An Instrumental Variable Approach to Confounded Off-Policy Evaluation

Yang Xu[1] Jin Zhu[2 3] Chengchun Shi[4] Shikai Luo[3] Rui Song[1]

Abstract

Off-policy evaluation (OPE) aims to estimate the return of a target policy using some pre-collected observational data generated by a potentially different behavior policy. In many cases, there exist unmeasured variables that confound the action-reward or action-next-state relationships, rendering many existing OPE approaches ineffective. This paper develops an instrumental variable (IV)-based method for consistent OPE in confounded sequential decision making. Similar to single-stage decision making, we show that IV enables us to correctly identify the target policy's value in infinite horizon settings as well. Furthermore, we propose a number of policy value estimators and illustrate their effectiveness through extensive simulations and real data analysis from a world-leading short-video platform.

1. Introduction

Off-policy evaluation (OPE) estimates the discounted cumulative reward following a given target policy with an offline dataset collected from another (possibly unknown) behavior policy. OPE is important in situations where it is impractical or too costly to directly evaluate the target policy via online experimentation, including robotics (Quillen et al., 2018), precision medicine (Murphy, 2003; Kosorok & Laber, 2019; Tsiatis et al., 2019), economics (Chernozhukov et al., 2018), quantitative social science (Abadie & Cattaneo, 2018), recommendation systems (Li et al., 2010; Kiyohara et al., 2022), etc.

Despite a large body of literature on OPE (see Section 2 for detailed discussions), many of them rely on the assumption of no unmeasured confounders (NUC), excluding the existence of unobserved variables that could potentially confound either the action-reward or action-next-state pair. This assumption, however, can be violated in some real-world applications such as healthcare and technological industries.

Our paper is partly motivated by the need to evaluate the long-term treatment effects of certain app download ads from a short-video platform. At each time, the platform may bid with many other companies to show their own ads to potential consumers. Unmeasured confounding poses a significant challenge in this data generating process. This is because other companies may win the auction and it remains unknown which ad is ultimately shown to the consumer. In addition, if the competitor's ad is displayed, the consumer may download their app instead. This lack of observability violates the no unmeasured confounders assumption, making it difficult to evaluate the effects of the ads consistently.

Recently, IV-based methods have stood out as a powerful approach to account for unmeasured confounding and measurement errors and have been applied in a range of studies (Angrist et al., 1996; Aronow & Carnegie, 2013; Tchetgen & Vansteelandt, 2013; Ogburn et al., 2015; Wang & Tchetgen, 2018; Qiu et al., 2021; Cui & Tchetgen Tchetgen, 2021). However, these methods are typically used in a single-stage setting and cannot be directly applied to general sequential decision making which is commonly encountered in the RL literature.

To fill in this gap, we propose an IV-based approach to OPE in confounded sequential decision making. The advances and contributions of our proposal are multi-fold.

First, to the best of our knowledge, this is one of the first papers to systematically examine the use of IVs for policy evaluation in infinite or long-horizon settings. Our proposal covers a range of models, including Markov decision processes with unmeasured confounders (MDPUCs), high-order MDPs with unmeasured confounders and POMDPs, allowing the Markov assumption to be potentially violated at different levels. Existing IV-based RL approaches are mainly designed for the purpose of policy optimization, not policy evaluation. Moreover, related studies either rely on the Markov assumption (Liao et al., 2021a; Li et al., 2021; Fu et al., 2022) or finite horizon settings with a few decision stages (Chen & Zhang, 2021). This narrows the scope of their findings.

[1]Department of Statistics, North Carolina State University, Raleigh, USA [2]Sun Yat-sen University, China [3]ByteDance, China [4]Department of Statistics, London School of Economics and Political Science, London, UK. Correspondence to: Rui Song <songray@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Second, when specialized to MDPUCs, we develop a doubly robust policy value estimator. This new estimator, as guaranteed by semiparametric theory (Tsiatis, 2006), achieves the efficiency bound and thus provides the most robust and efficient value estimate for OPE in confounded MDPs. Existing semiparametrically efficient estimators designed for MDPs (Kallus & Uehara, 2022) are biased in our setting, due to the existence of unmeasured confounders.

Finally, as illustrated in Section 7.2, our proposal offers valuable insights in helping tech industries to make sequential decisions in online digital advertising to improve consumers' conversion rates.

2. Related Works

2.1. Off-policy Evaluation

Over the past decades, OPE has been thoroughly researched in reinforcement learning (RL, see Uehara et al., 2022b, for an overview). Current estimators can be roughly divided into three categories. The first type is the direct method estimator (DM) which directly constructs the policy value estimator via an estimated Q- or value function (Lagoudakis & Parr, 2003; Le et al., 2019; Feng et al., 2020; Luckett et al., 2020; Hao et al., 2021; Liao et al., 2021b; Chen & Qi, 2022; Shi et al., 2022b). The second type is the importance sampling (IS)-based estimator that uses the (marginal) IS ratio to account for the distributional shift between the target and behavior policies (Thomas et al., 2015; Hallak & Mannor, 2017; Hanna et al., 2017; Liu et al., 2018; Schlegel et al., 2019; Xie et al., 2019; Dai et al., 2020; Zhang et al., 2020). The last type combines DM and IS for robust OPE (Jiang & Li, 2016; Thomas & Brunskill, 2016; Farajtabar et al., 2018; Tang et al., 2020; Uehara et al., 2020; Cai et al., 2021; Shi et al., 2021; Liao et al., 2022; Kallus & Uehara, 2022). However, none of the aforementioned methods can handle unmeasured confounding.

