}}(s')=\phi_{\pi^e}(s')^\top(I-\gamma A_{\widehat{\nu}})^{-1}\widehat{\omega}, \tag{13}
$$

$$
P_{\widehat{\Delta}}(f)=\int\widehat{\Delta}(s')^{-1}f(s')\mathrm{d}s'\quad\forall f\colon\mathcal{S}\to\mathbb{R}. \tag{14}
$$

We assume for now the nuisance estimate $\widehat{\eta}$ is given; in Appendix B.2 we discuss some proper ways to construct it.

We use the well-known *sample splitting* technique to construct our estimator. Specifically, we first divide the data in to K folds $\mathcal{D}_1,\mathcal{D}_2,\dots,\mathcal{D}_K$, each fold consisting of n/K samples. For $k=1,2,\dots,K$, we construct an estimate $\widehat{\eta}^{(k)}=(\widehat{\omega}^{(k)},\widehat{\nu}^{(k)},\widehat{\Omega}^{(k)},\widehat{\Delta}^{(k)})$ based on the data except for those in k-th fold, i.e., $\mathcal{D}\backslash\mathcal{D}_k$. Our final estimator is

$$
\widehat{v}_{\text{LMDP}}=\frac{1}{n}\sum_{k=1}^{K}\sum_{(s_i,a_i,r_i,s_i')\in\mathcal{D}_k}\Big[\underbrace{\widehat{\psi}_0^{(k)}}_{\text{plug-in}}+\underbrace{\widehat{\psi}_{1,i}^{(k)}+\widehat{\psi}_{2,i}^{(k)}}_{\text{augmentation}}\Big], \tag{15}
$$

where $\widehat{\psi}_0^{(k)}=\psi_0(\widehat{\eta}^{(k)})$, $\widehat{\psi}_{1,i}^{(k)}=\psi_1(s_i,a_i,r_i;\widehat{\eta}^{(k)},\mathcal{D}\backslash\mathcal{D}_k)$ and $\widehat{\psi}_{2,i}^{(k)}=\psi_2(s_i,a_i,s_i';\widehat{\eta}^{(k)})$. The whole estimating procedure is shown in Algorithm 1.

Algorithm 1 One-Step Estimator

Input: dataset $\mathcal{D}$, feature map $\phi(\cdot,\cdot)$, initial distribution $p^{(0)}_{\pi^e}(\cdot)$, target policy $\pi^e(\cdot\mid\cdot)$, discount factor γ
Output: one-step estimator $\widehat{v}_{\text{LMDP}}$
Divide the dataset into K folds $\mathcal{D}_1,\mathcal{D}_2,\dots,\mathcal{D}_K$, each consisting of n/K samples;
for $k=1,2,\dots,K$ **do**
  Construct estimates $\widehat{\eta}^{(k)}=(\widehat{\omega}^{(k)},\widehat{\nu}^{(k)},\widehat{\Omega}^{(k)},\widehat{\Delta}^{(k)})$ based on $\mathcal{D}\backslash\mathcal{D}_k$;
  Construct $A_{\widehat{\nu}^{(k)}}, V_{\pi^e;\widehat{\eta}^{(k)}}(s'), P_{\widehat{\Delta}^{(k)}}(1), P_{\widehat{\Delta}^{(k)}}(V_{\pi^e;\widehat{\eta}^{(k)}})$ according to (12)-(14);
end for
Construct the final estimator $\widehat{v}_{\text{LMDP}}$ according to (15):

There are several points to mention. First, the sample splitting technique is widely used in statistics and econometrics, particularly when the nuisance parameters to be estimated are infinite-dimensional. In such cases, many modern machine learning estimators suffer from high-complexity phenomena and fail to satisfy a strict Donsker-type condition, so traditional estimators based on full data fail to attain a parametric $\sqrt{n}$ convergence rate. Sample splitting, on the other hand, only requires a much weaker condition and thus overcomes such a problem. See (Chernozhukov et al., 2018) and Chapter 19 of (Van der Vaart, 2000) for details.

Second, our estimator $\widehat{v}_{\text{LMDP}}$ consists of two components: the plug-in part (ψ_0) and the augmentation part (ψ_1 and ψ_2). The plug-in part itself is a direct estimator of v_{π^e}, while the augmentation part is a sample analogue of the efficient influence functions (3)-(4), which serves as a stabilizing term and lowers the variance of the estimator. Such an estimator is often called the *one-step estimator* and is frequently used to attain efficiency in the field of semiparametric statistics.

Third, we discuss the relationship between our estimator and the generalized least squares (GLS) estimator. Recall that ψ_0 is the plug-in estimator for the value function, ψ_1 is an augmentation term for the reward mechanism, and ψ_2 is an augmentation term for the transition mechanism. If we

focus on the reward mechanism, adding up ψ_0 and ψ_1 over samples will result in

$$\psi_0(\widehat{\eta}) + \frac{1}{n}\sum_{i=1}^{n}\psi_1(s_i,a_i,r_i;\widehat{\eta},\mathcal{D})$$
$$= \Phi_{\pi^e}^\top (I-\gamma A_{\widehat{\nu}})^{-1}\left\{\mathbb{E}_{\mathcal{D}}[\phi(s,a)\widehat{\Omega}(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}$$
$$\cdot\, \mathbb{E}_{\mathcal{D}}[\phi(s,a)\widehat{\Omega}(s,a)^{-1}r]\,.$$

Combining the shaded two terms gives rise to the GLS estimator for the immediate reward. This is natural because for linear MDPs we impose a linear moment condition on the reward only, and the GLS estimator is known to be efficient when there is no extra information aside from this moment condition.

The following theorem establishes the efficiency of our estimator $\widehat{v}_{\text{LMDP}}$ in (15) under mild conditions. For any $\widetilde{\eta} = (\widetilde{\omega},\widetilde{\nu},\widetilde{\Omega},\widetilde{\Delta})$, define $\psi_{\text{eff}}(s,a,r,s';\widetilde{\eta})$ as the efficient influence function under the distribution parameterized by $\widetilde{\eta}$ (i.e., replacing $\omega_0,\nu_0,\Omega,\Delta$ with $\widetilde{\omega},\widetilde{\nu},\widetilde{\Omega},\widetilde{\Delta}$ in $\psi_{\text{eff}}(s,a,r,s')$ defined in Theorem 3.4), and $\Delta_{\widetilde{\nu}}(s') = \int p_{\pi^b}(s,a)\frac{\phi(s,a)\phi(s,a)^\top}{\phi(s,a)^\top\widetilde{\nu}(s')}\mathrm{d}s\mathrm{d}a$ (i.e., replacing ν_0 with $\widetilde{\nu}$ in $\Delta(s')$).

Theorem 4.1. *Suppose the true model is in $\mathcal{F}_{lin}$, and the following conditions hold for any $1\le k\le K$:*

1. *$\widehat{\eta}^{(k)}$ converges to its true value η in the sense that $\mathbb{E}\left[\{\psi_{\text{eff}}(s,a,r,s';\widehat{\eta}^{(k)})-\psi_{\text{eff}}(s,a,r,s';\eta)\}^2 \mid \widehat{\eta}^{(k)}\right] \xrightarrow{p} 0$;*
2. *the nuisance estimates satisfy $\|\widehat{\omega}^{(k)}-\omega_0\|_2 = o_p(\alpha_n^\omega)$, $\|A_{\widehat{\nu}^{(k)}}-A\|_2 = o_p(\alpha_n^\nu)$, $\|\widehat{\Delta}^{(k)}(s')^{-1}\Delta_{\widehat{\nu}^{(k)}}(s')-I\|_2 = o_p(\alpha_n^\Delta)$ and $\|P_{\widehat{\Delta}^{(k)}}(V_{\pi^e;\widehat{\eta}^{(k)}})P_{\widehat{\Delta}^{(k)}}(1)^{-1} - P_{\Delta_{\widehat{\nu}^{(k)}}}(V_{\pi^e;\widehat{\eta}^{(k)}})P_{\Delta_{\widehat{\nu}^{(k)}}}(1)^{-1}\|_2 = o_p(\widetilde{\alpha}_n^\Delta)$, with $\alpha_n^\nu = O(n^{-1/4})$, $\alpha_n^\omega\alpha_n^\nu = O(n^{-1/2})$, $\alpha_n^\nu\alpha_n^\Delta = O(n^{-1/2})$ and $\widetilde{\alpha}_n^\Delta = O(n^{-1/2})$;*
3. *$\widehat{\Omega}^{(k)} \in \mathcal{G}_\Omega$ such that $\{\phi(s,a)\widetilde{\Omega}(s,a)^{-1}\phi(s,a)^\top : \widetilde{\Omega}\in\mathcal{G}_\Omega\}$ is a Glivenko-Cantelli class.*

Then $\widehat{v}_{LMDP}$ is efficient w.r.t. $\mathcal{F}_{lin}$; that is, $\sqrt{n}(\widehat{v}_{LMDP}-v_{\pi^e}) \xrightarrow{d} \mathcal{N}(0,\mathcal{V}(v_{\pi^e}))$ where $\mathcal{V}(v_{\pi^e})$ is the efficiency bound defined in Theorem 3.4.

The proof of Theorem 4.1 is deferred to Appendix B.3. We discuss the meaning of each condition in Theorem 4.1. Condition 1 is a mild consistency requirement that $\widehat{\omega}^{(k)}$, $\widehat{\nu}^{(k)}$, $\widehat{\Omega}^{(k)}$ and $\widehat{\Delta}^{(k)}$ converge to their true values in a proper sense. Condition 2 specifies the convergence rate of each parameter. These convergence rate requirements are also easy to satisfy: the convergence rate of $\widehat{\nu}^{(k)}$, characterized by $\|A_{\widehat{\nu}^{(k)}}-A\|_2$, is enough at a $n^{-1/4}$ rate; since $\omega_0\in\mathbb{R}^d$ is a finite-dimensional parameter, $\widehat{\omega}^{(k)}$ often achieves a parametric $n^{-1/2}$ rate so that $\alpha_n^\omega\alpha_n^\nu = O(n^{-1/2})$ is automatically satisfied; α_n^Δ and $\widetilde{\alpha}_n^\Delta$ measure the difference between $\widehat{\Delta}^{(k)}$ and $\Delta_{\widehat{\nu}^{(k)}}$. When the data distribution $p_{\pi^b}(s,a)$ is known and numerical integration in the form of Δ in (5) is feasible, we can directly use $\Delta_{\widehat{\nu}^{(k)}}$ as an estimate of Δ. In this case, $\widehat{\Delta}^{(k)} = \Delta_{\widehat{\nu}^{(k)}}$ and $\alpha_n^\Delta = \widetilde{\alpha}_n^\Delta = 0$. Otherwise, α_n^Δ and $\widetilde{\alpha}_n^\Delta$ characterize the bias induced by estimation of $p_{\pi^b}(s,a)$ and/or numerical approximation of the integral in the form of Δ. Condition 3 is an artificial but mild assumption for ease of proof. It restricts the complexity of the estimation class $\mathcal{G}_\Omega$ for Ω, and thus ensures the sample mean of $\phi(s,a)\widehat{\Omega}^{(k)}(s,a)^{-1}\phi(s,a)^\top$ to converge to its population analogue. The definition of Glivenko-Cantelli classes along with a sufficient condition is discussed in Appendix D.3.

Remark 4.2. Condition 2 of Theorem 4.1 also indicates that for $\widehat{v}_{\text{LMDP}}$ to be consistent (i.e., $\widehat{v}_{\text{LMDP}} \xrightarrow{p} v_{\pi^e}$), it is necessary for the nuisance estimator $\widehat{\nu}^{(k)}$'s to be consistent. Therefore, our estimator does not enjoy the *doubly-robust property*, which often occurs in causal inference literature and only requires one out of all nuisance parameters to be consistent. Technically, the requirement that $\widehat{\nu}^{(k)}$'s should be consistent arises from an error term in the proof which solely depends on the error of $\widehat{\nu}^{(k)}$'s. Intuitively, absence of the doubly-robust property is reasonable since this nonparametric term enters the efficient influence functions (10)-(11) in a complex and nonlinear way, while for most doubly-robust estimators, efficient influence functions should be linear in each nuisance parameter (see e.g., Section 3.1 of (Robins et al., 2008)).

As mentioned earlier, in the $\widehat{\Delta}^{(k)} = \Delta_{\widehat{\nu}^{(k)}}$ case we have the following corollary.

Corollary 4.3. *Suppose the true model is in $\mathcal{F}_{lin}$, $\widehat{\Delta}^{(k)} = \Delta_{\widehat{\nu}^{(k)}}$, $\|\widehat{\omega}^{(k)}-\omega_0\|_2 = O_p(n^{-1/2})$ and $\|A_{\widehat{\nu}^{(k)}}-A\|_2 = o_p(n^{-1/4})$ for any $1\le k\le K$, and Conditions 1 and 3 of Theorem 4.1 hold. Then $\widehat{v}_{LMDP}$ is efficient w.r.t. $\mathcal{F}_{lin}$.*

5. Asymptotically Valid Confidence Intervals

In addition to an efficient point estimation, we are often interested in carrying out statistical inference for v_{π^e}. The asymptotic result $\sqrt{n}(\widehat{v}_{\text{LMDP}}-v_{\pi^e}) \xrightarrow{d} \mathcal{N}(0,\mathcal{V}(v_{\pi^e}))$ in Theorem 4.1 gives us a direct way of constructing an asymptotically valid confidence interval for v_{π^e}. Given a consistent variance estimator $\widehat{\mathcal{V}} \xrightarrow{p} \mathcal{V}(v_{\pi^e})$, we will always have $\mathbb{P}\left(|\widehat{v}_{\text{LMDP}}-v_{\pi^e}| \le z_{1-\alpha/2}\sqrt{\widehat{\mathcal{V}}/n}\right) \to 1-\alpha$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of a standard normal distribution. This means the confidence interval

$$\left[\widehat{v}_{\text{LMDP}} - z_{1-\alpha/2}\sqrt{\widehat{\mathcal{V}}/n},\ \widehat{v}_{\text{LMDP}} + z_{1-\alpha/2}\sqrt{\widehat{\mathcal{V}}/n}\right] \quad (16)$$

has an asymptotic $1-\alpha$ coverage rate.

It suffices to construct a consistent variance estimator $\widehat{\mathcal{V}}$. We propose the following estimator $\widehat{\mathcal{V}} = \widehat{\mathcal{V}}_1 + \widehat{\mathcal{V}}_2$, where

$$\begin{aligned}\widehat{\mathcal{V}}_1 &= \Phi_{\pi^e}^\top (I - \gamma A_{\widehat{\nu}})^{-1} \\ &\quad \cdot \Big\{ \mathbb{E}_{\mathcal{D}}[\phi(s,a)\widehat{\Omega}(s,a)^{-1}\phi(s,a)^\top] \Big\}^{-1} \\ &\quad \cdot (I - \gamma A_{\widehat{\nu}})^{-\top} \Phi_{\pi^e}, \qquad (17) \\ \widehat{\mathcal{V}}_2 &= \gamma^2 \Phi_{\pi^e}^\top (I - \gamma A_{\widehat{\nu}})^{-1} \\ &\quad \cdot \big\{ P_{\widehat{\Delta}}(V_{\pi^e;\widehat{\eta}}^2) - P_{\widehat{\Delta}}(V_{\pi^e;\widehat{\eta}}) P_{\widehat{\Delta}}(1)^{-1} P_{\widehat{\Delta}}(V_{\pi^e;\widehat{\eta}}) \big\} \\ &\quad \cdot (I - \gamma A_{\widehat{\nu}})^{-\top} \Phi_{\pi^e}. \qquad (18)\end{aligned}$$

Note that (17)-(18) are sample analogues of (7)-(8), so $\widehat{\mathcal{V}}$ is the plug-in estimator for $\mathcal{V}(\pi^e)$. The following theorem shows its consistency and thus the asymptotic validity of the confidence interval (16).

Theorem 5.1. *Suppose the following conditions hold:*

1. $\|\widehat{\omega} - \omega_0\|_2 \xrightarrow{p} 0$, $\|A_{\widehat{\nu}} - A\|_2 \xrightarrow{p} 0$, *and*
$$\mathbb{E}\left[\left| \widehat{\Omega}(s,a)^{-1} - \Omega(s,a)^{-1} \right| \mid \widehat{\Omega} \right] \xrightarrow{p} 0,$$
$$\int \left\| \widehat{\Delta}(s')^{-1} - \Delta(s')^{-1} \right\|_2 \mathrm{d}s' \xrightarrow{p} 0;$$
2. $\widehat{\Omega} \in \mathcal{G}_\Omega$ *such that* $\{\phi(s,a)\widetilde{\Omega}(s,a)^{-1}\phi(s,a)^\top : \widetilde{\Omega} \in \mathcal{G}_\Omega\}$ *is a Glivenko-Cantelli class.*

Then $\widehat{\mathcal{V}} \xrightarrow{p} \mathcal{V}(v_{\pi^e})$, *and* (16) *is an asymptotically valid confidence interval for* v_{π^e} *with a* $1-\alpha$ *coverage rate.*

The proof of Theorem 5.1 is deferred to Appendix B.4. The validity of our variance estimator $\widehat{\mathcal{V}}$ along with the corresponding inference procedure is illustrated in simulation studies.