2.2. Unmeasured Confounding

In observational studies, the no unmeasured confounders (NUC) assumption is often violated due to the presence of latent variables. Recently, there has been an increasing focus on developing RL methods in confounded contextual bandits and sequential decision to address this problem. Some related references in confounded contextual bandits include Bareinboim et al. (2015); Sen et al. (2017); Miao et al. (2018); Shi et al. (2020b); Kallus et al. (2021); Xu et al. (2021); Wang et al. (2022). In general sequential settings, existing works can be broadly grouped into three categories. The first category of work relies on the Markov assumption, models the observed data via an MDP with unmeasured confounders (MDPUC, Zhang & Bareinboim, 2016), and utilizes optimal balancing or certain proxy variables to handle the memoryless unobserved confounding (Bennett et al., 2021; Liao et al., 2021a; Wang et al., 2021; Shi et al., 2022c; Fu et al., 2022). The second category uses a confounded partially observable MDP (POMDP) for problem formulation, borrows the idea from proximal causal inference (see e.g., Tchetgen et al., 2020, for an overview) and extends the framework to sequential decision making (Tennenholtz et al., 2020; Bennett & Kallus, 2021; Nair & Jiang, 2021; Miao et al., 2022; Shi et al., 2022a). The last category develops partial identification bounds for policy learning and evaluation based on sensitivity analysis (Kallus & Zhou, 2020; Namkoong et al., 2020; Chen & Zhang, 2021).

2.3. POMDPs

Our work is also closely related to a line of works on policy learning and evaluation in unconfounded POMDPs (Boots et al., 2011; Anandkumar et al., 2014; Guo et al., 2016; Azizzadenesheli et al., 2016; Jin et al., 2020; Hu & Wager, 2021; Kwon et al., 2021). However, all the aforementioned methods are developed under settings without unmeasured confounders and are not directly applicable to our problem. Meanwhile, methods designed for confounded POMDPs require the action to be independent of the observation given the latent state (see e.g., Tennenholtz et al., 2020; Shi et al., 2022a), which are not applicable to settings when the behavior policy depends on both the state and the observation.

3. Preliminaries

To illustrate the idea, we start by working with the MDPUC setup where the Markov assumption is satisfied. Extensions to non-Markov settings will be discussed in Section 6.

Consider a single data trajectory where (S_t, A_t, R_t) denotes the state-action-reward triplet observed at time t. In the context of online digital advertising, both the action and the reward are binary variables. We denote $A_t = 1$ if the ad is indeed exposed to the consumer at time t, and $R_t = 1$ if the consumer is converted, i.e., downloaded our app at time t. S_t is a vector which contains both the consumer's baseline information and the behavioral data (e.g., the number of historical requests of consumers from different media channels). Let U_t denote the unobserved confounders at time t which may affect both the action and reward/next state. In this example, U_t includes the bidding strategies of other companies, as well as the information about the ad that is displayed to the consumer when $A_t = 0$. It is worth noting that although this example assumes R_t to be binary, the entire framework is applicable to settings where $R_t \in \mathbb{R}$.

As we have mentioned in the introduction, the bidding strategies of other companies can impact both the ad shown (A_t) and the consumer's conversion rate (R_t), resulting in a con-

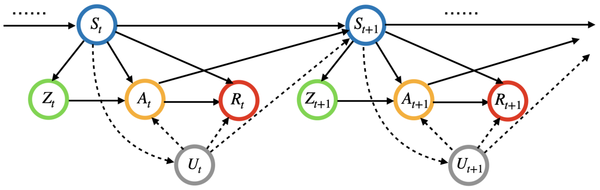

Figure 1. Causal diagram for IV-based MDPUC, where U_t denotes the unmeasured confounders in between $A_t \to (R_t, S_{t+1})$.

founded dataset. To address this problem, we leverage the IV (denoted by Z_t) to infer the long-term treatment effect of ads exposure. Throughout this paper, we assume the instrumental variable and action to be binary, where the value of Z_t depends on whether our company chooses to bid ($Z_t = 1$) at time t or not, and the value of A_t denotes the true ads exposure status to consumers. We will illustrate in Section 7.2 that Z_t defined in this real application is indeed a valid IV.

To summarize, the complete data under the IV-based MDPUC model is given by $\{(S_t, Z_t, A_t, R_t, U_t)\}_{t=0}^T$, where T can be very large or infinite. A causal diagram depiciting the resulting data generating process is given in Figure 1. The observed data contains N i.i.d. trajectories, given by

$$D_i = \{(S_{i,t}, Z_{i,t}, A_{i,t}, R_{i,t})\}_{t=1}^T, \quad i = \{1, \dots, N\}. \quad (1)$$

Let $\pi : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ denote the target policy we wish to evaluate, i.e., $\pi(a|s) = \mathbb{P}^\pi(A_t = a|S_t = s)$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. Likewise, let $b : \mathcal{S} \times \mathcal{U} \times \mathcal{A} \mapsto [0, 1]$ denote the behaviour policy that generates the data in (1). Due to unmeasured confounding, the behavior policy is allowed to depend on both the observed state S and the unobserved confounders U, and thus differs from π.

For a given discounted factor $0 \le \gamma < 1$, we define the value function $V^\pi(s_0)$ as the expected discounted sum of rewards starting from some initial state s_0 under policy π:

$$V^\pi(s_0) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}^\pi(R_t | S_0 = s_0),$$

where the superscript π in $\mathbb{E}^\pi$ denotes the expectation of potential outcome of R_t under policy π. We next define the aggregated value over the initial state distribution $\nu(s_0)$ as

$$\eta^\pi := \mathbb{E}_{S_0 \sim \nu}\big[V^\pi(S_0)\big].$$

Our objective lies in inferring η^π based on (1).

Directly applying existing OPE solutions in Section 2.1 will produce biased policy value estimators in the presence of unmeasured confounders. This is because $\mathbb{E}^\pi(R_t|S_0)$ is generally not equal to $\mathbb{E}(R_t|S_0, A_j \sim \pi, 0 \le j \le t)$. The former corresponds to the potential outcome generated by the causal diagram in Figure 1 with the arrows $\{U_t \to A_t\}_{0 \le t \le T}$ removed, whereas the latter corresponds to the observed outcome generated under the original causal diagram in Figure 1.

4. Identification

We show that the policy value can be consistently identified by Theorem 4.1 below.

Theorem 4.1. ***(Identifiability)*** *Under Assumptions 1-2 detailed in Appendix A, $V^\pi(s_0)$ equals*

$$\sum_{t,\tau_t} \gamma^t r_t \left\{ \prod_{j=0}^{t} p_{r,s}(r_j, s_{j+1}|a_j, z_j, s_j) p_a(a_j|z_j, s_j) c(z_j|s_j) \right\}, \quad (2)$$

where $\tau_t := \{z_j, a_j, r_j, s_{j+1}\}_{j=0}^t$ denotes the collection of all past (z, a, r, s') tuples up to time t, and

$$c(z_t|S_t) = \begin{cases} \dfrac{p_1^A(S_t) - \pi(1|S_t)}{p_1^A(S_t) - p_0^A(S_t)}, & \text{when } z_t = 0 \\ \dfrac{\pi(1|S_t) - p_0^A(S_t)}{p_1^A(S_t) - p_0^A(S_t)}, & \text{when } z_t = 1 \end{cases}, \quad (3)$$

where $p_1^A(S_t) := \mathbb{E}[A_t|Z_t = 1, S_t]$ and $p_0^A(S_t) := \mathbb{E}[A_t|Z_t = 0, S_t]$.

Remark 1. All the functions involved in (2) can be consistently estimated from the observed data, which thus implies the identifiability of $V^\pi(s_0)$. By taking expectation with respect to the initial state distribution, η^π is also identifiable.

Remark 2. The ratio function $c(z|s)$ in (3) measures the discrepancy between the behavior policy and the target π. In the special case where the target policy π equals the behavior policy b, $c(z_t|S_t)$ is reduced to $p_z(z_t|S_t)$, i.e. the conditional probability density/mass function of Z_t given S_t. In this case, it is immediate to see this equation holds since the product in the curly brackets of (2) corresponds to the joint probability density/mass function of the data trajectory up to time t. When $\pi \neq b$, $c(z|s)$ plays a similar role as the important sampling ratio to account for distributional shift.

Remark 3. Assumption 1 consists of a set of IV-related conditions, including IV dependence, IV relevance, exclusion restriction, no additive U–A interactions, etc. Similar assumptions are imposed in single-stage decision making as well (Angrist & Imbens, 1995; Abadie, 2003; Wang & Tchetgen, 2018; Qiu et al., 2021). Assumption 2 is composed of a set of conditional independence assumptions, which require $\{Z_t, U_t, A_t, R_t, S_{t+1}\}$ to be independent of the past data history given the current-stage information. Similar assumptions are imposed in RL when NUC is satisfied (Ertefaie, 2014; Sutton & Barto, 2018; Luckett et al., 2020). To save space, we relegate the details to Appendix A.

Remark 4. The main idea of the proof lies in first applying the conditional independence assumptions (Assumption 2) to decompose the cross-stage identification problem (i.e., $\mathbb{E}^\pi[R_t|S_0]$ for $t \geq 1$) into a sequence of single-stage problems, and then employ the IV-related conditions (Assumption 1) to replace the potential outcome distribution with the observed data distribution. More details about the proof can be found in Appendix E.