6. Simulation Studies

In this section we implement simulation experiments to demonstrate the efficiency of our estimator and the validity of our proposed inference procedure. We consider a linear MDP with discrete state and action spaces, where $|\mathcal{S}| = 30$, $|\mathcal{A}| = 10$, $d = 5$ and $\gamma = 0.8$. The feature map $\{\phi(s,a)\}_{s\in\mathcal{S}, a\in\mathcal{A}}$ is constructed by drawing i.i.d. $\mathrm{Exp}(1)$ random variables for each component of $\phi(s,a)$ and then normalizing it to satisfy $\sum_{i=1}^d \phi_i(s,a) = 1$. The reward parameter ω_0 has its components generated from i.i.d. $\mathrm{Unif}([0,1])$, and for each $s \in \mathcal{S}$, the transition parameter $\nu_0(s)$ has its components generated from i.i.d. $\mathrm{Exp}(1)$ followed by normalization to satisfy $\sum_{s\in\mathcal{S}} \nu_0(s) = 1_d$. The feature map and true parameters are kept fixed once they are generated. Denoting $\mathcal{S} = \{0, 1, \ldots, 29\}$ and $\mathcal{A} = \{0, 1, \ldots, 9\}$, we set the variance of the reward as $\Omega(s,a) = 1/100 + (10s + a)/600$, and the behavior and target policies are defined as

$$\pi^b(a \mid s) = \begin{cases} 0.2, & \text{if } a \equiv s-1, \\ 0.2, & \text{if } a \equiv s, \\ 0.6, & \text{if } a \equiv s+1, \\ 0, & \text{otherwise,} \end{cases} \quad \forall s \in \mathcal{S},$$
$$\pi^e(a \mid s) = 0.1, \quad \forall s \in \mathcal{S}, a \in \mathcal{A},$$

where $\equiv$ means equivalence in the sense of modulo 10. The initial state distribution is set as $p_{\pi_b}^{(0)}(s) = 1/30$, $\forall s \in \mathcal{S}$. Our aim is to evaluate the value function at $s_0 = 0$, i.e., $v_{\pi^e} = V_{\pi^e}(0)$. In the following, all simulation experiments are repeated by 1,000 times, and the number of samples used ranges from 5,000 to 100,000.

Efficiency of our estimator. We first compare the performance of our estimator with two other estimators: the direct method (DM) estimator and the double reinforcement learning (DRL) estimator. The DM estimator is constructed as $\widehat{v}_{\mathrm{DM}} = \Phi_{\pi^e}^\top (I - \gamma A_{\widehat{\nu}})^{-1} \widehat{\omega}$, which is often used in the case of linear MDPs and is also a consistent estimator for v_{π^e}. The DRL estimator is constructed according to $\widehat{v}_{\mathrm{DR}}$ in Theorem 2.1. Note that constructing $\widehat{v}_{\mathrm{DR}}$ needs two nuisance estimates for the density ratio $w(s,a)$ and the Q-function $Q_{\pi^e}(s,a)$. We describe our approach to such nuisance estimation in Appendix C.1.

Figure 1 illustrates the performance of the three estimators, all without sample splitting. The left plot shows the average, 75-th quantile and 25-th quantile of the 1,000 estimated biases under different numbers of samples, while the right plot shows the mean square error (MSE) of the corresponding 1,000 estimates. It is shown that the three estimators all converge to the true value v_{π^e}, while our estimator has the minimal fluctuation in bias as well as the minimal MSE. This demonstrates the efficiency of our estimator, and in turn, implies that both the DM and DRL estimators are not capable of capturing information in the linear structure of MDPs. Results of the same three estimators using 2-fold and 5-fold sample splitting are deferred to Appendix C.2, which exhibit similar patterns.

Necessity of sample splitting. We next explore whether the sample splitting (or cross-fitting) technique leads to any improvement in our estimator. We construct an estimator without sample splitting (i.e., all samples are used to construct nuisance estimates), a 2-fold sample splitting estimator and a 5-fold sample splitting estimator. Under different numbers of samples, the average, 75-th quantile and 25-th quantile of the 1,000 estimated biases, along with the mean square error (MSE) of the corresponding 1,000 estimates, are shown in Figure 2. We find that all three estimators share

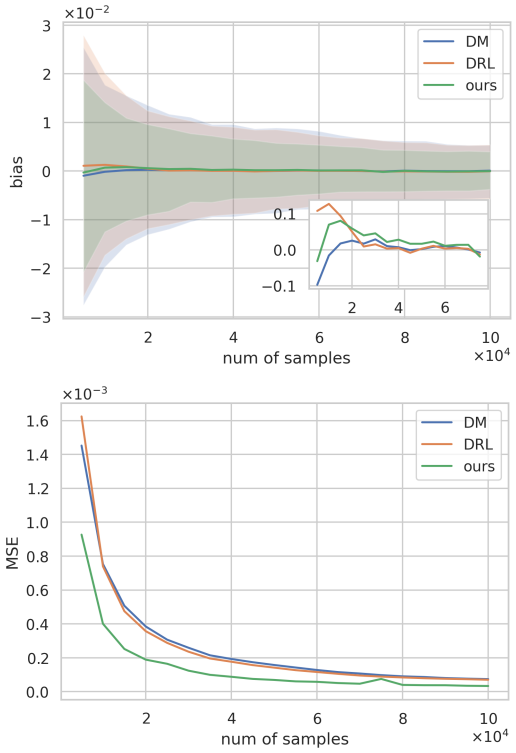

Figure 1. Up: the average, 75-th quantile and 25-th quantile of biases of three estimators. Down: the mean square errors (MSEs) of three estimators.

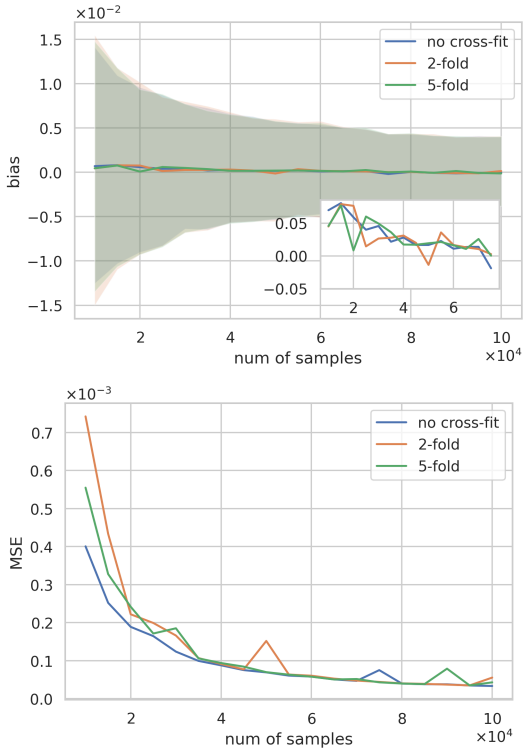

Figure 2. Up: the average, 75-th quantile and 25-th quantile of biases of estimators with/without sample splitting. Down: the mean square errors (MSEs) of estimators with/without sample splitting.

a similar scale of biases and MSEs. For two estimators using sample splitting, the 5-fold one slightly outperforms the 2-fold one when the sample size is not too large. This phenomenon may come from the fact that the former uses more data in constructing nuisance estimators than the latter, which may potentially increase stability in estimation. We also compare the performance of DRL estimators with and without sample splitting in Appendix C.2.

Validity of the inference procedure. We finally validate the confidence interval constructed in Section 5. Choosing $\alpha = 0.05$, we plot the coverage rates of the interval (16) for the estimator without sample splitting, the 2-fold sample splitting estimator and the 5-fold sample splitting estimator in Figure 3, and report the CI lengths in Table 1. It is shown that all three coverage rates achieves the nominal rate 0.95 when the number of samples is larger than 30,000, thus proving the validity of our inference procedure.

Table 1. CI lengths under different sample sizes.

NUM OF SAMPLES ($\times 10^4$)	2	4	6
CI LENGTH	0.0504	0.0359	0.0293
NUM OF SAMPLES ($\times 10^4$)	8	10	
CI LENGTH	0.0254	0.0228	

7. Concluding Remarks

In this paper we have established the semiparametric theory for linear MDPs as well as a complete approach to efficient estimation and inference, including an efficient estimator $\widehat{v}_{\text{LMDP}}$ for the value v_{π^e} and a corresponding asymptotically valid confidence interval for its inference. Simulation studies have confirmed the correctness of our theoretical results and proposed methods.

There are at least two further research directions. The first is to extend our estimator to the setting where the state and action spaces ($\mathcal{S}$ and $\mathcal{A}$) are continuous. In that case, integrals in the form of nuisance parameters Δ and P_Δ are usually intractable, so proper approximation techniques such as Monte Carlo and/or numerical integration should be considered. Another direction is to analyze semiparametric theory and propose efficient estimation for MDPs with different structures. The methodology we adopt enables one to make the best of the known structural information contained in the model.

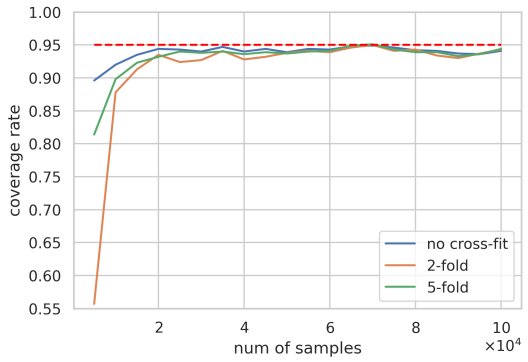

Figure 3. Coverage rates for estimators with/without sample splitting. The red line is the nominal coverage rate 0.95.

Acknowledgements

This work has been supported by the National Key Research and Development Project of China (No. 2022YFA1004002) and the National Natural Science Foundation of China (No. 12271011).

References

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005.

Bennett, A. and Kallus, N. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *arXiv preprint arXiv:2110.15332*, 2021.

Bertsekas, D. *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific, 2012.

Bickel, P. J., Klaassen, C., Ritov, Y., and Wellner, J. Efficient and adaptive inference in semiparametric models, 1993.

Chakraborty, B. and Moodie, E. Statistical methods for dynamic treatment regimes. *Springer-Verlag. doi*, 10: 978–1, 2013.

Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., and Chi, E. H. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 456–464, 2019.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen, E. T. Semiparametric proximal causal inference. *arXiv preprint arXiv:2011.08411*, 2020.

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

Gheshlaghi Azar, M., Munos, R., and Kappen, H. J. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91:325–349, 2013.

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

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167), 2020.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 2022.

Kennedy, E. H. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pp. 141–167. Springer, 2016.

Kennedy, E. H. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

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

Liu, J. S. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

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

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 2019.

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

Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., and Popovic, Z. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077, 2014.

Min, Y., Wang, T., Zhou, D., and Gu, Q. Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems*, 34: 7598–7610, 2021.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. Safe and efficient off-policy reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Murphy, S. A. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

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

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Ray, K. and van der Vaart, A. Semiparametric bayesian causal inference. *The Annals of Statistics*, 48(5):2999–3020, 2020.

Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and statistics: essays in honor of David A. Freedman*, 2:335–421, 2008.

Robins, J. M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121, 1995.

Schwartz, S. L., Li, F., and Mealli, F. A bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association*, 106(496):1331–1344, 2011.

Severini, T. A., Tripathi, G., et al. A survey of semiparametric efficiency bounds for some microeconometric models. Technical report, Department of Economics at the University of Luxembourg, 2013.

Shao, J. *Mathematical statistics*. Springer Science & Business Media, 2003.

Shi, C., Zhu, J., Shen, Y., Luo, S., Zhu, H., and Song, R. Off-policy confidence interval estimation with confounded markov decision process. *arXiv preprint arXiv:2202.10589*, 2022.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

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

Tsiatis, A. A. *Semiparametric theory and missing data*. Springer, 2006.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., and Xie, T. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

Uehara, M., Shi, C., and Kallus, N. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

Zhang, R., Dai, B., Li, L., and Schuurmans, D. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020.

A. Semiparametric Theory in Linear MDPs

A.1. Background of Semiparametric Theory

In this section, we review some precise definitions and results of semiparametric theory (Bickel et al., 1993; Van der Vaart, 2000; Tsiatis, 2006) that will be frequently used throughout the paper. We consider a semiparametric model $\mathcal{F}$ and denote the true data distribution by $F \in \mathcal{F}$. Our purpose is to estimate a functional of the data distribution, $\beta(F)$, which is also known as the parameter of interest, based on a set of i.i.d. observables $\{O_i\}_{i=1}^n$.

In the sequel, we denote $\mathcal{L}_2$ as the Hilbert space of all square-integrable functions, and $\mathcal{L}_2^0$ as the Hilbert space of all mean-zero functions in $\mathcal{L}_2$.

Definition A.1 (One-dimensional submodel and score function)**.** A parametric model $\mathcal{F}_{\text{sub}} = \{F_\theta \colon \theta \in \mathbb{R}\}$ is called a one-dimensional submodel of $\mathcal{F}$ passing through F, if (a) $F = F_0 \in \mathcal{F}_{\text{sub}}$, (b) $\mathcal{F}_{\text{sub}} \subset \mathcal{F}$, (c) the score function

$$s(O;\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log(\mathrm{d}F_\theta/\mathrm{d}\mu)(O)$$

exists and satisfies $\mathbb{E}[s(O;0)^2] < \infty$, and (d) $\mathbb{E}\sup_{\theta\in\mathbb{R}} |(\mathrm{d}F_\theta/\mathrm{d}\mu)(O)| < \infty$.

Definition A.2 (Tangent space)**.** The tangent space $\Lambda_{\mathcal{F}}(F)$ at F with respect to the model $\mathcal{F}$ is the linear closure of the score functions at F over all one-dimensional submodels with respect to the $\mathcal{L}_2$ space.

Definition A.3 (Pathwise differentiability, gradient, and efficient influence function)**.** A functional $\beta(F)$ is pathwise differentiable at F with respect to the model $\mathcal{F}$, if there exists a function $\psi_F(O) \in \mathcal{L}_2$ such that for any one-dimensional submodel $\mathcal{F}_{\text{sub}} = \{F_\theta \colon \theta \in \mathbb{R}\}$ of $\mathcal{F}$ passing through F at $\theta = 0$ with score function $s(O;\theta)$, it holds that

$$\left.\frac{\mathrm{d}\beta(F_\theta)}{\mathrm{d}\theta}\right|_{\theta=0} = \mathbb{E}[\psi_F(O)s(O;0)].$$

The function $\psi_F(O)$ is called a gradient of $\beta(F)$ at F with respect to the model $\mathcal{F}$. The efficient influence function (EIF) $\psi_{F,\text{eff}}(O)$ is defined as the unique mean-zero gradient that belongs to the tangent space $\Lambda_{\mathcal{F}}(F)$.

Definition A.4 (Regular estimators)**.** An estimator $\widehat{\beta}_n$ is regular for estimating $\beta(F)$ with respect to the model $\mathcal{F}$, if there exists a law G such that for any one-dimensional submodel $\mathcal{F}_{\text{sub}} = \{F_\theta \colon \theta \in \mathbb{R}\}$ of $\mathcal{F}$ passing through F, it holds that

$$\sqrt{n}\{\widehat{\beta}_n - \beta(F_{1/\sqrt{n}})\} \xrightarrow{D(F_{1/\sqrt{n}})} G,$$

where $D(F_{1/\sqrt{n}})$ means convergence in distribution under $F_{1/\sqrt{n}}$.

Definition A.5 (Asymptotically linear estimators)**.** An estimator $\widehat{\beta}_n$ is asymptotically linear with influence function $\psi(O) \in \mathcal{L}_2^0$ if

$$\sqrt{n}\{\widehat{\beta}_n - \beta(F)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(O_i) + o_p(1).$$

Pathwise differentiability ensures the parameter of interest $\beta(F)$ to be estimable at a $\sqrt{n}$ convergence rate, excluding unfavorable ones such as the density at a point. Regularity and asymptotic linearity ensure us to focus on those locally well-behaved estimators, excluding super-efficient estimators such as the Hodge estimator. If an estimator $\widehat{\beta}_n$ is both regular and asymptotically linear, then it is called a *RAL estimator*.

For a pathwise differentiable $\beta(F)$, the *efficiency bound* is defined as the variance of its efficient influence function, i.e., $\mathcal{V}_{\mathcal{F}}(\beta, F) = \text{var}\{\psi_{F,\text{eff}}(O)\}$. It is the minimal asymptotic variance that can be achieved by any regular estimator $\widehat{\beta}_n$. We then say $\widehat{\beta}_n$ is *efficient* if it is regular and the limiting distribution of $\sqrt{n}\{\widehat{\beta}_n - \beta(F)\}$ is $\mathcal{N}(0, \mathcal{V}_{\mathcal{F}}(\beta, F))$. The next theorem states a necessary and sufficient condition for an estimator $\widehat{\beta}_n$ to be efficient.

Lemma A.6 ((Van der Vaart, 2000), Lemma 25.23)**.** *Let $\beta(F)$ be pathwise differentiable at F with respect to the model $\mathcal{F}$ with efficient influence function $\psi_{F,\text{eff}}(O)$. Then an estimator $\widehat{\beta}_n$ is efficient, if and only if it is asymptotically linear with influence function $\psi_{F,\text{eff}}(O)$.*

A.2. Proof of Theorem 3.4

Proof of Theorem 3.4. We denote $\mathcal{F} = \mathcal{F}_{\text{lin}}$ for notational simplicity. To derive the efficient influence function of v_{π^e} with respect to the model $\mathcal{F}$, we proceed with the following three steps: (a) calculate a mean-zero gradient $\psi(O)$ of v_{π^e}; (b) calculate the tangent space $\Lambda_{\mathcal{F}}$ of $\mathcal{F}$; (c) project $\psi(O)$ onto the tangent space $\Lambda_{\mathcal{F}}$ to obtain the efficient influence function $\psi_{\text{eff}}(O)$.

Calculating a mean-zero gradient $\psi(O)$. In (Kallus & Uehara, 2022), the authors consider the nonparametric model, containing all distributions induced by arbitrary initial state distributions, behavior policy distributions, reward distributions and transition probability kernels, i.e.,

$$\mathcal{F}_{\text{np}} = \{p(s,a,r,s') \colon p(s,a,r,s') = p_s(s)p_{a|s}(a \mid s)p_{r|s,a}(r \mid s,a)p_{s'|s,a}(s' \mid s,a)\}.$$

A mean-zero gradient of v_{π^e} with respect to the model $\mathcal{F}_{\text{np}}$ is given by

$$\psi(s,a,r,s') = \frac{1}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}\left(r + \gamma V_{\pi^e}(s') - Q_{\pi^e}(s,a)\right).$$

Since $\mathcal{F} \subset \mathcal{F}_{\text{np}}$ and reducing the semiparametric model size will only expand the set of gradients, $\psi(s,a,r,s')$ is still a mean-zero gradient of v_{π^e} with respect to the model $\mathcal{F}$. Recall that $\varepsilon = r - \phi(s,a)^\top\omega_0 = r - \mathbb{E}[r \mid s,a]$. By the Bellman equation

$$Q_{\pi^e}(s,a) = \mathbb{E}[r \mid s,a] + \gamma\mathbb{E}[V_{\pi^e}(s') \mid s,a],$$

the gradient can also be written as

$$\psi(s,a,\varepsilon,s') = \frac{1}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}\left\{\varepsilon + \gamma V_{\pi^e}(s') - \gamma\mathbb{E}[V_{\pi^e}(s') \mid s,a]\right\}.$$

Calculating the tangent space $\Lambda_{\mathcal{F}}$. Next, we derive the tangent space $\Lambda_{\mathcal{F}}$.

Lemma A.7. *The tangent space $\Lambda_{\mathcal{F}}$ of $\mathcal{F}$ is given by*

$$\Lambda_{\mathcal{F}} = \Bigg\{ g(s,a,\varepsilon) + \frac{\phi(s,a)^\top\dot\nu(s')}{\phi(s,a)^\top\nu_0(s')} \colon \int \dot\nu(s')\mathrm{d}s' = 0_d,\ \mathbb{E}[g(s,a,\varepsilon)] = 0,$$
$$\mathbb{E}[\varepsilon g(s,a,\varepsilon) \mid s,a] = c^\top\phi(s,a) \textit{ for some } c \in \mathbb{R}^d \Bigg\} \cap \mathcal{L}_2^0.$$

Proof of Lemma A.7. For any regular parametric submodel

$$\mathcal{F}_{\text{sub}} = \{p(s,a,r,s';\theta,\omega) = p_\theta(s)p_\theta(a \mid s)p_\theta(\varepsilon(\omega) \mid s,a)p_\theta(s' \mid s,a) \colon$$
$$\varepsilon(\omega) = r - \phi(s,a)^\top\omega,\ \mathbb{E}_\theta[\varepsilon(\omega)] = 0,\ p_\theta(s' \mid s,a) = \phi(s,a)^\top\nu_\theta(s')\},$$

the score function at $(\theta,\omega) = (0,\omega_0)$ is

$$g_\theta(s,a,\varepsilon,s') = g_s(s) + g_{a|s}(s,a) + g_{\varepsilon|s,a}(s,a,\varepsilon) + \frac{\phi(s,a)^\top\dot\nu(s')}{\phi(s,a)^\top\nu_0(s')},$$
$$g_\omega(s,a,\varepsilon,s') = -\kappa(\varepsilon \mid s,a)\phi(s,a),$$

where $\varepsilon = r - \phi(s,a)^\top\omega_0$, $g_s(s) = \frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(s)\big|_{\theta=0}$, $g_{a|s}(s,a) = \frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(a \mid s)\big|_{\theta=0}$, $g_{\varepsilon|s,a}(s,a,\varepsilon) = \frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(\varepsilon \mid s,a)\big|_{\theta=0}$, $\dot\nu(s') = \frac{\mathrm{d}}{\mathrm{d}\theta}\nu_\theta(s')\big|_{\theta=0}$ and $\kappa(\varepsilon \mid s,a) = \frac{\mathrm{d}\log p_0(\varepsilon|s,a)}{\mathrm{d}\varepsilon}$. Since $\varepsilon \in \mathcal{L}_2^0 \subset \mathcal{L}_1$, we have $\varepsilon p(\varepsilon \mid s,a) \xrightarrow{\varepsilon\to\pm\infty} 0$. Integration by parts yields $\mathbb{E}[\kappa(\varepsilon \mid s,a) \mid s,a] = 0$ and $\mathbb{E}[\varepsilon\kappa(\varepsilon \mid s,a) \mid s,a] = -1$, so $\mathbb{E}[\varepsilon g_\omega(s,a,\varepsilon,s') \mid s,a] = \phi(s,a)$. Thus, for any score function

$$g_{\theta,\omega}(s,a,\varepsilon,s') = c_1 g_\theta(s,a,\varepsilon,s') + c_2^\top g_\omega(s,a,\varepsilon,s') \in \Lambda_{\mathcal{F}_{\text{sub}}},$$

we can let

$$g(s,a,\varepsilon) = c_1\{g_s(s) + g_{a|s}(s,a) + g_{\varepsilon|s,a}(s,a,\varepsilon)\} - \kappa(\varepsilon \mid s,a)c_2^\top \phi(s,a),$$
$$\dot{\nu}(s') = \frac{\mathrm{d}}{\mathrm{d}\theta}\nu_\theta(s')\bigg|_{\theta=0},$$

so the conditions $\int \dot{\nu}(s')\mathrm{d}s' = 0_d$, $\mathbb{E}[g(s,a,\varepsilon)] = 0$ and $\mathbb{E}[\varepsilon g(s,a,\varepsilon) \mid s,a] = c_2^\top\phi(s,a)$ are satisfied. Since $\Lambda_{\mathcal{F}}$ is the closed linear span of all regular $\Lambda_{\mathcal{F}_{\text{sub}}}$'s, this implies

$$\Lambda_{\mathcal{F}} \subseteq \bigg\{g(s,a,\varepsilon) + \frac{\phi(s,a)^\top\dot{\nu}(s')}{\phi(s,a)^\top\nu_0(s')} \colon \int \dot{\nu}(s')\mathrm{d}s' = 0_d,\ \mathbb{E}[g(s,a,\varepsilon)] = 0,$$
$$\mathbb{E}[\varepsilon g(s,a,\varepsilon) \mid s,a] = c^\top\phi(s,a) \text{ for some } c \in \mathbb{R}^d\bigg\} \cap \mathcal{L}_2^0. \tag{19}$$

On the other hand, for any $g(s,a,\varepsilon)$ and $\dot{\nu}(s')$ satisfying these conditions, we let

$$\widetilde{g}(s,a,\varepsilon) = g(s,a,\varepsilon) + \kappa(\varepsilon \mid s,a)c^\top\phi(s,a),$$
$$g_s(s) = \mathbb{E}[\widetilde{g}(s,a,\varepsilon) \mid s],$$
$$g_{a|s}(s,a) = \mathbb{E}[\widetilde{g}(s,a,\varepsilon) \mid s,a] - \mathbb{E}[\widetilde{g}(s,a,\varepsilon) \mid s],$$
$$g_{\varepsilon|s,a}(s,a,\varepsilon) = \widetilde{g}(s,a,\varepsilon) - \mathbb{E}[\widetilde{g}(s,a,\varepsilon) \mid s,a],$$

so they satisfy $\mathbb{E}[g_s(s)] = 0$, $\mathbb{E}[g_{a|s}(s,a) \mid s] = 0$, $\mathbb{E}[g_{\varepsilon|s,a}(s,a,\varepsilon) \mid s,a] = 0$ and $\mathbb{E}[\varepsilon g_{\varepsilon|s,a}(s,a,\varepsilon) \mid s,a] = 0$. If $\widetilde{g}$ is bounded, then g_s, $g_{a|s}$ and $g_{\varepsilon|s,a}$ are bounded, and we can construct a regular parametric submodel $\mathcal{F}_{\text{sub}}$ with

$$p_\theta(s) = p_0(s)(1 + \theta g_s(s))c_s(\theta),$$
$$p_\theta(a \mid s) = p_0(a \mid s)(1 + \theta g_{a|s}(s,a))c_{a|s}(\theta),$$
$$p_\theta(\varepsilon \mid s,a) = p_0(\varepsilon \mid s,a)(1 + \theta g_{\varepsilon|s,a}(s,a,\varepsilon))c_{\varepsilon|s,a}(\theta),$$
$$\nu_\theta(s') = \nu_0(s') + \theta\dot{\nu}(s'),$$

where $c_s(\theta)$, $c_{a|s}(\theta)$ and $c_{\varepsilon|s,a}(\theta)$ are normalizing constants. Since g_s, $g_{a|s}$ and $g_{\varepsilon|s,a}$ are bounded, we can restrict $\theta \in (-\delta,\delta)$ with sufficiently small $\delta > 0$ such that the densities are positive and well-defined. Thus, the tangent space $\Lambda_{\mathcal{F}_{\text{sub}}}$ contains

$$g_s(s) + g_{a|s}(s,a) + g_{\varepsilon|s,a}(s,a,\varepsilon) - \kappa(\varepsilon \mid s,a)c^\top\phi(s,a) + \frac{\phi(s,a)^\top\dot{\nu}(s')}{\phi(s,a)^\top\nu_0(s')}$$
$$= g(s,a,\varepsilon) + \frac{\phi(s,a)^\top\dot{\nu}(s')}{\phi(s,a)^\top\nu_0(s')}.$$

If $\widetilde{g}$ is unbounded, then we can construct a sequence of bounded score functions $\{\widetilde{g}^{(n)}\}_{n=1}^\infty$ converging to $\widetilde{g}$ in the $\mathcal{L}_2$ space, and then construct $\{g_s^{(n)}\}_{n=1}^\infty$, $\{g_{a|s}^{(n)}\}_{n=1}^\infty$ and $\{g_{\varepsilon|s,a}^{(n)}\}_{n=1}^\infty$ with the same procedure as before:

$$g_s^{(n)}(s) = \mathbb{E}[\widetilde{g}^{(n)}(s,a,\varepsilon) \mid s],$$
$$g_{a|s}^{(n)}(s,a) = \mathbb{E}[\widetilde{g}^{(n)}(s,a,\varepsilon) \mid s,a] - \mathbb{E}[\widetilde{g}^{(n)}(s,a,\varepsilon) \mid s],$$
$$g_{\varepsilon|s,a}^{(n)}(s,a,\varepsilon) = \widetilde{g}^{(n)}(s,a,\varepsilon) - \mathbb{E}[\widetilde{g}^{(n)}(s,a,\varepsilon) \mid s,a].$$

By the same construction of $\mathcal{F}_{\text{sub}}$ as above, we can construct a sequence of regular parametric submodels $\{\mathcal{F}_{\text{sub}}^{(n)}\}_{n=1}^\infty$, with the tangent space of each $\mathcal{F}_{\text{sub}}^{(n)}$ containing

$$\widetilde{g}^{(n)}(s,a,\varepsilon) - \kappa(\varepsilon \mid s,a)c^\top\phi(s,a) + \frac{\phi(s,a)^\top\dot{\nu}(s')}{\phi(s,a)^\top\nu_0(s')},$$

which converges to $g(s,a,\varepsilon)+\frac{\phi(s,a)^\top\dot\nu(s')}{\phi(s,a)^\top\nu_0(s')}$ in the $\mathcal{L}_2$ space. Since $\Lambda_{\mathcal{F}}$ is a closed linear span, we obtain

$$\Lambda_{\mathcal{F}} \supseteq \Bigg\{ g(s,a,\varepsilon)+\frac{\phi(s,a)^\top\dot\nu(s')}{\phi(s,a)^\top\nu_0(s')} \colon \int \dot\nu(s')\mathrm{d}s' = 0_d,\ \mathbb{E}[g(s,a,\varepsilon)]=0, \\ \mathbb{E}[\varepsilon g(s,a,\varepsilon)\mid s,a] = c^\top\phi(s,a) \text{ for some } c\in\mathbb{R}^d \Bigg\} \cap \mathcal{L}_2^0. \tag{20}$$

Combining (19) and (20) concludes the proof. □

Projecting $\psi(O)$ onto $\Lambda_{\mathcal{F}}$. Note that the tangent space $\Lambda_{\mathcal{F}}$ can be decomposed into two orthogonal parts, $\Lambda_{\mathcal{F}} = \Lambda_{\mathcal{F},1}\oplus\Lambda_{\mathcal{F},2}$, where

$$\Lambda_{\mathcal{F},1} = \big\{ g(s,a,\varepsilon)\colon \mathbb{E}[g(s,a,\varepsilon)]=0,\ \mathbb{E}[\varepsilon g(s,a,\varepsilon)\mid s,a] = c^\top\phi(s,a) \text{ for some } c\in\mathbb{R}^d \big\}\cap\mathcal{L}_2^0,$$
$$\Lambda_{\mathcal{F},2} = \left\{ \frac{\phi(s,a)^\top\dot\nu(s')}{\phi(s,a)^\top\nu_0(s')} \colon \int\dot\nu(s')\mathrm{d}s'=0_d \right\}\cap\mathcal{L}_2^0.$$

Therefore, to calculate the efficient influence function $\psi_{\text{eff}}(s,a,\varepsilon,s') = \Pi[\psi(s,a,\varepsilon,s')\mid\Lambda_{\mathcal{F}}]$, it suffices to calculate $\Pi[\psi(s,a,\varepsilon,s')\mid\Lambda_{\mathcal{F},1}]$ and $\Pi[\psi(s,a,\varepsilon,s')\mid\Lambda_{\mathcal{F},2}]$ respectively, and then add them up.

To calculate $\Pi[\psi(s,a,\varepsilon,s')\mid\Lambda_{\mathcal{F},1}]$, we first note that

$$\Pi[\psi(s,a,\varepsilon,s')\mid\Lambda_{\mathcal{F},1}] = \Pi[\mathbb{E}[\psi(s,a,\varepsilon,s')\mid s,a,\varepsilon]\mid\Lambda_{\mathcal{F},1}] = \Pi\left[\frac{1}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}\varepsilon\mid\Lambda_{\mathcal{F},1}\right].$$

Also note that $\Lambda_{\mathcal{F},1}=\Lambda_{\mathcal{F},1,1}\oplus\Lambda_{\mathcal{F},1,2}$, where

$$\Lambda_{\mathcal{F},1,1} = \{g(s,a)\colon \mathbb{E}[g(s,a)]=0\}\cap\mathcal{L}_2^0,$$
$$\Lambda_{\mathcal{F},1,2} = \big\{g(s,a,\varepsilon)\colon \mathbb{E}[g(s,a,\varepsilon)\mid s,a]=0,\ \mathbb{E}[\varepsilon g(s,a,\varepsilon)\mid s,a]=c^\top\phi(s,a)\text{ for some } c\in\mathbb{R}^d\big\}\cap\mathcal{L}_2^0.$$

The term $\frac{1}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}\varepsilon$ is orthogonal to $\Lambda_{\mathcal{F},1,1}$. Next, we directly apply Lemma D.3 with $Z=(s,a,r)$, $X=(s,a)$, $\theta_0=\omega_0$, $g(Z;\theta_0)=r-\phi(s,a)^\top\omega_0=\varepsilon$ and $\Lambda=\Lambda_{\mathcal{F},1,2}$ to obtain

$$\begin{aligned}\Pi[\psi(s,a,\varepsilon,s')\mid\Lambda_{\mathcal{F},1}] &= \Pi\left[\frac{1}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}\varepsilon\mid\Lambda_{\mathcal{F},1}\right]\\ &= \Pi\left[\frac{1}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}\varepsilon\mid\Lambda_{\mathcal{F},1,2}\right]\\ &= \frac{1}{1-\gamma}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}[\phi(s,a)^\top]\left\{\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\phi(s,a)\Omega(s,a)^{-1}\varepsilon.\end{aligned} \tag{21}$$

To calculate $\Pi[\psi(s,a,\varepsilon,s')\mid\Lambda_{\mathcal{F},2}]$, we first suppose the desired projection is

$$\Pi[\psi(s,a,\varepsilon,s')\mid\Lambda_{\mathcal{F},2}] = \frac{\phi(s,a)^\top\dot\nu_0(s')}{\phi(s,a)^\top\nu_0(s')},\quad \int\dot\nu_0(s')\mathrm{d}s'=0_d.$$

For any $\dot\nu(s')$ satisfying $\int\dot\nu(s')\mathrm{d}s'=0_d$, it must hold that

$$\begin{aligned}&\mathbb{E}\left\{\left[\frac{1}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}\{r+\gamma V_{\pi^e}(s')-Q_{\pi^e}(s,a)\}-\frac{\phi(s,a)^\top\dot\nu_0(s')}{\phi(s,a)^\top\nu_0(s')}\right]\frac{\phi(s,a)^\top\dot\nu(s')}{\phi(s,a)^\top\nu_0(s')}\right\}=0\\ \iff&\mathbb{E}\left\{\left[\frac{\gamma}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}V_{\pi^e}(s')-\frac{\phi(s,a)^\top\dot\nu_0(s')}{\phi(s,a)^\top\nu_0(s')}\right]\frac{\phi(s,a)^\top\dot\nu(s')}{\phi(s,a)^\top\nu_0(s')}\right\}=0\\ \iff&\int\left[\frac{\gamma}{1-\gamma}p^{(\infty)}_{\pi^e,\gamma}(s,a)V_{\pi^e}(s')-p_{\pi^b}(s,a)\frac{\phi(s,a)^\top\dot\nu_0(s')}{\phi(s,a)^\top\nu_0(s')}\right]\phi(s,a)^\top\dot\nu(s')\mathrm{d}s\mathrm{d}a\mathrm{d}s'=0.\end{aligned}$$

Since this holds for arbitrary $\dot{\nu}(s')$ satisfying $\int \dot{\nu}(s')\mathrm{d}s' = 0_d$, we have that

$$\int \left[\frac{\gamma}{1-\gamma} p^{(\infty)}_{\pi^e,\gamma}(s,a) V_{\pi^e}(s') - p_{\pi^b}(s,a)\frac{\phi(s,a)^\top \dot{\nu}_0(s')}{\phi(s,a)^\top \nu_0(s')}\right] \phi(s,a)\mathrm{d}s\mathrm{d}a = c, \tag{22}$$

for some $c \in \mathbb{R}^d$. Now, we define

$$\Delta(s') = \int p_{\pi^b}(s,a)\frac{\phi(s,a)\phi(s,a)^\top}{\phi(s,a)^\top \nu_0(s')}\mathrm{d}s\mathrm{d}a \in \mathbb{R}^{d\times d}.$$

Multiplying both sides of (22) by $\Delta(s')^{-1}$ from the left and integrate over s', we obtain

$$\frac{\gamma}{1-\gamma}\left[\int \Delta(s')^{-1} V_{\pi^e}(s')\mathrm{d}s'\right]\left[\int p^{(\infty)}_{\pi^e,\gamma}(s,a)\phi(s,a)\mathrm{d}s\mathrm{d}a\right] = \left[\int \Delta(s')^{-1}\mathrm{d}s'\right] c,$$

where we have used the fact that $\int \dot{\nu}_0(s')\mathrm{d}s' = 0_d$. Therefore we have solved

$$c = \frac{\gamma}{1-\gamma}\left[\int \Delta(s')^{-1}\mathrm{d}s'\right]^{-1}\left[\int \Delta(s')^{-1} V_{\pi^e}(s')\mathrm{d}s'\right]\left[\int p^{(\infty)}_{\pi^e,\gamma}(s,a)\phi(s,a)\mathrm{d}s\mathrm{d}a\right].$$