5. Estimation

In this section, we discuss how to efficiently estimate η^π under IV-based MDPUCs. We begin with introducing a direct method estimator and a marginal importance sampling estimator. Lastly, we present a doubly robust estimator in the presence of model misspecifications, which can be proven to be semiparametrically efficient.

5.1. Direct Method Estimator

We first introduce the DM estimator which constructs the policy value estimator based on an estimated Q-function. Toward that end, we define the Q-function in IV-based MDPUCs as

$$Q^\pi(s,z,a) = \mathbb{E}^\pi\left[\sum_{k=0}^{\infty} \gamma^t R_{t+k} | S_t = s, Z_t = z, A_t = a\right].$$

Different from the standard Q-function which is a function of the state-action pair only, our Q-function additionally depends on the IV to handle the unmeasured confounding.

Based on Theorem 4.1, it is immediate to see that the value function can be represented as a weighted average of the Q-function, i.e.,

$$V^\pi(s) = \sum_{z,a} c(z|s) p_a(a|z,s) Q^\pi(s,z,a), \tag{4}$$

where $p_a(a|z,s) := \mathbb{P}(A_t = a | Z_t = z, S_t = s)$. Aggregating (4) over the empirical initial state distribution yields the DM estimator, which is given by

$$\widehat{\eta}_{\text{DM}}^\pi = \frac{1}{N} \sum_{i,z,a} \widehat{c}(z|S_{i,0}) \cdot \widehat{p}_a(a|z,S_{i,0}) \widehat{Q}^\pi(S_{i,0},z,a),$$

where $\widehat{c}$, $\widehat{p}_a$ and $\widehat{Q}^\pi$ denote certain consistent estimators for c, p_a and Q^π, respectively. The estimators $\widehat{c}$ and $\widehat{p}_a$ can be computed via supervised learning, and $\widehat{Q}^\pi$ can be obtained by solving a Bellman equation for IV-based MDPUCs. To save space, we relegate the detailed estimation procedures to Appendix B.

5.2. Marginal Importance Sampling Estimator

The second estimator is the marginal importance sampling (MIS) estimator. The traditional sequential IS estimator, constructed based on the product of individual importance sampling ratios at each time, is known to suffer from the curse of horizon (Liu et al., 2018) and becomes very inefficient in the long-horizon settings.

To break the curse of horizon, we borrow ideas from Liu et al. (2018) and define the marginal importance sampling ratio as below:

$$\omega^\pi(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \frac{p_t^\pi(s)}{p_\infty(s)},$$

where p_t^π denotes the probability density/mass function of S_t when the system follows π, and $p_\infty(s)$ to denote the stationary distribution of the stochastic process $\{S_t\}_{t\geq 0}$. Thus, it follows from the change of measure theorem that

$$\eta^\pi = (1-\gamma)^{-1} \mathbb{E}_{S_t \sim p_\infty}[\omega^\pi(S_t) \mathbb{E}^\pi(R_t|S_t)].$$

By applying the IV-based importance sampling trick detailed in Section 4.2 of Wang & Tchetgen (2018), we can represent $\mathbb{E}^\pi(R_t|S_t)$ with the observed data distribution and obtain

$$\eta^\pi = \frac{1}{1-\gamma} \mathbb{E}_{S_t \sim p_\infty, Z_t \sim p_z}\Big[\omega^\pi(S_t) \rho(S_t, Z_t) \mathbb{E}[R_t | Z_t, S_t]\Big],$$

where p_∞ is the stationary distribution of S_t, p_z denotes the sampling distribution of Z_t in observational data, and $\rho(s,z) = c(z|s)/p_z(z|s)$. As such, an MIS estimator can be constructed as below:

$$\widehat{\eta}_{\text{MIS}} = (1-\gamma)^{-1} \frac{1}{\sum_i T_i} \sum_{i,t} \widehat{\omega}^\pi(S_{i,t}) \widehat{\rho}(S_{i,t}, Z_{i,t}) R_{i,t}, \tag{5}$$

where $\widehat{\rho}$ and $\widehat{\omega}^\pi$ denote some consistent estimators of ρ and ω^π, respectively. These estimators can be learned from the observed data, as detailed in Appendix B.

In Formula (5), the expression for the IS estimator consists of two ratios: $\omega^\pi(S_t)$ and $\rho(S_t, Z_t)$. The second ratio $\rho(S_t, Z_t)$ relies on the function c which accounts for the distributional shift, as we have discussed in Remark 2. In the special case where $\pi = b$, we have $\rho(s,z) = 1$.

Finally, let us conclude this section by briefly discussing the drawbacks of the DM and MIS estimators. Both estimators may be seriously biased due to model misspecifications. Specifically, the consistency of DM requires correct specification of c, p_a and Q^π whereas the consistency of MIS requires correct specification of the two ratio functions. In the next section, we will develop a doubly robust (DR) estimator that combines the strength of both estimators.