Using (22) again, we obtain

$$\begin{aligned}
\dot{\nu}_0(s') &= \frac{\gamma}{1-\gamma}\Delta(s')^{-1} V_{\pi^e}(s')\left[\int p^{(\infty)}_{\pi^e,\gamma}(s,a)\phi(s,a)\mathrm{d}s\mathrm{d}a\right] - \Delta(s')^{-1} c \\
&= \frac{\gamma}{1-\gamma}\Delta(s')^{-1}\left\{V_{\pi^e}(s') I - \left[\int \Delta(s')^{-1}\mathrm{d}s'\right]^{-1}\left[\int \Delta(s')^{-1} V_{\pi^e}(s')\mathrm{d}s'\right]\right\} \\
&\quad \cdot \left[\int p^{(\infty)}_{\pi^e,\gamma}(s,a)\phi(s,a)\mathrm{d}s\mathrm{d}a\right],
\end{aligned}$$

and consequently

$$\begin{aligned}
\Pi[\psi(s,a,\varepsilon,s') \mid \Lambda_{\mathcal{F},2}] &= \frac{\phi(s,a)^\top \dot{\nu}_0(s')}{\phi(s,a)^\top \nu_0(s')} \\
&= \frac{\gamma}{1-\gamma}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}[\phi(s,a)^\top] \\
&\quad \cdot \left\{V_{\pi^e}(s') I - \left[\int V_{\pi^e}(s')\Delta(s')^{-1}\mathrm{d}s'\right]\left[\int \Delta(s')^{-1}\mathrm{d}s'\right]^{-1}\right\} \\
&\quad \cdot \frac{\Delta(s')^{-1}\phi(s,a)}{\phi(s,a)^\top \nu_0(s')}.
\end{aligned} \tag{23}$$

Finally, combining (21), (23) and the definition of $P_\Delta(f)$ yields the efficient influence function result of Theorem 3.4. Since $\Lambda_{\mathcal{F},1}$ and $\Lambda_{\mathcal{F},2}$ are orthogonal, the efficiency bound is given by

$$\begin{aligned}
\mathcal{V}(v_{\pi^e}) &= \mathrm{var}\{\psi_{\text{eff},1}(s,a,\varepsilon) + \psi_{\text{eff},2}(s,a,s')\} \\
&= \mathrm{var}\{\psi_{\text{eff},1}(s,a,\varepsilon)\} + \mathrm{var}\{\psi_{\text{eff},2}(s,a,s')\},
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{var}\{\psi_{\text{eff},1}(s,a,\varepsilon)\} &= \frac{1}{(1-\gamma)^2}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}[\phi(s,a)^\top]\left\{\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}[\phi(s,a)], \\
\mathrm{var}\{\psi_{\text{eff},2}(s,a,s')\} &= \frac{\gamma^2}{(1-\gamma)^2}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}[\phi(s,a)^\top]\left\{P_\Delta(V^2_{\pi^e}) - P_\Delta(V_{\pi^e})P_\Delta(1)^{-1}P_\Delta(V_{\pi^e})\right\}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}[\phi(s,a)].
\end{aligned}$$

□

A.3. Theorem 3.4 in the Tabular Case

We assume that $\mathcal{S}$ and $\mathcal{A}$ are both finite sets, $d = |\mathcal{S}||\mathcal{A}|$, and the features $\{\phi(s,a)\}_{s\in\mathcal{S},a\in\mathcal{A}}$ form an orthonormal basis in $\mathbb{R}^d$. We will show that the efficient influence function $\psi_{\text{eff}}(s,a,\varepsilon,s')$ given in Theorem 3.4 degenerates into

$$\psi_{\text{eff,np}}(s,a,r,s') = \frac{1}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}\left(r+\gamma V_{\pi^e}(s') - Q_{\pi^e}(s,a)\right).$$

By assumption, there exists an orthonormal matrix $U\in\mathbb{R}^{d\times d}$ such that $\phi_0(s,a) = U^{-1}\phi(s,a)$ is a standard unit vector with its (s,a)-th component being 1 and any other component being 0, for any state-action pair (s,a). Thus we have

$$\begin{aligned}
\Delta(s') &= \int p_{\pi^b}(s,a)\frac{\phi(s,a)\phi(s,a)^\top}{\phi(s,a)^\top\nu_0(s')}\mathrm{d}s\mathrm{d}a \\
&= \int p_{\pi^b}(s,a)\frac{U\phi_0(s,a)\phi_0(s,a)^\top U^\top}{\phi_0(s,a)^\top U^\top\nu_0(s')}\mathrm{d}s\mathrm{d}a \\
&=: U\Delta_0(s')U^\top,
\end{aligned}$$

and consequently,

$$\begin{aligned}
\psi_{\text{eff},1}(s,a,\varepsilon) &= \frac{1}{1-\gamma}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}\left[\phi(s,a)^\top\right]\left\{\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\phi(s,a)\Omega(s,a)^{-1}\varepsilon \\
&= \frac{1}{1-\gamma}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}\left[\phi_0(s,a)^\top U^\top\right]\left\{\mathbb{E}[U\phi_0(s,a)\Omega(s,a)^{-1}\phi_0(s,a)^\top U^\top]\right\}^{-1}U\phi_0(s,a)\Omega(s,a)^{-1}\varepsilon \\
&= \frac{1}{1-\gamma}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}\left[\phi_0(s,a)^\top\right]\left\{\mathbb{E}[\phi_0(s,a)\Omega(s,a)^{-1}\phi_0(s,a)^\top]\right\}^{-1}\phi_0(s,a)\Omega(s,a)^{-1}\varepsilon, \\
\psi_{\text{eff},2}(s,a,s') &= \frac{\gamma}{1-\gamma}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}\left[\phi(s,a)^\top\right] \\
&\quad\cdot\left\{V_{\pi^e}(s')I - \left[\int V_{\pi^e}(s')\Delta(s')^{-1}\mathrm{d}s'\right]\left[\int\Delta(s')^{-1}\mathrm{d}s'\right]^{-1}\right\}\frac{\Delta(s')^{-1}\phi(s,a)}{\phi(s,a)^\top\nu_0(s')} \\
&= \frac{\gamma}{1-\gamma}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}\left[\phi_0(s,a)^\top U^\top\right] \\
&\quad\cdot\left\{V_{\pi^e}(s')I - \left[\int V_{\pi^e}(s')U^{-\top}\Delta_0(s')^{-1}U^{-1}\mathrm{d}s'\right]\left[\int U^{-\top}\Delta_0(s')^{-1}U^{-1}\mathrm{d}s'\right]^{-1}\right\} \\
&\quad\cdot\frac{U^{-\top}\Delta_0(s')^{-1}U^{-1}U\phi_0(s,a)}{\phi_0(s,a)^\top U^\top\nu_0(s')} \\
&= \frac{\gamma}{1-\gamma}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}\left[\phi_0(s,a)^\top\right] \\
&\quad\cdot\left\{V_{\pi^e}(s')I - \left[\int V_{\pi^e}(s')\Delta_0(s')^{-1}\mathrm{d}s'\right]\left[\int\Delta_0(s')^{-1}\mathrm{d}s'\right]^{-1}\right\}\frac{\Delta_0(s')^{-1}\phi_0(s,a)}{P(s'\mid s,a)}.
\end{aligned}$$

This means if we keep the expected reward and the transition probability unchanged (so the value function remains unchanged), and only change the feature map $\phi(s,a)$ into $\phi_0(s,a)$, then the efficient influence function $\psi_{\text{eff}}(s,a,\varepsilon,s') = \psi_{\text{eff},1}(s,a,\varepsilon) + \psi_{\text{eff},2}(s,a,s')$ remains unchanged. Therefore, without loss of generality, it suffices to consider the case where $\phi(s,a)$'s are standard unit vectors.

In this special case, $\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}[\phi(s,a)]$ is a $d\times 1$ vector with its (s,a)-th component given by $p^{(\infty)}_{\pi^e,\gamma}(s,a)$, and $\Delta(s')$ is a $d\times d$ diagonal matrix with its (s,a)-th diagonal given by $\frac{p_{\pi^b}(s,a)}{\phi(s,a)^\top\nu_0(s')}$. Furthermore, $\int\Delta(s')^{-1}\mathrm{d}s'$ and $\int V_{\pi^e}(s')\Delta(s')^{-1}\mathrm{d}s'$ are also $d\times d$ diagonal matrices, with their (s,a)-th diagonals given by $\frac{1}{p_{\pi^b}(s,a)}$ and $\frac{\mathbb{E}[V_{\pi^e}(s')|s,a]}{p_{\pi^b}(s,a)}$, respectively. Consequently,

the second part of the efficient influence function becomes

$$\psi_{\text{eff},2}(s,a,s') = \frac{\gamma}{1-\gamma}\begin{bmatrix}\cdots & p^{(\infty)}_{\pi^e,\gamma}(s,a) & \cdots\end{bmatrix}\begin{bmatrix}\ddots & & \\ & V_{\pi^e}(s') - \mathbb{E}[V_{\pi^e}(s') \mid s,a] & \\ & & \ddots\end{bmatrix}\begin{bmatrix}0\\ \vdots \\ 0 \\ \frac{1}{p_{\pi^b}(s,a)} \\ 0 \\ \vdots \\ 0\end{bmatrix}$$

$$= \frac{\gamma}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}\left\{V_{\pi^e}(s') - \mathbb{E}[V_{\pi^e}(s') \mid s,a]\right\}.$$

In addition, $\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]$ is a $d \times d$ diagonal matrix with its (s,a)-th diagonal given by $p_{\pi^b}(s,a)\Omega(s,a)^{-1}$, so the first part of the efficient influence function is simply given by

$$\psi_{\text{eff},1}(s,a,\varepsilon) = \frac{1}{1-\gamma}\mathbb{E}_{p^{(\infty)}_{\pi^e,\gamma}}[\phi(s,a)^\top]\left\{\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\phi(s,a)\Omega(s,a)^{-1}\varepsilon$$

$$= \frac{1}{1-\gamma}\begin{bmatrix}\cdots & p^{(\infty)}_{\pi^e,\gamma}(s,a) & \cdots\end{bmatrix}\begin{bmatrix}\ddots & & \\ & \frac{\Omega(s,a)}{p_{\pi^b}(s,a)} & \\ & & \ddots\end{bmatrix}\begin{bmatrix}0\\ \vdots \\ 0 \\ \Omega(s,a)^{-1}\varepsilon \\ 0 \\ \vdots \\ 0\end{bmatrix}$$

$$= \frac{1}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}\left\{r - \mathbb{E}[r \mid s,a]\right\}.$$

Summing them up and applying the Bellman equation

$$Q_{\pi^e}(s,a) = \mathbb{E}[r \mid s,a] + \gamma\mathbb{E}[V_{\pi^e}(s') \mid s,a],$$

we obtain

$$\begin{aligned}\psi_{\text{eff}}(s,a,\varepsilon,s') &= \psi_{\text{eff},1}(s,a,\varepsilon) + \psi_{\text{eff},2}(s,a,s') \\ &= \frac{1}{1-\gamma}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)}{p_{\pi^b}(s,a)}\left\{r + \gamma V_{\pi^e}(s') - Q_{\pi^e}(s,a)\right\},\end{aligned}$$

which concludes our claim.

A.4. Comparison between Efficiency Bounds in Corollary 3.6

Recall from Theorem 2.1 that

$$\mathcal{V}_{\text{np}} = \frac{1}{(1-\gamma)^2}\mathbb{E}\left[w(s,a)^2\left(r + \gamma V(s') - Q(s,a)\right)^2\right],$$

where $w(s,a)$ is the density ratio, $Q(s,a)$ and $V(s')$ are the Q-function and value function with respect to the target policy π^e, respectively. By independence of r and s' given (s,a), this efficiency bound can also be decomposed into the reward and transition parts just like (7)-(8): $\mathcal{V}_{\text{np}} = \mathcal{V}_{\text{np},1} + \mathcal{V}_{\text{np},2}$, where

$$\mathcal{V}_{\text{np},1} = \frac{1}{(1-\gamma)^2}\mathbb{E}\left[w(s,a)^2\left(r - \mathbb{E}[r \mid s,a]\right)^2\right], \quad \mathcal{V}_{\text{np},2} = \frac{\gamma^2}{(1-\gamma)^2}\mathbb{E}\left[w(s,a)^2\left(V(s') - \mathbb{E}[V(s') \mid s,a]\right)^2\right].$$

Below we show that in the linear MDP case, such two terms are respectively larger than their correspondences in the decomposition of $\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2$, as defined in (7)-(8).

Specifically, for the first term we have

$$\begin{aligned}
\mathcal{V}_{\text{np},1} &= \frac{1}{(1-\gamma)^2}\mathbb{E}\left[w(s,a)^2\varepsilon^2\right] = \frac{1}{(1-\gamma)^2}\mathbb{E}\left[\mathbb{E}\left[w(s,a)^2\varepsilon^2 \mid s,a\right]\right] = \frac{1}{(1-\gamma)^2}\mathbb{E}\left[w(s,a)^2\Omega(s,a)\right], \\
\mathcal{V}_1 &= \Phi_{\pi^e,\gamma}^\top \left\{\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\Phi_{\pi^e,\gamma} \\
&= \frac{1}{(1-\gamma)^2}\mathbb{E}[w(s,a)\phi(s,a)^\top]\left\{\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\mathbb{E}[w(s,a)\phi(s,a)].
\end{aligned}$$

Let $\mathbf{a} = \Omega(s,a)^{-1/2}\phi(s,a) \in \mathbb{R}^d, \mathbf{b} = w(s,a)\phi(s,a) \in \mathbb{R}^d, c = w(s,a)\Omega(s,a)^{1/2} \in \mathbb{R}$. For any $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbf{x}^\top(\mathbf{a}\mathbf{a}^\top)\mathbf{x} + 2\mathbf{b}^\top\mathbf{x} + c^2 = (\mathbf{a}^\top\mathbf{x} + c)^2 \geq 0.$$

Taking expectation with respect to (s,a) yields $\mathbf{x}^\top\mathbb{E}[\mathbf{a}\mathbf{a}^\top]\mathbf{x} + 2\mathbb{E}[\mathbf{b}^\top]\mathbf{x} + \mathbb{E}[c^2] \geq 0$ for any $x \in \mathbb{R}^d$, which further leads to

$$\begin{bmatrix}\mathbb{E}[\mathbf{a}\mathbf{a}^\top] & \mathbb{E}[\mathbf{b}] \\ \mathbb{E}[\mathbf{b}^\top] & \mathbb{E}[c^2]\end{bmatrix} \succeq 0 \quad \Longrightarrow \quad \mathbb{E}[c^2] - \mathbb{E}[\mathbf{b}^\top]\left(\mathbb{E}[\mathbf{a}\mathbf{a}^\top]\right)^{-1}\mathbb{E}[\mathbf{b}] \geq 0.$$

By definitions of $\mathbf{a}, \mathbf{b}, c$, the last inquality essentially states that $\mathcal{V}_{\text{np},1} \geq \mathcal{V}_1$, and the difference is at least $\frac{1}{(1-\gamma)^2}$ times the minimum eigenvalue of $\mathbb{E}\left(\begin{bmatrix}\Omega(s,a)^{-1/2}\phi(s,a) \\ \Omega(s,a)^{1/2}w(s,a)\end{bmatrix}\begin{bmatrix}\Omega(s,a)^{-1/2}\phi(s,a)^\top & \Omega(s,a)^{1/2}w(s,a)\end{bmatrix}\right)$.

For the second term, we illustrate $\mathcal{V}_{\text{np},2} \geq \mathcal{V}_2$ and the difference between efficiency bounds via a numerical experiment. We set $\gamma = 0.5, |\mathcal{S}| = 100, |\mathcal{A}| = 1$ and $p_{\pi^b}^{(0)}(s) = p_{\pi^e}^{(0)}(s) = 1/100$. The dimension of the feature map varies from 1 to 100. We consider three types of the feature map: 1). each component of $\phi(s,a)$ is generated randomly from $\text{Exp}(1)$ and then normalized to satisfy $\phi(s,a)^\top\mathbf{1}_d = 1$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$; 2). $\phi(s_j,a) = \mathbf{e}_j$ (the j-th unit vector) for $2 \leq j \leq d$ and $\phi(s,a) = \mathbf{e}_1$ otherwise; 3). the numbers of feature maps $\phi(s,a)$ equaling to $\mathbf{e}_j$ are roughly the same ($\approx 100/d$). We call the three cases "random", "unbalanced", "balanced", respectively. Figure 4 plots the absolute and relative differences between $\mathcal{V}_{\text{np},2}$ and $\mathcal{V}_2$ when the dimension of the feature map d varied from 1 to 100.

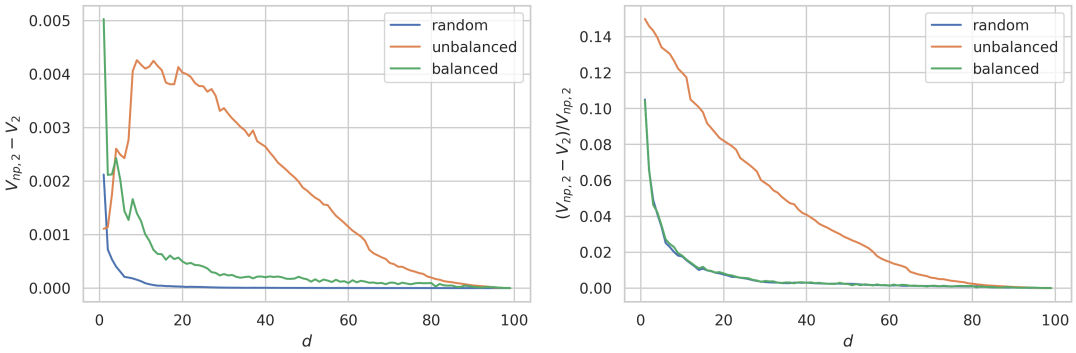

Figure 4. absolute and relative differences between $\mathcal{V}_{\text{np},2}$ and $\mathcal{V}_2$ when the feature map is random, unbalanced and balanced respectively.

It is shown that the absolute difference $\mathcal{V}_{\text{np},2} - \mathcal{V}_2$ is always above 0. As the dimension of the feature map d approaches the size of the state-action space $|\mathcal{S}||\mathcal{A}|$, the difference generally becomes smaller. In addition, the absolute and relative differences are similar in the "random" and "balanced" cases, while apparently larger in the "unbalanced" case. Therefore, it is likely that the efficiency gain in the transition part of the efficiency bound is larger if $\phi(s,a)$ is unbalancedly distributed in $\mathbb{R}^d$.

A.5. Proof of Proposition 3.7

Let $\underline{\sigma} = \lambda_{\min}\left(\mathbb{E}[\phi(s,a)\phi(s,a)^\top]\right)$. We prove that $\mathcal{V}_1 \leq O\left(\frac{1}{\underline{\sigma}(1-\gamma^2)}\right)$ and $\mathcal{V}_2 \leq O\left(\frac{1}{\underline{\sigma}^2(1-\gamma)^4}\right)$.

Due to Assumption 3.1 and $r \in [0, R_{\max}]$, we have $\|\phi(s,a)\|_2 \le 1$ and $\Omega(s,a) \le R_{\max}^2$ and therefore

$$\begin{aligned}
\mathcal{V}_1 &= \frac{1}{(1-\gamma)^2}\mathbb{E}_{p_{\pi^e}^{(\infty)}}[\phi(s,a)^\top]\left\{\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\mathbb{E}_{p_{\pi^e}^{(\infty)}}[\phi(s,a)] \\
&\le \frac{1}{(1-\gamma)^2}\lambda_{\max}\left(\left\{\mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\right) \\
&\le \frac{R_{\max}^2}{(1-\gamma)^2}\lambda_{\max}\left(\left\{\mathbb{E}[\phi(s,a)\phi(s,a)^\top]\right\}^{-1}\right) \\
&= \frac{R_{\max}^2}{\underline{\sigma}(1-\gamma)^2}.
\end{aligned}$$

For any vector $\alpha \in \mathbb{R}^d$ and $\|\alpha\|_2 = 1$, we have

$$\begin{aligned}
\alpha^\top\Delta(s')\alpha &= \sum_{a\in\mathcal{S},a\in\mathcal{A}}\frac{p_{\pi^b}(s,a)}{P(s' \mid s,a)}[\phi(s,a)^\top\alpha]^2 \\
&\ge \frac{\left(\sum_{s\in\mathcal{S},a\in\mathcal{A}} p_{\pi^b}(s,a)[\phi(s,a)^\top\alpha]^2\right)^2}{\sum_{s\in\mathcal{S},a\in\mathcal{A}} P(s' \mid s,a)p_{\pi^b}(s,a)[\phi(s,a)^\top\alpha]^2} \\
&\ge \frac{\left(\alpha^\top\mathbb{E}[\phi(s,a)\phi(s,a)^\top]\alpha\right)^2}{\sum_{s\in\mathcal{S},a\in\mathcal{A}} P(s' \mid s,a)p_{\pi^b}(s,a)[\phi(s,a)^\top\alpha]^2} \\
&\ge \frac{\underline{\sigma}^2}{\sum_{s\in\mathcal{S},a\in\mathcal{A}} P(s' \mid s,a)p_{\pi^b}(s,a)},
\end{aligned}$$

where we have used Cauchy's inequality and $|\phi(s,a)^\top\alpha| \le 1$. Thus, for any vector $\beta \in \mathbb{R}^d$ and $\|\beta\|_2 = 1$, we have

$$\beta^\top\left(\sum_{s'\in\mathcal{S}}\Delta(s')^{-1}\right)\beta \le \sum_{s'\in\mathcal{S}}\frac{\sum_{s\in\mathcal{S},a\in\mathcal{A}} P(s' \mid s,a)p_{\pi^b}(s,a)}{\underline{\sigma}^2} = \frac{1}{\underline{\sigma}^2}.$$

Thus by $V(s') \in [0, R_{\max}/(1-\gamma)]$, we can bound $\mathcal{V}_2$ by

$$\begin{aligned}
\mathcal{V}_2 &= \frac{\gamma^2}{(1-\gamma)^2}\mathbb{E}_{p_{\pi^e}^{(\infty)}}[\phi(s,a)^\top]\left\{P_\Delta(V^2) - P_\Delta(V)P_\Delta(1)^{-1}P_\Delta(V)\right\}\mathbb{E}_{p_{\pi^e}^{(\infty)}}[\phi(s,a)] \\
&\le \frac{\gamma^2}{(1-\gamma)^2}\mathbb{E}_{p_{\pi^e}^{(\infty)}}[\phi(s,a)^\top]P_\Delta(V^2)\mathbb{E}_{p_{\pi^e}^{(\infty)}}[\phi(s,a)] \\
&\le \frac{\gamma^2}{(1-\gamma)^2}\lambda_{\max}\left(P_\Delta(V^2)\right) \\
&\le \frac{\gamma^2}{(1-\gamma)^2}\frac{R_{\max}^2}{\underline{\sigma}^2(1-\gamma)^2} \\
&= \frac{\gamma^2 R_{\max}^2}{\underline{\sigma}^2(1-\gamma)^4},
\end{aligned}$$

which completes the proof.

Remark A.8. Now we give a corresponding upper bound for the efficiency bound $\mathcal{V}_{\text{np}}$ in the tabular case. This case is equivalent to the linear MDP case where $\phi(s,a) = \mathbf{e}_{(s,a)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is a unit vector, as illustrated in Appendix A.3. When the offline dataset is of good coverage, $\underline{\sigma} = \lambda_{\min}\left(\mathbb{E}[\phi(s,a)\phi(s,a)^\top]\right) = \min_{s,a} p_{\pi^b}(s,a) \sim \frac{1}{|\mathcal{S}||\mathcal{A}|}$, so the bound on the

reward term is roughly $\mathcal{V}_{\text{np},1} = O\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}\right)$, and the bound on the transition term is roughly

$$\begin{aligned}
\mathcal{V}_{\text{np},2} &= \frac{\gamma^2}{(1-\gamma)^2}\mathbb{E}\left[w(s,a)^2 \text{var}\left(V(s') \mid s,a\right)\right] \\
&\leq \frac{\gamma^2}{(1-\gamma)^2}\sum_{s,a}\frac{p^{(\infty)}_{\pi^e,\gamma}(s,a)^2}{p_{\pi^b}(s,a)}\text{var}\left(V(s') \mid s,a\right) \\
&\leq \frac{\gamma^2}{(1-\gamma)^2}\max_{s,a}\frac{\text{var}\left(V(s') \mid s,a\right)}{p_{\pi^b}(s,a)} \\
&\leq \frac{\gamma^2}{(1-\gamma)^2}\frac{\max_{s,a}\text{var}\left(V(s') \mid s,a\right)}{\min_{s,a} p_{\pi^b}(s,a)}.
\end{aligned}$$

The denominator is $\min_{s,a} p_{\pi^b}(s,a) = \underline{\sigma}$. For the numerator, Lemma 8 of (Gheshlaghi Azar et al., 2013) indicates that $\max_{s,a}\text{var}\left(V(s') \mid s,a\right)$ (corresponding to $\|\sigma_{V^\pi}\|_\infty$ in the lemma) is roughly of order $O(1/(1-\gamma))$, so we have $\mathcal{V}_{\text{np},2} = O\left(\frac{\gamma^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\right)$. Putting these together we obtain $\mathcal{V}_{\text{np}} = \mathcal{V}_{\text{np},1} + \mathcal{V}_{\text{np},2} = O\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\right)$. Therefore, knowledge of a specific linear MDP structure reduces the efficiency bound roughly from $O(|\mathcal{S}||\mathcal{A}|)$ to $O(d^2)$.

B. Efficient Estimation and Inference in Linear MDPs

B.1. Derivation of the Matrix Form

We show that

$$\begin{aligned}
&v_{\pi^e} = \mathbb{E}_{p^{(0)}_{\pi^e}}[V_{\pi^e}(s)] = \Phi_{\pi^e}^\top(I-\gamma A)^{-1}\omega_0, \quad V_{\pi^e}(s) = \phi_{\pi^e}(s)^\top(I-\gamma A)^{-1}\omega_0, \\
&Q_{\pi^e}(s,a) = \phi(s,a)^\top(I-\gamma A)^{-1}\omega_0, \quad \mathbb{E}_{p^{(\infty)}_{\pi^e}}[\phi(s,a)^\top] = (1-\gamma)\Phi_{\pi^e}^\top(I-\gamma A)^{-1},
\end{aligned}$$

where $\phi_{\pi^e}(s) = \mathbb{E}_{\pi^e}[\phi(s,a) \mid s]$, $\Phi_{\pi^e} = \mathbb{E}_{p^{(0)}_{\pi^e}}[\phi_{\pi^e}(s)]$ and $A = \int \nu_0(s)\phi_{\pi^e}(s)^\top \mathrm{d}s$.

We first rewrite the value function using the Bellman equation, i.e.,

$$\begin{aligned}
V_{\pi^e}(s) &= \mathbb{E}_{\pi^e}[r^{(0)} \mid s^{(0)} = s] + \gamma\mathbb{E}_{\pi^e}[V_{\pi^e}(s^{(1)}) \mid s^{(0)} = s] \\
&= \int \pi^e(a \mid s)\phi(s,a)^\top\omega_0 \mathrm{d}a + \gamma\int \pi^e(a \mid s)\phi(s,a)^\top\nu_0(s')V_{\pi^e}(s')\mathrm{d}a\mathrm{d}s' \\
&= \phi_{\pi^e}(s)^\top\omega_0 + \gamma\phi_{\pi^e}(s)^\top\int \nu_0(s')V_{\pi^e}(s')\mathrm{d}s'. \qquad (24)
\end{aligned}$$

Multiplying both sides by $\nu_0(s)$ from the left and integrate over s, we obtain

$$\begin{aligned}
&\int \nu_0(s)V_{\pi^e}(s)\mathrm{d}s = A\omega_0 + \gamma A\int \nu_0(s')V_{\pi^e}(s')\mathrm{d}s' \\
\iff &\int \nu_0(s)V_{\pi^e}(s)\mathrm{d}s = (I-\gamma A)^{-1}A\omega_0. \qquad (25)
\end{aligned}$$

where we have used the definition of A. Here, the invertibility of $I-\gamma A$ is guaranteed by Assumption 3.2. In fact, since $0 < \gamma < 1$ and

$$\begin{aligned}
\|A\|_2 &= \left\|\int \nu_0(s)\phi(s,a)^\top\pi^e(a \mid s)\mathrm{d}s\mathrm{d}a\right\|_2 \\
&\leq \sup_{a\in\mathcal{A}}\left\|\int \nu_0(s)\phi(s,a)^\top\mathrm{d}s\right\|_2 \\
&\leq 1,
\end{aligned}$$

the matrix $I-\gamma A$ cannot have zero as its eigenvalue, so it is invertible.

Plugging (25) back to the Bellman equation (24), we obtain

$$V_{\pi^e}(s) = \phi_{\pi^e}(s)^\top \omega_0 + \gamma \phi_{\pi^e}(s)^\top (I - \gamma A)^{-1} A \omega_0 = \phi_{\pi^e}(s)^\top (I - \gamma A)^{-1} \omega_0.$$

Consequently, we also have $v_{\pi^e} = \mathbb{E}_{p_{\pi^e}^{(0)}}[V_{\pi^e}(s)] = \Phi_{\pi^e}^\top (I - \gamma A)^{-1} \omega_0$. The forms of $Q_{\pi^e}(s,a)$ and $\mathbb{E}_{p_{\pi^e}^{(\infty)}}[\phi(s,a)^\top]$ can be derived in a similar way.

B.2. Construction of Nuisance Estimates

Here we discuss some possible choices to construct the nuisance estimate $\widehat{\eta} = (\widehat{\omega}, \widehat{\nu}, \widehat{\Omega}, \widehat{\Delta})$. Specifically, we focus on the case where $|\mathcal{S}|$ and $|\mathcal{A}|$ are finite. Estimates below can be extended to the infinite case with nonparametric estimation and/or supervised learning techniques.

Estimating $\widehat{w}$. Note that $r = \phi(s,a)^\top w_0 + \varepsilon$, so we can simply use the OLS estimator as an initial estimate of w_0:

$$\widehat{w} = \left\{ \sum_{i=1}^n \phi(s_i, a_i) \phi(s_i, a_i)^\top \right\}^{-1} \sum_{i=1}^n \phi(s_i, a_i) r_i.$$

Estimating $\widehat{\nu}$. Note that $P(s' \mid s, a) = \phi(s,a)^\top \nu_0(s')$, so we can use the MLE as an initial estimate of ν_0:

$$\widehat{\nu} = \underset{\nu:\ \sum_{s \in \mathcal{S}} \nu(s) = 1_d}{\operatorname{argmax}} \sum_{i=1}^n \log \left(\phi(s_i, a_i)^\top \nu(s_i') \right).$$

An alternative is to use the least squares estimator

$$\widehat{\nu} = \underset{\nu:\ \sum_{s \in \mathcal{S}} \nu(s) = 1_d}{\operatorname{argmin}} \sum_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \left(\phi(s,a)^\top \nu(s') - \widehat{P}(s' \mid s, a) \right)^2,$$

where $\widehat{P}(s' \mid s, a) = \frac{\sum_{i=1}^n \mathbb{I}(s_i = s, a_i = a, s_i' = s')}{\sum_{i=1}^n \mathbb{I}(s_i = s, a_i = a)}$ is the empirical transition probability.

Estimating $\widehat{A}$. In Section 4 we use $A_{\widehat{\nu}} = \int \widehat{\nu}(s) \phi_{\pi^e}(s)^\top \mathrm{d}s = \sum_{s \in \mathcal{S}} \widehat{\nu}(s) \phi_{\pi^e}(s)^\top$ as an estimate of A. This requires summing up $|\mathcal{S}|$ matrices, which may be infeasible particularly when the state space $\mathcal{S}$ is large. As an alternative, the following estimate $\widehat{A}$ leverages the structure of A and lowers the computation complexity. Note that

$$\begin{aligned} \phi(s,a)^\top A &= \int \phi(s,a)^\top \nu_0(s') \phi_{\pi^e}(s')^\top \mathrm{d}s' = \mathbb{E}[\phi_{\pi^e}(s')^\top \mid s, a] \\ \implies \quad A &= \left\{ \mathbb{E}[\phi(s,a) \phi(s,a)^\top] \right\}^{-1} \mathbb{E}[\phi(s,a) \phi_{\pi^e}(s')^\top]. \end{aligned}$$

Therefore we can construct a plug-in estimate for A:

$$\widehat{A} = \left\{ \sum_{i=1}^n \phi(s_i, a_i) \phi(s_i, a_i)^\top \right\}^{-1} \sum_{i=1}^n \phi(s_i, a_i) \phi_{\pi^e}(s_i')^\top.$$

Estimating $\widehat{\Omega}$. $\Omega(s,a)$ is the variance of the reward given the state-action pair is (s,a). We can use the conditional sample average of residual squares as an estimate of Ω:

$$\widehat{\Omega}(s,a) = \frac{\sum_{i=1}^n \mathbb{I}(s_i = s, a_i = a) \left(r_i - \phi(s_i, a_i)^\top \widehat{w} \right)^2}{\sum_{i=1}^n \mathbb{I}(s_i = s, a_i = a)}.$$

Estimating $\widehat{\Delta}$. When the data distribution $p_{\pi^b}(s,a)$ is known and summing over $\mathcal{S} \times \mathcal{A}$ is feasible, we can use $\Delta_{\widehat{\nu}}(s') = \int p_{\pi^b}(s,a) \frac{\phi(s,a) \phi(s,a)^\top}{\phi(s,a)^\top \widehat{\nu}(s')} \mathrm{d}s \mathrm{d}a = \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} p_{\pi^b}(s,a) \frac{\phi(s,a) \phi(s,a)^\top}{\phi(s,a)^\top \widehat{\nu}(s')}$ as an estimate of $\Delta(s')$. Otherwise when $p_{\pi^b}(s,a)$ is unknown, it is reasonable to use the sample average to estimate $\Delta(s')$:

$$\widehat{\Delta}(s') = \frac{1}{n} \sum_{i=1}^n \frac{\phi(s_i, a_i) \phi(s_i, a_i)^\top}{\phi(s_i, a_i)^\top \widehat{\nu}(s')}.$$

B.3. Proof of Theorem 4.1

Proof of Theorem 4.1. We focus on the $K = 2$ case with two folds $\mathcal{D}_1 = \{O_1, \ldots, O_m\}$ and $\mathcal{D}_2 = \{O_{m+1}, \ldots, O_n\}$, where $m = \lfloor n/2 \rfloor$ and $O_i = (s_i, a_i, r_i, s_i')$. Cases with a general K can be proved similarly. Define

$$\widehat{v}_1 = \frac{1}{m} \sum_{i=1}^{m} \left[\psi_0(\widehat{\eta}^{(1)}) + \psi_1(s_i, a_i, r_i; \widehat{\eta}^{(1)}, \mathcal{D}_2) + \psi_2(s_i, a_i, s_i'; \widehat{\eta}^{(1)}) \right],$$
$$\widehat{v}_2 = \frac{1}{n-m} \sum_{i=m+1}^{n} \left[\psi_0(\widehat{\eta}^{(2)}) + \psi_1(s_i, a_i, r_i; \widehat{\eta}^{(2)}, \mathcal{D}_1) + \psi_2(s_i, a_i, s_i'; \widehat{\eta}^{(2)}) \right].$$

We aim to show that

$$\sqrt{m}\{\widehat{v}_1 - v_{\pi^e}\} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \psi_{\text{eff}}(s_i, a_i, r_i, s_i') + o_p(1), \tag{26}$$
$$\sqrt{n-m}\{\widehat{v}_2 - v_{\pi^e}\} = \frac{1}{\sqrt{n-m}} \sum_{i=m+1}^{n} \psi_{\text{eff}}(s_i, a_i, r_i, s_i') + o_p(1), \tag{27}$$

where $\psi_{\text{eff}}(s, a, r, s')$ is defined in Theorem 3.4, so that $\widehat{v}_{\text{LMDP}} = \frac{m\widehat{v}_1 + (n-m)\widehat{v}_2}{n}$ satisfies

$$\sqrt{n}\{\widehat{v}_{\text{LMDP}} - v_{\pi^e}\} = \frac{1}{n} \sum_{i=1}^{n} \psi_{\text{eff}}(s_i, a_i, r_i, s_i') + o_p(1),$$

and the result follows by applying Lemma A.6. Without loss of generality we only prove (26).