5.3. Our Proposal

We begin by deriving the efficient influence function (EIF) for η^π, which corresponds to the canonical gradient of a statistical estimand and plays a central role in constructing doubly robust (DR) and semiparametrically efficient estimators (Tsiatis, 2006). The idea of using EIF to develop efficient estimators has been widely used in the statistics and machine learning literature (see e.g., Wang & Tchetgen, 2018; Kallus & Uehara, 2022).

Theorem 5.1. ***(Efficient Influence Function)*** *The EIF for $\eta^\pi = \mathbb{E}_{S_0\sim\nu}[V^\pi(S_0)]$ is given by*

$$\begin{aligned} EIF_{\eta^\pi} = {} & (1-\gamma)^{-1}\omega^\pi(S_t)\Big[\rho(S_t,Z_t)\Big\{Y_t \\ & - \mathbb{E}\big[Y_t\big|Z_t,S_t\big] - \big(A_t - \mathbb{E}\left[A_t\big|Z_t,S_t\right]\big)\cdot\Delta(S_t)\Big\} \\ & + \sum_{z_t} c(z_t|S_t)\cdot\mathbb{E}[R_t|z_t,S_t]\Big] - \eta^\pi, \end{aligned} \tag{6}$$

where $\Delta(S_t)$ is defined as the cumulative conditional Wald estimand (cumulative CWE), where

$$\Delta(S_t) = \frac{\mathbb{E}\left[Y_t\big|Z_t=1,S_t\right] - \mathbb{E}\left[Y_t\big|Z_t=0,S_t\right]}{\mathbb{E}\left[A_t\big|Z_t=1,S_t\right] - \mathbb{E}\left[A_t\big|Z_t=0,S_t\right]},$$

and $Y_t := R_t + \gamma\cdot V^\pi(S_{t+1})$.

Remark 5. The classical CWE plays a key role in identifying the conditional average treatment effect in single-stage decision making. In MDPUCs, we extend the original definition by using Y_t to account for the long-term offline causal effect of executing policy π. When the discounted factor $\gamma = 0$, cumulative CWE will degenerate to the classical CWE.

Remark 6. We notice that a recent concurrent work by Fu et al. (2022) also developed a DR estimator in IV-based MDPUCs. However, their estimator is not constructed based on the EIF, which is less efficient compared to our proposed DR estimator that will be introduced below.

Define $D_t = (S_t, Z_t, A_t, R_t, S_{t+1})$. Based on the result of Theorem 5.1, we propose a DR estimator $\widehat{\eta}_{\text{DR}}$ for aggregated value η^π, given by

$$\widehat{\eta}_{\text{DR}} = \widehat{\eta}^\pi_{\text{DM}} + (NT)^{-1}\sum_{i,t}\widehat{\phi}(D_{i,t}), \tag{7}$$

where $\widehat{\phi}$ denotes some plug-in estimator for the augmentation function ϕ:

$$\begin{aligned} \phi(D_t) = {} & (1-\gamma)^{-1}\omega^\pi(S_t)\rho(S_t,Z_t)\Big\{Y_t \\ & - \mathbb{E}\big[Y_t\big|Z_t,S_t\big] - \big(A_t - \mathbb{E}\left[A_t\big|Z_t,S_t\right]\big)\cdot\Delta(S_t)\Big\}. \end{aligned} \tag{8}$$

According to (7), the proposed estimator is essentially the sum of the DM estimator and an estimated augmentation function $\widehat{\phi}$ which offers additional protection to the final estimator against potential model misspecifications of Q^π. To compute $\widehat{\phi}$, we need to estimate ω^π, ρ, $\mathbb{E}\big[Y_t\big|Z_t,S_t\big]$, p_a and Δ, or equivalently, ω^π, p_z, p_a and Q^π. Since $\mathbb{E}\big[Y_t\big|Z_t,S_t\big] = \sum_{a_t} p_a(a_t|S_t,Z_t)\cdot Q^\pi(S_t,Z_t,a_t)$, Δ and ρ can be determined by p_z, p_a and Q^π. We will discuss the estimation details of these nuisance functions in Section B.

Our final estimator $\widehat{\eta}_{\text{DR}}$, as shown in (7), enjoys the double robustness property. Firstly, recall that the consistency of $\widehat{\eta}_{\text{DM}}$ relies on the correct specification of p_a and Q^π. When both are correctly specified, so are $\mathbb{E}[Y_t|Z_t,S_t]$ and $\mathbb{E}[A_t|Z_t,S_t]$. As such, it is immediate to see that the augmentation term is mean zero regardless of whether the two IS ratios are correctly specified or not. Therefore, the DR estimator is consistent.

Secondly, when the two IS ratios and p_a are correctly specified, it can be shown that no matter whether Q^π is correctly specified or not, we have

$$\begin{aligned} & \mathbb{E}\left[\widehat{\eta}^\pi_{\text{DM}}\right] + (1-\gamma)^{-1}\mathbb{E}\Big[\omega^\pi(S_t)\cdot\rho(S_t,Z_t)\cdot\Big\{\gamma\widehat{V}^\pi(S_{t+1}) \\ & \qquad - \sum_{a_t} p_a(a_t|Z_t,S_t)\widehat{Q}^\pi(S_t,Z_t,a_t)\Big\}\Big] = 0, \end{aligned}$$

where $\widehat{V}^\pi$ depends on $\widehat{Q}^\pi$ through (4). [See Equation (26) in Appendix for details.] It follows that the DR estimator becomes equivalent to the MIS estimator with correctly specified IS ratios

$$(NT)^{-1}\sum_{i,t}(1-\gamma)^{-1}\omega^\pi(S_{i,t})\rho(S_{i,t},Z_{i,t})\cdot R_{i,t},$$

and is thus consistent.

We empirically verify the doubly robustness property in Figure 2. In particular, we apply the proposed method to a toy numerical example detailed in Appendix C.1. It can be seen that the relative absolute bias and MSE of the proposed estimator are fairly small when one set of the models are correctly specified. To the contrary, the resulting estimator is seriously biased when both sets of models are misspecified.

The following theorem states that $\widehat{\eta}_{\text{DR}}$ is not only doubly robust, but semiparametrically efficient as well (e.g., it achieves the minimum variance, i.e., efficiency bound, among all regular and asymptotically linear estimators).

Theorem 5.2. *Suppose that the nuisance function classes are bounded and belong to VC type classes (Van Der Vaart et al., 1996) with VC indices upper bounded by $v = O(N^k)$ for some $0 \le k < 1/2$. Define two model classes as below:*

$\mathcal{M}_1$: $Q^\pi(s,z,a)$ is correctly specified.

$\mathcal{M}_2$: $p_z(z|s)$ and $\omega^\pi(s)$ are correctly specified.

Suppose $p_a(a|s,z)$ is always correctly specified. Then

(a) as long as either $\mathcal{M}_1$ or $\mathcal{M}_2$ holds, $\widehat{\eta}_{DR}$ is a consistent estimator of η^π;

(b) when $\widehat{Q}^{\pi}$, $\widehat{p}_a$, $\widehat{p}_z$ and $\widehat{\omega}^{\pi}$ converge in L_2 norm (see Appendix G.2 for the detailed definition) to their oracle values at a rate of $o(N^{-\alpha})$ with $\alpha \geq 1/4$, we have

$$\sqrt{N}(\widehat{\eta}_{DR} - \eta^{\pi}) \xrightarrow{d} \mathcal{N}(0, \sigma_T^2),$$

where σ_T^2 is the efficiency bound of η^{π}, given by

$$Var\Big\{V^{\pi}(S_0)\Big\} + \frac{1}{T^2}\sum_{t=1}^{T} Var\Big\{\phi(D_t)\Big\}. \tag{9}$$

Remark 7. Theorem 5.2(a) proves the doubly robustness property and (b) proves the semiparametric efficiency. In addition, (b) also establishes the asymptotic normality of $\widehat{\eta}_{\mathrm{DR}}$, based on which the following Wald-type confidence interval (CI) can be constructed for η^{π},

$$\Big[\widehat{\eta}_{\mathrm{DR}} \pm z_{\alpha/2}\frac{\widehat{\sigma}_T}{\sqrt{N}}\Big],$$

where $\widehat{\sigma}_T^2$ is a sampling variance estimator of σ_T^2.

Remark 8. The rate conditions in Theorem 5.2(b) are rather lenient. In fact, for most supervised learning algorithms that exist, the convergence rate is typically $(NT)^{-\alpha}$ for some $\alpha \leq 1/2$. As a result, our condition is automatically met when $\alpha > 1/4$. Furthermore, even in cases where $\alpha \leq 1/4$, this condition can be satisfied when T is relatively large compared to N.

Remark 9. It can be seen from (9) that the semiparametric efficiency bound σ_T^2 generally decays with T, as we have more data for policy value estimation. In particular, as $T \to \infty$, the variance of the augmentation term will vanish, resulting the variance bound to be reduced to $\mathrm{Var}[V^{\pi}(S_0)]$.

Corollary 5.3. *Define another doubly robust estimator $\widehat{\eta}_{DR} = \widehat{\eta}^{\pi}_{DM} + (NT)^{-1}\sum_{i,t}\widehat{\psi}(D_{i,t})$ with a new augmentation term*

$$\psi(D_t) = (1-\gamma)^{-1}\omega^{\pi}(S_t)\rho(S_t, Z_t)\Big\{Y_t - \mathbb{E}\big[Y_t \big| Z_t, S_t\big]\Big\}.$$

After removing the assumption that p_a is correctly specified while maintaining the other assumptions unchanged in Theorem 5.2, all of the results in 5.2(a)-(b) remain valid.