For any $\eta' = (\omega', \nu', \Omega', \Delta')$, define $e(\eta') = \mathbb{E}_\eta[\psi_{\text{eff}}(s, a, r, s'; \eta')]$, where $\mathbb{E}_\eta$ is the expectation under the true distribution[4] (parameterized by $\eta = (\omega_0, \nu_0, \Omega, \Delta)$) and $\psi_{\text{eff}}(s, a, r, s'; \eta')$ as the efficient influence function under the distribution parameterized by η' (i.e., replacing $\omega_0, \nu_0, \Omega, \Delta$ with $\omega', \nu', \Omega', \Delta'$ in $\psi_{\text{eff}}(s, a, r, s')$ defined in Theorem 3.4). It is clear that $e(\eta) = 0$ by the mean-zero property of efficient influence functions.

Recall that $\widehat{\eta}^{(1)} = (\widehat{\omega}^{(1)}, \widehat{\nu}^{(1)}, \widehat{\Omega}^{(1)}, \widehat{\Delta}^{(1)})$ is the nuisance estimate based on $\mathcal{D}_2$. We now decompose $\sqrt{m}\{\widehat{v}_1 - v_{\pi^e}\}$ into the following four terms:

$$\begin{aligned}
\sqrt{m}\{\widehat{v}_1 - v_{\pi^e}\} = & \underbrace{\frac{1}{\sqrt{m}} \sum_{i=1}^{m} [\psi_{\text{eff}}(s_i, a_i, r_i, s_i'; \widehat{\eta}^{(1)}) - e(\widehat{\eta}^{(1)})] - \frac{1}{\sqrt{m}} \sum_{i=1}^{m} [\psi_{\text{eff}}(s_i, a_i, r_i, s_i'; \eta) - e(\eta)]}_{A_m} \\
& + \underbrace{\frac{1}{\sqrt{m}} \sum_{i=1}^{m} [\psi_{\text{eff}}(s_i, a_i, r_i, s_i'; \eta) - e(\eta)]}_{B_m} \\
& + \underbrace{\sqrt{m}\{\psi_0(\widehat{\eta}^{(1)}) - v_{\pi^e}\} + \sqrt{m} e(\widehat{\eta}^{(1)})}_{C_m} \\
& + \underbrace{\frac{1}{\sqrt{m}} \sum_{i=1}^{m} [\psi_1(s_i, a_i, r_i; \widehat{\eta}^{(1)}, \mathcal{D}_2) + \psi_2(s_i, a_i, s_i'; \widehat{\eta}^{(1)}) - \psi_{\text{eff}}(s_i, a_i, r_i, s_i'; \widehat{\eta}^{(1)})]}_{D_m}
\end{aligned}$$

[4] $\mathbb{E}_\eta$ is equivalent to $\mathbb{E}$; we add a subscript η only to emphasize its dependence on the true parameter $\eta = (\omega_0, \nu_0, \Omega, \Delta)$.

Bounding A_m. Since $\widehat{\eta}^{(1)}$ is independent of $\mathcal{D}_1$, we have

$$\begin{aligned}\mathbb{E}_\eta[A_m \mid \mathcal{D}_2] &= \frac{1}{\sqrt{m}}\sum_{i=1}^m \Big\{\mathbb{E}_\eta[\psi_{\text{eff}}(s_i,a_i,r_i,s_i';\widehat{\eta}^{(1)}) - \psi_{\text{eff}}(s_i,a_i,r_i,s_i';\eta) \mid \mathcal{D}_2] - [e(\widehat{\eta}^{(1)}) - e(\eta)]\Big\}\\ &= \sqrt{m}\Big\{\mathbb{E}_\eta[\psi_{\text{eff}}(s,a,r,s';\widehat{\eta}^{(1)}) - \psi_{\text{eff}}(s,a,r,s';\eta) \mid \mathcal{D}_2] - [e(\widehat{\eta}^{(1)}) - e(\eta)]\Big\}\\ &= \sqrt{m}\Big\{[e(\widehat{\eta}^{(1)}) - e(\eta)] - [e(\widehat{\eta}^{(1)}) - e(\eta)]\Big\}\\ &= 0.\end{aligned}$$

Furthermore, by Condition 1 of the theorem,

$$\begin{aligned}\mathrm{var}_\eta[A_m \mid \mathcal{D}_2] &= \mathrm{var}_\eta[\psi_{\text{eff}}(s,a,r,s';\widehat{\eta}^{(1)}) - \psi_{\text{eff}}(s,a,r,s';\eta) \mid \mathcal{D}_2]\\ &\le \mathbb{E}_\eta\left[\{\psi_{\text{eff}}(s,a,r,s';\widehat{\eta}^{(1)}) - \psi_{\text{eff}}(s,a,r,s';\eta)\}^2 \mid \mathcal{D}_2\right] \xrightarrow[m\to\infty]{p} 0.\end{aligned}$$

Therefore, by Chebyshev's inequality, we have

$$Q_m := \mathbb{P}_\eta[|A_m| > \delta \mid \mathcal{D}_2] \le \frac{\mathbb{E}_\eta[A_m^2 \mid \mathcal{D}_2]}{\delta^2} \xrightarrow[m\to\infty]{p} 0.$$

Since $|Q_m| \le 1$, by the bounded convergence theorem we get

$$\mathbb{P}_\eta[|A_m| > \delta] = \mathbb{E}_\eta[Q_m] \xrightarrow[m\to\infty]{} 0,$$

which implies $A_m = o_p(1)$.

Bounding B_m. Since $e(\eta) = 0$ and $\psi_{\text{eff}}(s_i,a_i,r_i,s_i';\eta) = \psi_{\text{eff}}(s_i,a_i,r_i,s_i')$, we have

$$B_m = \frac{1}{\sqrt{m}}\sum_{i=1}^m \psi_{\text{eff}}(s_i,a_i,r_i,s_i').$$

Bounding C_m. Recall that $\psi_0(\widehat{\eta}^{(1)}) = \Phi_{\pi^e}^\top(I - \gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)}$, $v_{\pi^e} = \Phi_{\pi^e}^\top(I - \gamma A_{\nu_0})^{-1}\omega_0$, where $A_{\nu_0} = A = \int \nu_0(s)\phi_{\pi^e}(s)^\top \mathrm{d}s$; and

$$\begin{aligned}e(\widehat{\eta}^{(1)}) &= \mathbb{E}_\eta[\psi_{\text{eff}}(s,a,r,s';\widehat{\eta}^{(1)})]\\ &\overset{(1)}{=} \frac{1}{1-\gamma}(1-\gamma)\Phi_{\pi^e}^\top(I - \gamma A_{\widehat{\nu}^{(1)}})^{-1}\Big\{\mathbb{E}_\eta[\phi(s,a)\widehat{\Omega}^{(1)}(s,a)^{-1}\phi(s,a)^\top]\Big\}^{-1}\\ &\quad\cdot \mathbb{E}_\eta[\phi(s,a)\widehat{\Omega}^{(1)}(s,a)^{-1}(r - \phi(s,a)^\top\widehat{\omega}^{(1)})]\\ &\quad + \frac{\gamma}{1-\gamma}(1-\gamma)\Phi_{\pi^e}^\top(I - \gamma A_{\widehat{\nu}^{(1)}})^{-1}\\ &\quad\cdot \mathbb{E}_\eta\left\{\{V_{\pi^e;\widehat{\eta}^{(1)}}(s')I - P_{\widehat{\Delta}^{(1)}}(V_{\pi^e;\widehat{\eta}^{(1)}})P_{\widehat{\Delta}^{(1)}}(1)^{-1}\}\frac{\widehat{\Delta}^{(1)}(s')^{-1}\phi(s,a)}{\phi(s,a)^\top\widehat{\nu}^{(1)}(s')}\right\}\\ &\overset{(2)}{=} \Phi_{\pi^e}^\top(I - \gamma A_{\widehat{\nu}^{(1)}})^{-1}\Big\{\mathbb{E}_\eta[\phi(s,a)\widehat{\Omega}^{(1)}(s,a)^{-1}\phi(s,a)^\top]\Big\}^{-1}\\ &\quad\cdot \mathbb{E}_\eta[\phi(s,a)\widehat{\Omega}^{(1)}(s,a)^{-1}\phi(s,a)^\top](\omega_0 - \widehat{\omega}^{(1)})\\ &\quad + \gamma\Phi_{\pi^e}^\top(I - \gamma A_{\widehat{\nu}^{(1)}})^{-1}\\ &\quad\cdot \int\{V_{\pi^e;\widehat{\eta}^{(1)}}(s')I - P_{\widehat{\Delta}^{(1)}}(V_{\pi^e;\widehat{\eta}^{(1)}})P_{\widehat{\Delta}^{(1)}}(1)^{-1}\}\frac{\widehat{\Delta}^{(1)}(s')^{-1}\phi(s,a)\phi(s,a)^\top\nu_0(s')}{\phi(s,a)^\top\widehat{\nu}^{(1)}(s')}p_{\pi^b}(s,a)\mathrm{d}s\mathrm{d}a\mathrm{d}s'\\ &\overset{(3)}{=} \Phi_{\pi^e}^\top(I - \gamma A_{\widehat{\nu}^{(1)}})^{-1}(\omega_0 - \widehat{\omega}^{(1)})\\ &\quad + \gamma\Phi_{\pi^e}^\top(I - \gamma A_{\widehat{\nu}^{(1)}})^{-1}\underbrace{\int\{V_{\pi^e;\widehat{\eta}^{(1)}}(s') - P_{\widehat{\Delta}^{(1)}}(V_{\pi^e;\widehat{\eta}^{(1)}})P_{\widehat{\Delta}^{(1)}}(1)^{-1}\}\widehat{\Delta}^{(1)}(s')^{-1}\Delta_{\widehat{\nu}^{(1)}}(s')\nu_0(s')\mathrm{d}s'}_{E_m},\end{aligned}$$

where $\mathbb{E}_\eta$ is taken over the randomness of (s,a,r,s') (not the randomness of $\widehat{\eta}^{(1)}$), (1) uses $\mathbb{E}_{p^{(\infty)}_{\pi^e;\widehat{\eta}^{(1)}}}[\phi(s,a)^\top] = (1-\gamma)\Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}$, (2) uses $\mathbb{E}_\eta[r \mid s,a] = \phi(s,a)^\top\omega_0$, and (3) uses the definition of $\Delta_{\widehat{\nu}^{(1)}}$.

We next bound E_m. We let $\Phi_{\pi^b} = \mathbb{E}[\phi(s,a)]$. Then, we have

$$
\begin{aligned}
E_m &\overset{(1)}{=} \int \{V_{\pi^e;\widehat{\eta}^{(1)}}(s') - P_{\widehat{\Delta}^{(1)}}(V_{\pi^e;\widehat{\eta}^{(1)}})P_{\widehat{\Delta}^{(1)}}(1)^{-1}\}\nu_0(s')\mathrm{d}s' \cdot (1+o_p(\alpha_n^\Delta)) \\
&\overset{(2)}{=} \left\{\left(\int \nu_0(s')\phi_{\pi^e}(s')^\top \mathrm{d}s'\right)(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)} - P_{\widehat{\Delta}^{(1)}}(V_{\pi^e;\widehat{\eta}^{(1)}})P_{\widehat{\Delta}^{(1)}}(1)^{-1}1_d\right\}\cdot(1+o_p(\alpha_n^\Delta)) \\
&\overset{(3)}{=} \left\{A_{\nu_0}(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)} - P_{\widehat{\Delta}^{(1)}}(V_{\pi^e;\widehat{\eta}^{(1)}})P_{\widehat{\Delta}^{(1)}}(1)^{-1}1_d\right\}\cdot(1+o_p(\alpha_n^\Delta)) \\
&\overset{(4)}{=} \left\{A_{\nu_0}(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)} - P_{\Delta_{\widehat{\nu}^{(1)}}}(V_{\pi^e;\widehat{\eta}^{(1)}})P_{\Delta_{\widehat{\nu}^{(1)}}}(1)^{-1}1_d\right\}\cdot(1+o_p(\alpha_n^\Delta)) + o_p(\widetilde{\alpha}_n^\Delta) \\
&\overset{(5)}{=} \left\{A_{\nu_0}(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)} - P_{\Delta_{\widehat{\nu}^{(1)}}}(V_{\pi^e;\widehat{\eta}^{(1)}})\Phi_{\pi^b}\right\}\cdot(1+o_p(\alpha_n^\Delta)) + o_p(\widetilde{\alpha}_n^\Delta) \\
&\overset{(6)}{=} \left\{(A_{\nu_0}-A_{\widehat{\nu}^{(1)}})(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)}\right\}\cdot(1+o_p(\alpha_n^\Delta)) + o_p(\widetilde{\alpha}_n^\Delta),
\end{aligned}
$$

where (1) uses the convergence rate of $\|\widehat{\Delta}^{(1)}(s')^{-1}\Delta_{\widehat{\nu}^{(1)}}(s') - I\|_2$, (2) uses $V_{\pi^e;\widehat{\eta}^{(1)}}(s') = \phi_{\pi^e}(s')^\top(I-\gamma A_{\widehat{\nu}^{(1)}})\widehat{\omega}^{(1)}$ and $\int\nu_0(s')\mathrm{d}s' = 1_d$, (3) uses the definition of A_{ν_0}, (4) uses the convergence rate of $\|P_{\widehat{\Delta}^{(1)}}(V_{\pi^e;\widehat{\eta}^{(1)}})P_{\widehat{\Delta}^{(1)}}(1)^{-1} - P_{\Delta_{\widehat{\nu}^{(1)}}}(V_{\pi^e;\widehat{\eta}^{(1)}})P_{\Delta_{\widehat{\nu}^{(1)}}}(1)^{-1}\|_2$, (5) and (6) use the results in Lemma B.1 below.

Combining together, we get

$$
\begin{aligned}
C_m &= \sqrt{m}\{\psi_0(\widehat{\eta}^{(1)}) - v_{\pi^e} + e(\widehat{\eta}^{(1)})\} \\
&= \sqrt{m}\left\{\Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)} - \Phi_{\pi^e}^\top(I-\gamma A_{\nu_0})^{-1}\omega_0 + \Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}(\omega_0-\widehat{\omega}^{(1)})\right\} \\
&\quad + \sqrt{m}\cdot\gamma\Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\left\{(A_{\nu_0}-A_{\widehat{\nu}^{(1)}})(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)}\right\}\cdot(1+o_p(\alpha_n^\Delta)) + o_p(\sqrt{m}\widetilde{\alpha}_n^\Delta) \\
&= \sqrt{m}\cdot\Phi_{\pi^e}^\top\left\{(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1} - (I-\gamma A_{\nu_0})^{-1}\right\}\omega_0 \\
&\quad + \sqrt{m}\cdot\gamma\Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\left\{(A_{\nu_0}-A_{\widehat{\nu}^{(1)}})(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)}\right\}\cdot(1+o_p(\alpha_n^\Delta)) + o_p(\sqrt{m}\widetilde{\alpha}_n^\Delta) \\
&= \sqrt{m}\cdot\gamma\Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}(A_{\widehat{\nu}^{(1)}}-A_{\nu_0})(I-\gamma A_{\nu_0})^{-1}(\omega_0-\widehat{\omega}^{(1)}) \\
&\quad + \sqrt{m}\cdot\gamma\Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}(A_{\nu_0}-A_{\widehat{\nu}^{(1)}})\left\{(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1} - (I-\gamma A_{\nu_0})^{-1}\right\}\widehat{\omega}^{(1)} \\
&\quad + \sqrt{m}\cdot\gamma\Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\left\{(A_{\nu_0}-A_{\widehat{\nu}^{(1)}})(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)}\right\}\cdot o_p(\alpha_n^\Delta) + o_p(\sqrt{m}\widetilde{\alpha}_n^\Delta) \\
&\lesssim \sqrt{m}\{\|A_{\widehat{\nu}^{(1)}}-A_{\nu_0}\|_2\|\widehat{\omega}^{(1)}-\omega_0\|_2 + \|A_{\widehat{\nu}^{(1)}}-A_{\nu_0}\|_2^2 + \|A_{\widehat{\nu}^{(1)}}-A_{\nu_0}\|_2\cdot o_p(\alpha_n^\Delta)\} + o_p(\sqrt{m}\widetilde{\alpha}_n^\Delta) \\
&= o_p(\sqrt{m}\{\alpha_n^\nu\alpha_n^\omega + (\alpha_n^\nu)^2 + \alpha_n^\nu\alpha_n^\Delta + \widetilde{\alpha}_n^\Delta\}) = o_p(1).
\end{aligned}
$$