In Equation (7), the only term that involves the estimation of p_a is $\big(A_t - \mathbb{E}\left[A_t \middle| Z_t, S_t\right]\big) \cdot \Delta(S_t)$. If we do not estimate p_a (or $\mathbb{E}\left[A_t \middle| Z_t, S_t\right]$) and replace it with its unbiased observation A_t, this term will disappear. As such, the original augmentation term $\phi(D_t)$ is reduced to $\psi(D_t)$, which does not explicitly rely on p_a. However, the estimator in Corollary 5.3 is no longer semiparametrically efficient. To summarize, there exsits a trade-off between the estimation of nuisance functions and the efficiency of the final estimator. The proof, similar to Theorem 5.2, is omitted for brevity.

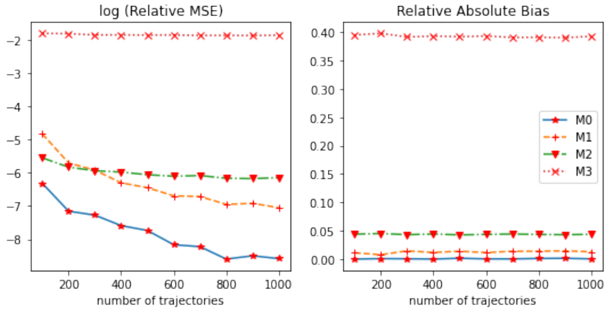

Figure 2. The logarithmic relative MSEs (left panel) and relative absolute biases (right panel) comparison under different model specifications. Specifically, the blue solid line depicts the estimator where the two set of models $\mathcal{M}_1$ and $\mathcal{M}_2$ are correctly specified. The yellow dashed and green dash-dotted lines depict the estimators where one set of the models is correctly specified and the other set misspecified. The red dotted line depicts the estimator where both set of models are misspecified. More details about the data generating process are provided in Appendix C.1.

6. Extensions to Non-Markov Settings

Our proposal in Section 5.3 relies on the set of conditional independence assumptions imposed in Assumption 2. In particular, it requires the states to satisfy the Markov assumption, yielding a memoryless unobserved confounding condition (Kallus & Zhou, 2020). This assumption essentially excludes the existence of directed edges from past observed data or U_{t-1} to U_t in Figure 1 and is likely to be violated in practice. In this section, we discuss two potential relaxations of Assumption 2 to accommodate non-Markov settings. Throughout this section, we will use O_t (instead of S_t) to denote the time-varying observation measured at time t due to the violation of Markovianity.

6.1. High-order MDPs with Unmeasured Confounders

One approach to relax Markov assumption is to impose a high-order memoryless unobserved confounding condition. Specifically, a kth order memoryless unobserved confounding assumption requires U_t to be conditionally independent of the past data history (including $\{U_j\}_{j<t}$) given O_t and the observation-IV-action triplets collected from time $t-k+1$ to $t-1$. When $k = 1$, high-order MDPs will reduce to the memoryless unobserved confounding case. When $k \geq 2$, it allows for the conditional dependence of U_t on the observed data history.

A key observation is that, under the kth order memoryless unobserved confounding assumption, the system forms a kth order MDP with unmeasured confounders. Specifically, let S_t denote the union of O_t and the observation-IV-action triplets collected from time $t-k+1$ to $t-1$. By doing so, the newly-defined state satisfies the Markov assumption, i.e., S_t is independent of the past data history given

$(S_{t-1}, Z_{t-1}, A_{t-1})$. As such, our proposal developed in Section 5.3 can be directly applied here to address the kth order MDPUC.

6.2. Partially Observable MDP

To further relax the high-order memoryless assumption, the second approach is to adopt an IV-based POMDP model $\mathcal{M} = \langle \mathcal{S}, \mathcal{O}, \mathcal{Z}, \mathcal{A}, \mathcal{R} \rangle$ for policy evaluation. Here, $\mathcal{S}, \mathcal{O}, \mathcal{Z}, \mathcal{A}, \mathcal{R}$ denote the spaces of latent states, observed features, IVs, actions and rewards, respectively. At a given time, suppose the environment is in a state $S \in \mathcal{S}$. Although S is not directly observable, we have access to an observation $O \sim p_o(\cdot|S) \in \mathcal{O}$. An IV $Z \sim p_z(\cdot|O)$ is generated whose distribution is independent of S. Next, based on the action $A \in \mathcal{A}$ of the agent, the environment responds by providing an immediate reward R and transitioning to a new state S'. Since A, R and S' are all allowed to depend on S, the dataset we observed is thus confounded. To proceed, we further denote H and F as the multi-step history and future observations, given by

$$H = (O_{t-M_H:t-1}, A_{t-M_H:t-1})$$
$$F = (O_{t:t+M_F-1}, A_{t:t+M_F-2}),$$

where M_H and M_F are two positive integers denoting the number of steps tracing back or forward.

As discussed in Section 2.3, several methods have been developed in the literature to handle POMDPs. Here, we extend the proposal developed by Uehara et al. (2022a) to IV-based POMDPs to deal with confounders.

For illustration purpose, we will focus on evaluating memoryless target policies $\pi : \mathcal{O} \to \mathcal{A}$, but the entire framework can easily be extended to accommodate M-memory policies where the decision rule depends on the last M observations.