Bounding D_m. Note that

$$
\begin{aligned}
D_m &= \frac{1}{\sqrt{m}}\sum_{i=1}^m[\psi_1(s_i,a_i,r_i;\widehat{\eta}^{(1)},\mathcal{D}_2) + \psi_2(s_i,a_i,s_i';\widehat{\eta}^{(1)}) - \psi_{\text{eff}}(s_i,a_i,r_i,s_i';\widehat{\eta}^{(1)})] \\
&= \frac{1}{\sqrt{m}}\sum_{i=1}^m \Phi_{\pi^e}^\top(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1} \\
&\quad\cdot\left(\left\{\mathbb{E}_{\mathcal{D}_2}[\phi(s,a)\widehat{\Omega}^{(1)}(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1} - \left\{\mathbb{E}_\eta[\phi(s,a)\widehat{\Omega}^{(1)}(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\right) \\
&\quad\cdot\phi(s_i,a_i)\widehat{\Omega}^{(1)}(s_i,a_i)^{-1}(r_i - \phi(s_i,a_i)^\top\widehat{\omega}^{(1)}),
\end{aligned}
$$

i.e., the only difference between $\psi_1(s_i,a_i,r_i;\widehat{\eta}^{(1)},\mathcal{D}_2) + \psi_2(s_i,a_i,s_i';\widehat{\eta}^{(1)})$ and $\psi_{\text{eff}}(s_i,a_i,r_i,s_i';\widehat{\eta}^{(1)})$ is whether to use empirical/population expectation of the quantity $\phi(s,a)\widehat{\Omega}^{(1)}(s,a)^{-1}\phi(s,a)^\top$. Furthermore, by the central limit theorem,

we have

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\phi(s_i,a_i)\widehat{\Omega}^{(1)}(s_i,a_i)^{-1}(r_i-\phi(s_i,a_i)^\top\widehat{\omega}^{(1)})=O_p(1).$$

If we additionally have

$$\left\|\left\{\mathbb{E}_{\mathcal{D}_2}[\phi(s,a)\widehat{\Omega}^{(1)}(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}-\left\{\mathbb{E}_{\eta}[\phi(s,a)\widehat{\Omega}^{(1)}(s,a)^{-1}\phi(s,a)^\top]\right\}^{-1}\right\|_2=o_p(1),$$

then $D_m=o_p(1)$. Define

$$F_{m,\text{emp}}=\mathbb{E}_{\mathcal{D}_2}[\phi(s,a)\widehat{\Omega}^{(1)}(s,a)^{-1}\phi(s,a)^\top],\quad F_{m,\text{pop}}=\mathbb{E}_{\eta}[\phi(s,a)\widehat{\Omega}^{(1)}(s,a)^{-1}\phi(s,a)^\top].$$

Since $\|F_{m,\text{emp}}^{-1}-F_{m,\text{pop}}^{-1}\|_2\le\|F_{m,\text{emp}}^{-1}\|_2\|F_{m,\text{emp}}-F_{m,\text{pop}}\|_2\|F_{m,\text{pop}}^{-1}\|_2$, $\|F_{m,\text{emp}}^{-1}\|_2=O_p(1)$ and $\|F_{m,\text{pop}}^{-1}\|_2=O_p(1)$, it suffices to show $\|F_{m,\text{emp}}-F_{m,\text{pop}}\|_2=o_p(1)$. In fact, this can be ensured by Condition 4 of the theorem. Since $\{\phi(s,a)\widetilde{\Omega}(s,a)^{-1}\phi(s,a)^\top:\widetilde{\Omega}\in\mathcal{G}_\Omega\}$ is a Glivenko-Cantelli class,

$$\begin{aligned}&\|F_{m,\text{emp}}-F_{m,\text{pop}}\|_2\\&\le\sup_{\widetilde{\Omega}\in\mathcal{G}_\Omega}\left\|\mathbb{E}_{\mathcal{D}_2}[\phi(s,a)\widetilde{\Omega}(s,a)^{-1}\phi(s,a)^\top]-\mathbb{E}_{\eta}[\phi(s,a)\widetilde{\Omega}(s,a)^{-1}\phi(s,a)^\top]\right\|_2\\&\xrightarrow[m\to\infty]{p}0,\end{aligned}$$

which yields the claim.

Finally, combining A_m, B_m, C_m and D_m, we obtain

$$\sqrt{m}\{\widehat{v}_1-v_{\pi^e}\}=A_m+B_m+C_m+D_m=\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\psi_{\text{eff}}(s_i,a_i,r_i,s_i')+o_p(1),$$

so the proof is complete. □

Lemma B.1. *Suppose the conditions in Theorem 4.1 hold, and let $\Phi_{\pi^b}=\mathbb{E}[\phi(s,a)]$. Then*

$$P_{\Delta_{\widehat{\nu}^{(1)}}}(1)\Phi_{\pi^b}=1_d,\quad P_{\Delta_{\widehat{\nu}^{(1)}}}(V_{\pi^e;\widehat{\eta}^{(1)}})\Phi_{\pi^b}=A_{\widehat{\nu}^{(1)}}(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)}.$$

Proof of Lemma B.1. Noting that

$$\Delta_{\widehat{\nu}^{(1)}}(s')\widehat{\nu}^{(1)}(s')=\int p_{\pi^b}(s,a)\frac{\phi(s,a)\phi(s,a)^\top\widehat{\nu}^{(1)}(s')}{\phi(s,a)^\top\widehat{\nu}^{(1)}(s')}\mathrm{d}s\mathrm{d}a=\int p_{\pi^b}(s,a)\phi(s,a)\mathrm{d}s\mathrm{d}a=\Phi_{\pi^b},$$

we have

$$P_{\Delta_{\widehat{\nu}^{(1)}}}(1)\Phi_{\pi^b}=\int\Delta_{\widehat{\nu}^{(1)}}(s')^{-1}\Phi_{\pi^b}\mathrm{d}s'=\int\widehat{\nu}^{(1)}(s')\mathrm{d}s'=1_d,$$

and

$$\begin{aligned}P_{\Delta_{\widehat{\nu}^{(1)}}}(V_{\pi^e;\widehat{\eta}^{(1)}})\Phi_{\pi^b}&=\int\Delta_{\widehat{\nu}^{(1)}}(s')^{-1}\Phi_{\pi^b}V_{\pi^e;\widehat{\eta}^{(1)}}(s')\mathrm{d}s'\\&=\int\widehat{\nu}^{(1)}(s')\phi_{\pi^e}(s')^\top(I-\gamma A_{\widehat{\nu}^{(1)}})\widehat{\omega}^{(1)}\mathrm{d}s'\\&=A_{\widehat{\nu}^{(1)}}(I-\gamma A_{\widehat{\nu}^{(1)}})^{-1}\widehat{\omega}^{(1)}.\end{aligned}$$

□

B.4. Proof of Theorem 5.1

Proof of Theorem 5.1. Recall from (17)-(18) that $\widehat{\mathcal{V}} = \widehat{\mathcal{V}}_1 + \widehat{\mathcal{V}}_2$, where

$$\widehat{\mathcal{V}}_1 = \Phi_{\pi^e}^\top (I - \gamma A_{\widehat{\nu}})^{-1} \left\{ \mathbb{E}_{\mathcal{D}}[\phi(s,a)\widehat{\Omega}(s,a)^{-1}\phi(s,a)^\top] \right\}^{-1} (I - \gamma A_{\widehat{\nu}})^{-\top} \Phi_{\pi^e},$$
$$\widehat{\mathcal{V}}_2 = \gamma^2 \Phi_{\pi^e}^\top (I - \gamma A_{\widehat{\nu}})^{-1} \left\{ P_{\widehat{\Delta}}(V_{\pi^e;\widehat{\eta}}^2) - P_{\widehat{\Delta}}(V_{\pi^e;\widehat{\eta}}) P_{\widehat{\Delta}}(1)^{-1} P_{\widehat{\Delta}}(V_{\pi^e;\widehat{\eta}}) \right\} (I - \gamma A_{\widehat{\nu}})^{-\top} \Phi_{\pi^e}.$$

By Condition 1 of the theorem, we have

$$\begin{aligned} \left\| (I - \gamma A_{\widehat{\nu}})^{-1} - (I - \gamma A)^{-1} \right\|_2 &= \left\| \gamma (I - \gamma A_{\widehat{\nu}})^{-1} (A_{\widehat{\nu}} - A)(I - \gamma A)^{-1} \right\|_2 \\ &\le \gamma \|(I - \gamma A_{\widehat{\nu}})^{-1}\|_2 \|A_{\widehat{\nu}} - A\|_2 \|(I - \gamma A)^{-1}\|_2 \\ &= o_p(1). \end{aligned}$$

We aim to prove $\widehat{\mathcal{V}}_1 \xrightarrow{p} \mathrm{var}\{\psi_{\text{eff},1}(s,a,\varepsilon)\}$ and $\widehat{\mathcal{V}}_2 \xrightarrow{p} \mathrm{var}\{\psi_{\text{eff},2}(s,a,s')\}$, so $\widehat{\mathcal{V}} \xrightarrow{p} \mathcal{V}(v_{\pi^e})$ follows by adding them up. On the one hand, by Condition 2 of the theorem,

$$\begin{aligned} &\left\| \mathbb{E}_{\mathcal{D}}[\phi(s,a)\widehat{\Omega}(s,a)^{-1}\phi(s,a)^\top] - \mathbb{E}[\phi(s,a)\widehat{\Omega}(s,a)^{-1}\phi(s,a)^\top \mid \widehat{\Omega}] \right\|_2 \\ \le &\sup_{\widetilde{\Omega} \in \mathcal{G}_\Omega} \left\| \mathbb{E}_{\mathcal{D}}[\phi(s,a)\widetilde{\Omega}(s,a)^{-1}\phi(s,a)^\top] - \mathbb{E}[\phi(s,a)\widetilde{\Omega}(s,a)^{-1}\phi(s,a)^\top] \right\|_2 \\ = &\, o_p(1). \end{aligned}$$

In addition, by Jensen's inequality and consistency of $\widehat{\Omega}$ in Condition 1 of the theorem, we have

$$\begin{aligned} &\left\| \mathbb{E}[\phi(s,a)\widehat{\Omega}(s,a)^{-1}\phi(s,a)^\top \mid \widehat{\Omega}] - \mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top] \right\|_2 \\ = &\left\| \mathbb{E}\left[\phi(s,a) \left(\widehat{\Omega}(s,a)^{-1} - \Omega(s,a)^{-1} \right) \phi(s,a)^\top \mid \widehat{\Omega} \right] \right\|_2 \\ \le &\, \mathbb{E}\left[\left\| \phi(s,a) \left(\widehat{\Omega}(s,a)^{-1} - \Omega(s,a)^{-1} \right) \phi(s,a)^\top \right\|_2 \mid \widehat{\Omega} \right] \\ = &\, \mathbb{E}\left[\left| \widehat{\Omega}(s,a)^{-1} - \Omega(s,a)^{-1} \right| \left\| \phi(s,a)\phi(s,a)^\top \right\|_2 \mid \widehat{\Omega} \right] \\ = &\, \mathbb{E}\left[\left| \widehat{\Omega}(s,a)^{-1} - \Omega(s,a)^{-1} \right| \|\phi(s,a)\|_2^2 \mid \widehat{\Omega} \right] \\ \le &\, d \cdot \mathbb{E}\left[\left| \widehat{\Omega}(s,a)^{-1} - \Omega(s,a)^{-1} \right| \mid \widehat{\Omega} \right] \\ = &\, o_p(1). \end{aligned}$$

Combining the above two together, we obtain

$$\left\| \mathbb{E}_{\mathcal{D}}[\phi(s,a)\widehat{\Omega}(s,a)^{-1}\phi(s,a)^\top] - \mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top] \right\|_2 = o_p(1).$$

This together with consistency of $(I - \gamma A_{\widehat{\nu}})^{-1}$ yields

$$\widehat{\mathcal{V}}_1 \xrightarrow{p} \Phi_{\pi^e}^\top (I - \gamma A)^{-1} \left\{ \mathbb{E}[\phi(s,a)\Omega(s,a)^{-1}\phi(s,a)^\top] \right\}^{-1} (I - \gamma A)^{-\top} \Phi_{\pi^e} = \mathrm{var}\{\psi_{\text{eff},1}(s,a,\varepsilon)\}. \tag{28}$$

On the other hand, to prove $\widehat{\mathcal{V}}_2 \xrightarrow{p} \mathrm{var}\{\psi_{\text{eff},2}(s,a,s')\}$ it suffices to show

$$P_{\widehat{\Delta}}(1) \xrightarrow{p} P_{\Delta}(1), \quad P_{\widehat{\Delta}}(V_{\pi^e;\widehat{\eta}}) \xrightarrow{p} P_{\Delta}(V_{\pi^e}), \quad P_{\widehat{\Delta}}(V_{\pi^e;\widehat{\eta}}^2) \xrightarrow{p} P_{\Delta}(V_{\pi^e}^2). \tag{29}$$

This along with $(I - \gamma A_{\widehat{\nu}})^{-1} \xrightarrow{p} (I - \gamma A)^{-1}$ yields

$$\widehat{\mathcal{V}}_2 \xrightarrow{p} \gamma^2 \Phi_{\pi^e}^\top (I - \gamma A)^{-1} \left\{ P_{\Delta}(V_{\pi^e}^2) - P_{\Delta}(V_{\pi^e}) P_{\Delta}(1)^{-1} P_{\Delta}(V_{\pi^e}) \right\} (I - \gamma A)^{-\top} \Phi_{\pi^e} = \mathrm{var}\{\psi_{\text{eff},2}(s,a,s')\}. \tag{30}$$

In fact, by Jensen's inequality, Condition 1 of the theorem and Lemma B.2 below, we have

$$\begin{aligned}\|P_{\widehat{\Delta}}(1) - P_{\Delta}(1)\|_2 &= \left\| \int \left(\widehat{\Delta}(s')^{-1} - \Delta(s')^{-1} \right) \mathrm{d}s' \right\|_2 \\ &\leq \int \left\| \widehat{\Delta}(s')^{-1} - \Delta(s')^{-1} \right\|_2 \mathrm{d}s' \\ &= o_p(1),\end{aligned}$$

$$\begin{aligned}\|P_{\widehat{\Delta}}(V_{\pi^e;\widehat{\eta}}) - P_{\Delta}(V_{\pi^e})\|_2 &= \left\| \int \left(\widehat{\Delta}(s')^{-1} V_{\pi^e;\widehat{\eta}}(s') - \Delta(s')^{-1} V_{\pi^e}(s') \right) \mathrm{d}s' \right\|_2 \\ &\leq \left\| \int \left(\widehat{\Delta}(s')^{-1} - \Delta(s')^{-1} \right) V_{\pi^e;\widehat{\eta}}(s') \mathrm{d}s' \right\|_2 \\ &\quad + \left\| \int \Delta(s')^{-1} \left(V_{\pi^e;\widehat{\eta}}(s') - V_{\pi^e}(s') \right) \mathrm{d}s' \right\|_2 \\ &\leq \int \left\| \widehat{\Delta}(s')^{-1} - \Delta(s')^{-1} \right\|_2 |V_{\pi^e;\widehat{\eta}}(s')| \mathrm{d}s' \\ &\quad + \int \left\| \Delta(s')^{-1} \right\|_2 |V_{\pi^e;\widehat{\eta}}(s') - V_{\pi^e}(s')| \mathrm{d}s' \\ &= \int \left\| \widehat{\Delta}(s')^{-1} - \Delta(s')^{-1} \right\|_2 \mathrm{d}s' \cdot \sup_{s' \in \mathcal{S}} |V_{\pi^e;\widehat{\eta}}(s')| \\ &\quad + \int \left\| \Delta(s')^{-1} \right\|_2 \mathrm{d}s' \cdot \sup_{s' \in \mathcal{S}} |V_{\pi^e;\widehat{\eta}}(s') - V_{\pi^e}(s')| \\ &= \sup_{s' \in \mathcal{S}} |V_{\pi^e;\widehat{\eta}}(s')| \cdot o_p(1) + \int \left\| \Delta(s')^{-1} \right\|_2 \mathrm{d}s' \cdot o_p(1) \\ &= o_p(1),\end{aligned}$$

and

$$\begin{aligned}\|P_{\widehat{\Delta}}(V^2_{\pi^e;\widehat{\eta}}) - P_{\Delta}(V^2_{\pi^e})\|_2 &= \left\| \int \left(\widehat{\Delta}(s')^{-1} V_{\pi^e;\widehat{\eta}}(s')^2 - \Delta(s')^{-1} V_{\pi^e}(s')^2 \right) \mathrm{d}s' \right\|_2 \\ &\leq \left\| \int \left(\widehat{\Delta}(s')^{-1} - \Delta(s')^{-1} \right) V_{\pi^e;\widehat{\eta}}(s')^2 \mathrm{d}s' \right\|_2 \\ &\quad + \left\| \int \Delta(s')^{-1} \left(V_{\pi^e;\widehat{\eta}}(s')^2 - V_{\pi^e}(s')^2 \right) \mathrm{d}s' \right\|_2 \\ &\leq \int \left\| \widehat{\Delta}(s')^{-1} - \Delta(s')^{-1} \right\|_2 |V_{\pi^e;\widehat{\eta}}(s')|^2 \mathrm{d}s' \\ &\quad + \int \left\| \Delta(s')^{-1} \right\|_2 |V_{\pi^e;\widehat{\eta}}(s')^2 - V_{\pi^e}(s')^2| \mathrm{d}s' \\ &= \int \left\| \widehat{\Delta}(s')^{-1} - \Delta(s')^{-1} \right\|_2 \mathrm{d}s' \cdot \sup_{s' \in \mathcal{S}} |V_{\pi^e;\widehat{\eta}}(s')|^2 \\ &\quad + \int \left\| \Delta(s')^{-1} \right\|_2 \mathrm{d}s' \cdot \sup_{s' \in \mathcal{S}} |V_{\pi^e;\widehat{\eta}}(s')^2 - V_{\pi^e}(s')^2| \\ &\leq \sup_{s' \in \mathcal{S}} |V_{\pi^e;\widehat{\eta}}(s')|^2 \cdot o_p(1) \\ &\quad + \int \left\| \Delta(s')^{-1} \right\|_2 \mathrm{d}s' \cdot \left(\sup_{s' \in \mathcal{S}} |V_{\pi^e;\widehat{\eta}}(s')| + \sup_{s' \in \mathcal{S}} |V_{\pi^e}(s')| \right) \cdot o_p(1) \\ &= o_p(1),\end{aligned}$$

which prove (29). Finally, combining (28) and (30) yields $\widehat{\mathcal{V}} \xrightarrow{p} \mathcal{V}(v_{\pi^e})$, and by Slutsky's theorem we obtain

$$\mathbb{P}\left(|\widehat{v}_{\text{LMDP}} - v_{\pi^e}| \le z_{1-\alpha/2}\sqrt{\widehat{\mathcal{V}}/n}\right) \to 1-\alpha,$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of a standard normal distribution.

□

Lemma B.2. *Suppose the conditions in Theorem 5.1 hold. Then*

$$\sup_{s'\in\mathcal{S}} |V_{\pi^e;\widehat{\eta}}(s') - V_{\pi^e}(s')| = o_p(1), \quad \sup_{s'\in\mathcal{S}} |V_{\pi^e;\widehat{\eta}}(s')| = O_p(1), \quad \sup_{s'\in\mathcal{S}} |V_{\pi^e}(s')| = O_p(1).$$

Proof of Lemma B.2. For the first we have

$$\begin{aligned}
\sup_{s'\in\mathcal{S}} |V_{\pi^e;\widehat{\eta}}(s') - V_{\pi^e}(s')| &= \sup_{s'\in\mathcal{S}} \left|\phi_{\pi^e}(s')(I-\gamma A_{\widehat{\nu}})^{-1}\widehat{\omega} - \phi_{\pi^e}(s')(I-\gamma A)^{-1}\omega_0\right| \\
&\le \sup_{s'\in\mathcal{S}} \Big\{ \left|\phi_{\pi^e}(s')\left[(I-\gamma A_{\widehat{\nu}})^{-1} - (I-\gamma A)^{-1}\right]\widehat{\omega}\right| \\
&\qquad + \left|\phi_{\pi^e}(s')(I-\gamma A)^{-1}(\widehat{\omega}-\omega_0)\right| \Big\} \\
&\le \sup_{s'\in\mathcal{S}} \Big\{ \|\phi_{\pi^e}(s')\|_2 \left\|(I-\gamma A_{\widehat{\nu}})^{-1} - (I-\gamma A)^{-1}\right\|_2 \|\widehat{\omega}\|_2 \\
&\qquad + \|\phi_{\pi^e}(s')\|_2 \|(I-\gamma A)^{-1}\|_2 \|\widehat{\omega}-\omega_0\|_2 \Big\} \\
&\le d\|\widehat{\omega}\|_2 \cdot o_p(1) + d\|(I-\gamma A)^{-1}\|_2 \cdot o_p(1) \\
&= o_p(1),
\end{aligned}$$

where we use the fact that each component of $\phi_{\pi^e}(s')$ is no more than 1 by Assumption 3.1 and the definition of ϕ_{π^e}.

Similarly, for the second and the third we have

$$\begin{aligned}
\sup_{s'\in\mathcal{S}} |V_{\pi^e;\widehat{\eta}}(s')| &= \sup_{s'\in\mathcal{S}} \left|\phi_{\pi^e}(s')(I-\gamma A_{\widehat{\nu}})^{-1}\widehat{\omega}\right| \\
&\le \sup_{s'\in\mathcal{S}} \|\phi_{\pi^e}(s')\|_2 \|(I-\gamma A_{\widehat{\nu}})^{-1}\|_2 \|\widehat{\omega}\|_2 \\
&\le d\|(I-\gamma A_{\widehat{\nu}})^{-1}\|_2 \|\widehat{\omega}\|_2 \\
&= o_p(1),
\end{aligned}$$

and

$$\begin{aligned}
\sup_{s'\in\mathcal{S}} |V_{\pi^e}(s')| &= \sup_{s'\in\mathcal{S}} \left|\phi_{\pi^e}(s')(I-\gamma A)^{-1}\omega_0\right| \\
&\le \sup_{s'\in\mathcal{S}} \|\phi_{\pi^e}(s')\|_2 \|(I-\gamma A)^{-1}\|_2 \|\omega_0\|_2 \\
&\le d\|(I-\gamma A)^{-1}\|_2 \|\omega_0\|_2 \\
&= o_p(1).
\end{aligned}$$

□

C. Additional Details for Simulation Studies

C.1. Construction of Nuisance Estimators in DRL

Here we discuss how to produce nuisance estimates $\widehat{w}(s,a)$ and $\widehat{Q}_{\pi^e}(s,a)$ so as to construct the DRL estimator. For the Q-function, we know from Appendix B.1 that $Q_{\pi^e}(s,a) = \phi(s,a)^\top (I-\gamma A)^{-1}\omega_0$, so we can use its plug-in estimate

$\widehat{Q}_{\pi^e}(s,a) = \phi(s,a)^\top (I - \gamma A_{\widehat{p}})^{-1}\widehat{\omega}$. For the density ratio, we use the fact that for any test function $f\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, we have

$$\mathbb{E}[w(s,a)\{-f(s,a)+\gamma\tilde{f}(s')\}] + (1-\gamma)\mathbb{E}_{p^{(0)}_{\pi^e}}[\tilde{f}(s)] = 0, \tag{31}$$

where $\tilde{f}(s) = \mathbb{E}_{\pi^e}[f(s,a)\mid s]$. This fact is widely used for estimating the density ratio in the OPE literature (see e.g., Section 3 in (Uehara et al., 2021)).

In our simulation studies, both $|\mathcal{S}|$ and $|\mathcal{A}|$ are finite, so it requires $|\mathcal{S}||\mathcal{A}|$ moment conditions like (31) to ensure the unique characterization of $w(s,a)$. We can choose the test functions to be $f_{s_0,a_0}(s,a) = \mathbb{I}(s=s_0, a=a_0)$, with (s_0,a_0) ranging over $\mathcal{S}\times\mathcal{A}$. In this case, we have $\tilde{f}_{s_0,a_0}(s) = \mathbb{I}(s=s_0)\pi^e(a_0\mid s_0)$. These lead to a natural GMM estimate for $w(s,a)$ by plugging the empirical mean and the test functions into (31):

$$\widehat{w} = \underset{w\colon w(s,a)\geq 0, \mathbb{E}[w(s,a)]=1}{\operatorname{argmin}} \sum_{s\in\mathcal{S},a\in\mathcal{A}} \left[\left(\frac{n_{s,a,s}}{n}\gamma\pi^e(a\mid s) - \frac{n_{s,a}}{n}\right) w(s,a) + (1-\gamma)p^{(0)}_{\pi^e}(s)\pi^e(a\mid s)\right]^2,$$

where $n_{s,a,s} = \#\{1\leq i\leq n\colon s_i=s, a_i=a, s_i'=s\}$ and $n_{s,a} = \#\{1\leq i\leq n\colon s_i=s, a_i=a\}$. The resulting minimizer $\widehat{w}(s,a)$ is chosen to be our nuisance estimate for $w(s,a)$.

C.2. Additional Simulation Results

Here we present additional experiment results omitted in Section 6. We compare the performance of the DM, DRL and our proposed estimator, all using 2-fold (resp. 5-fold) sample splitting, in Figure 5 (resp. Figure 6). Both cases exhibit the superiority of our estimator in aspect of smaller variation in the estimator.

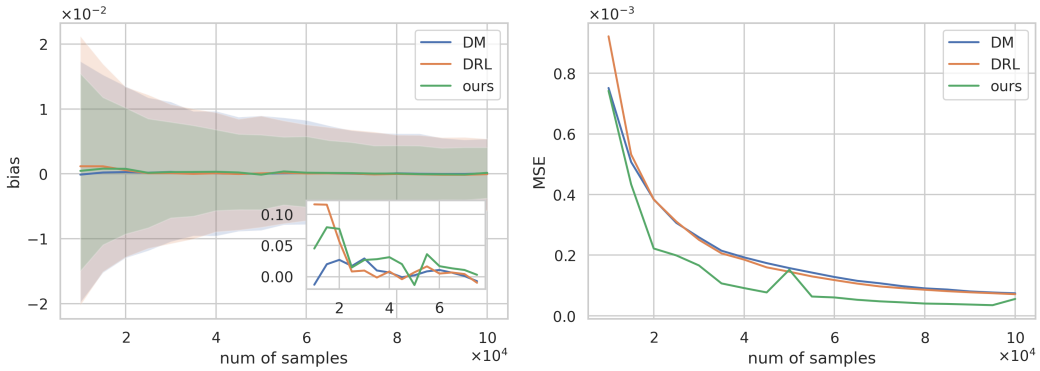

Figure 5. Left: the average, 75-th quantile and 25-th quantile of biases of three estimators. Right: the mean square errors (MSEs) of three estimators.

In addition, we plot in Figure 7 the biases and MSEs of DRL estimators without sample splitting, as well as with 2-fold and 5-fold sample splitting. All three estimators share similar performance. For sample splitting estimators, the 5-fold one has a smaller MSE than the 2-fold one, implying that increasing the number of folds may increase the stability of the final estimator.

D. Auxiliary Results

D.1. Double Reinforcement Learning

Here we provide a rigorous statement of Theorem 2.1 as proposed in (Kallus & Uehara, 2022). The result is composed of two parts.

Theorem D.1 (Theorem 4 of (Kallus & Uehara, 2022)). *Consider the fully nonparametric model for the data distribution* $\mathcal{F}_{np} = \{p\colon p(s,a,r,s') = p_s(s)p_{a|s}(a\mid s)p_{r|s,a}(r\mid s,a)p_{s'|s,a}(s'\mid s,a)\}$. *The efficient influence function and the*

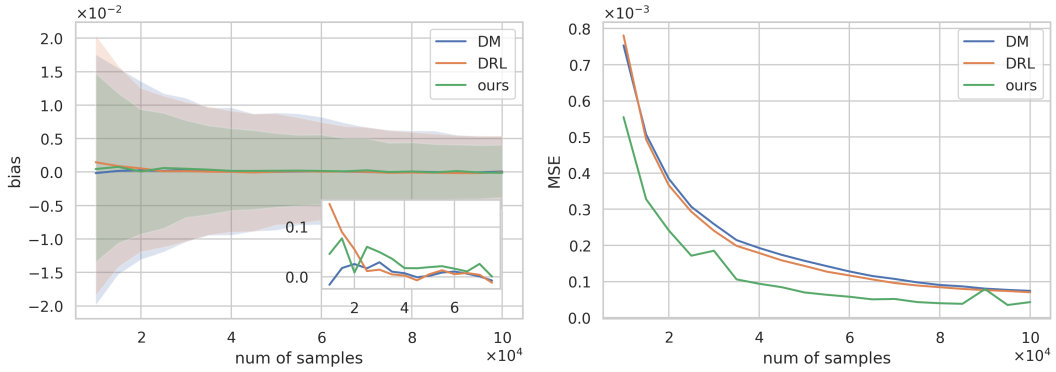

Figure 6. Left: the average, 75-th quantile and 25-th quantile of biases of three estimators. Right: the mean square errors (MSEs) of three estimators.

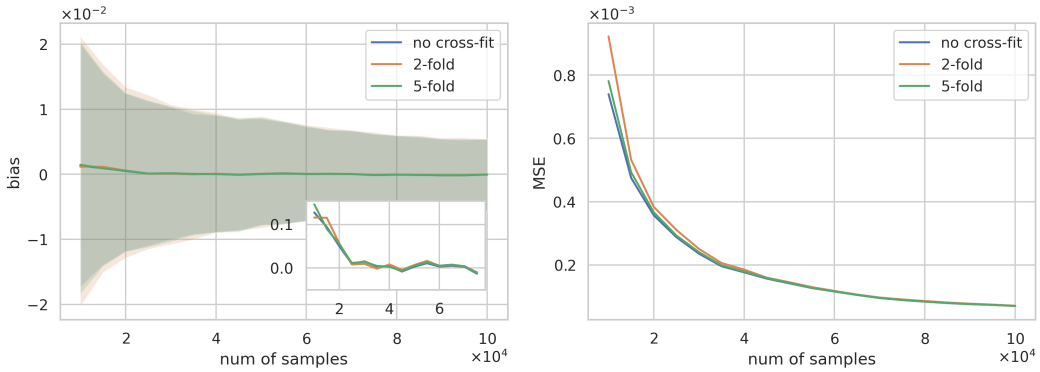

Figure 7. Left: the average, 75-th quantile and 25-th quantile of biases of estimators with/without sample splitting. Right: the mean square errors (MSEs) of estimators with/without sample splitting.

efficiency bound with respect to the model $\mathcal{F}_{np}$ for estimating v_{π^e} are given by

$$\psi_{eff,np}(s,a,r,s') = \frac{1}{1-\gamma} w(s,a)(r + \gamma V_{\pi^e}(s') - Q_{\pi^e}(s,a)),$$
$$\mathcal{V}_{np}(v_{\pi^e}) = \frac{1}{(1-\gamma)^2} \mathbb{E}\left[w(s,a)^2 (r + \gamma V_{\pi^e}(s') - Q_{\pi^e}(s,a))^2\right].$$

Theorem D.2 (Theorem 8 of (Kallus & Uehara, 2022))**.** *Suppose $\widehat{w}(s,a)$ and $\widehat{Q}_{\pi^e}(s,a)$ are some estimates of $w(s,a)$ and $Q_{\pi^e}(s,a)$. Define κ_n^w, κ_n^q such that $\|\widehat{w} - w\|_2 \le \kappa_n^w$ and $\|\widehat{Q}_{\pi^e} - Q_{\pi^e}\|_2 \le \kappa_n^q$. Suppose that*

1. *$w \le C_w$ and $p_{b,s'}(\cdot)/p_{b,s}(\cdot) \le C_{s'}$, where $p_{b,s'}(\cdot)$ and $p_{b,s}(\cdot)$ are marginal densities of $p_{\pi^b}(s,a,r,s')$ with respect to s' and s;*
2. *$0 \le \widehat{Q}_{\pi^e} \le (1-\gamma)^{-1} R_{max}$ and $0 \le \widehat{w} \le C_w$;*
3. *$\kappa_n^w \vee \kappa_n^q = o_p(1)$ and $\kappa_n^w \kappa_n^q = o_p(n^{-1/2})$;*
4. *$\widehat{w} \in \mathcal{F}_w, \widehat{Q}_{\pi^e} \in \mathcal{F}_q$ such that $\log \mathcal{N}(\tau, \mathcal{F}_w, \mathcal{L}_\infty) = O(1/\tau^2)$ and $\log \mathcal{N}(\tau, \mathcal{F}_q, \mathcal{L}_\infty) = O(1/\tau^2)$, where $\mathcal{N}(\tau, \mathcal{F}, \mathcal{L}_\infty)$ is the τ-covering number of $\mathcal{F}$ with respect to $\mathcal{L}_\infty$.*

Then the following estimator is efficient:

$$\widehat{v}_{DR} = \mathbb{E}_{p_{\pi^e}^{(0)}}[\widehat{V}_{\pi^e}(s)] + \frac{1}{n(1-\gamma)} \sum_{i=1}^{n} \widehat{w}(s_i,a_i)(r_i + \gamma \widehat{V}_{\pi^e}(s_i') - \widehat{Q}_{\pi^e}(s_i,a_i)),$$

where $\widehat{V}_{\pi^e}(s)$ is defined in terms of $\widehat{Q}_{\pi^e}(s,a)$ by taking expectation over $a \sim \pi^e(\cdot \mid s)$. In particular, $\sqrt{n}(\widehat{v}_{DR} - v_{\pi^e}) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}_{np}(v_{\pi^e}))$.

D.2. Projection Formula

Lemma D.3 characterizes the explicit formula of the projection of a function to a specific space in conditional moment models.

Lemma D.3 ((Severini et al., 2013), Lemma G.1)**.** *Suppose the random vectors Z and X satisfy the conditional moment restriction $\mathbb{E}[g(Z;\theta_0) \mid X] = 0$, where $\theta_0 \in \mathbb{R}^p$ and $g(Z;\theta_0)$ is a $q \times 1$ vector of functions known up to θ_0. Define*

$$\Omega(X) = \mathbb{E}[g(Z;\theta_0)g(Z;\theta_0)^\top \mid X],\ D(X) = \frac{\partial \mathbb{E}[g(Z;\theta_0) \mid X]}{\partial \theta^\top},\ V = \mathbb{E}[D(X)^\top \Omega(X)^{-1} D(X)].$$

Suppose in addition that $\mathbb{E}\|g(Z;\theta_0)\|^2 \vee \mathbb{E}\|D(X)\|^2 \vee \|\Omega(X)^{-1}\|_\infty < \infty$ and V is invertible. Consider the space

$$\Lambda = \{s(Z,X) \colon \mathbb{E}[s(Z,X) \mid X] = 0,\ \mathbb{E}[g(Z;\theta_0)s(Z,X) \mid X] \in \mathcal{R}(D(X))\} \cap \mathcal{L}_2^0,$$

where $\mathcal{R}(D(X))$ is the column space of $D(X) \in \mathbb{R}^{q \times p}$. Then, for any $h(Z,X) \in \mathcal{L}_2$, it holds that

$$\Pi[h \mid \Lambda] = h - \mathbb{E}[h \mid X] - g^\top \Omega(X)^{-1} \left\{ \mathbb{E}[gh \mid X] - D(X)V^{-1}\mathbb{E}[D(X)^\top \Omega(X)^{-1}\mathbb{E}[gh \mid X]] \right\}.$$

D.3. Glivenko-Cantelli Property

Definition D.4 defines P-Glivenko-Cantelli function classes. We omit "P-" when it refers to the true distribution and makes no confusion. Lemma D.5 gives a sufficient condition for a function class to be P-Glivenko-Cantelli. Many commonly used estimation classes such as *pointwise compact classes*, *smooth function classes* and *Sobolev classes* satisfy this condition (see Chapter 19 of (Van der Vaart, 2000) for details).

Definition D.4 (P-Glivenko-Cantelli, (Van der Vaart, 2000))**.** A class $\mathcal{F}$ of measurable functions $f \colon \mathcal{X} \to \mathbb{R}$ is called P-Glivenko-Cantelli if

$$\|\mathbb{P}_n f - Pf\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \xrightarrow{a.s.} 0,$$

where

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i), \quad Pf = \int f \mathrm{d}P.$$

Lemma D.5 ((Van der Vaart, 2000), Theorem 19.4)**.** *Every class $\mathcal{F}$ of measurable functions such that $N_{[]}(\epsilon, \mathcal{F}, \mathcal{L}_1(P)) < \infty$ for every $\epsilon > 0$ is P-Glivenko-Cantelli, where $N_{[]}(\epsilon, \mathcal{F}, \mathcal{L}_1(P))$ is the ϵ-bracketing number of $\mathcal{F}$ with respect to $\mathcal{L}_1(P)$.*