In IV-based POMDPs, the Q-function defined in Section 5.1 is not directly estimable since the state is not observable. However, due to the temporal dependence, the multi-step history and future observations contain rich information to infer the latent state. These variables serve as proxies for policy value identification. Toward that end, we define a future-dependent Q-function g_Q as the solution to the following conditional moment equation:

$$\mathbb{E}\Big\{R + \gamma \sum_{z,a} c(z|O')p_a(a|z,O')g_Q(F',z,a) - g_Q(F,Z,A)\big|H,Z,A\Big\} = 0,$$

where O' and F' denote the next-step observation and the next-step future, respectively. Intuitively, g_Q can be viewed as a projection of the Q-function onto the multi-step future. This idea is similar to spectral learning in predictive state representations (PSRs), in which future information is employed to infer the latent state.

The following theorem shows that the policy value can be consistently identified based on g_Q.

Theorem 6.1. *Suppose the following three conditions hold:*

1. *There exists a future-dependent Q function g_Q.*
2. *Invertibility: for any $g : \mathcal{S} \times \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$, if $\mathbb{E}[g(S,Z,A)|H,Z,A] = 0$, then $g(S,Z,A) = 0$, a.s..*
3. *Overlap condition: $|c(Z|O)| < \infty$, a.s..*

Then for any g_Q, we have

$$\eta^\pi = \mathbb{E}_{F\sim\nu_F}\Big[\sum_{z,a} c(z|O)p_a(a|z,O)g_Q(F,z,a)\Big], \quad (10)$$

where ν_F denotes the initial future distribution.

Remark 9. The first two conditions require the cardinality of the future and the history to be at least greater than or equal to the latent state, respectively. These conditions are weaker than requiring the cardinality of the observation to be greater than or equal to the latent state, which is needed in confounded POMDPs without IVs (Nair & Jiang, 2021).

Next, we develop a minimax learning approach to estimate η^π from the observed data. According to the result of Theorem 6.1, as long as we can learn g_Q from the data, a direct method estimator can be naturally constructed by Equation (10). To address so, we consider the following loss function

$$\mathcal{L}(q,\xi) := \Big\{R + \gamma \sum_{z,a} c(z|O')p_a(a|z,O')q(F',z,a) - q(F,Z,A)\Big\}\xi(H,Z,A)$$

for any functions q and ξ. Given some prespecified function classes $q \in \mathcal{Q}$ and $\xi \in \Xi$, we can solve the following minimax problem to obtain an estimator for g_Q:

$$\widehat{g}_Q = \arg\min_{q\in\mathcal{Q}}\max_{\xi\in\Xi} \mathbb{E}_{\mathcal{D}}\Big[\mathcal{L}(q,\xi) - 0.5\lambda\xi^2(H)\Big] + 0.5\alpha'\|q\|^2_{\mathcal{Q}} - 0.5\alpha\|\xi\|^2_{\Xi},$$

where $\|\cdot\|^2_{\mathcal{Q}}$ and $\|\cdot\|^2_{\Xi}$ are certain function norms defined on the spaces of $\mathcal{Q}$ and Ξ, and λ, α and α' are some positive constants. Closed-form solutions are available when using reproducing kernel Hilbert spaces or linear models to parameterize $\mathcal{Q}$ and Ξ (Uehara et al., 2020). Given $\widehat{g}_Q$, $\widehat{c}$ and $\widehat{p}_a$, the resulting DM estimator under POMDP is given by

$$\widehat{\eta}^\pi = \frac{1}{N}\sum_{i=1}^{N}\Big[\sum_{z,a}\widehat{c}(z|O_{i,0})\widehat{p}_a(a|z,O_{i,0})\widehat{g}_Q(F_{i,0},z,a)\Big]. \quad (11)$$

Algorithm 1 Model Selection for IV-based confounded OPE

```
Input: Data trajectories $\{D_i\}_{1\leq i\leq n}$, parameter $K$.
for $k = 1$ to $K$ do
    Apply forward-backward learning procedure in Algorithm 1 of Shi et al. (2020a).
    if $H_0$ is not rejected then
        Conclude the system follows a $k$-th order MDP.
        Apply Section 6.1 and (7) to estimate $\eta^\pi$; Break.
    end if
end for
Conclude the system is most likely a POMDP.
Apply Section 6.2 and (11) to estimate $\eta^\pi$.
```

6.3. Model Selection

So far, we have discussed two approaches to relax the memoryless unobserved confounding assumption, one with the high-order memoryless assumption and the other with the POMDP formulation. These assumptions are not directly testable, since they rely on the unmeasured confounders. However, as commented in Section 6.1, under the kth order memoryless assumption, the observed data satisfy a kth order Markov assumption. When $k = \infty$, this data process becomes a POMDP. This motivates us to apply the sequential testing procedure developed by Shi et al. (2020a) for model selection. Specifically, we consider a hypothesis testing problem where

$$H_0 : \text{The system follows an MDP, v.s.}$$
$$H_1 : \text{The system is a high-order MDP or POMDP.}$$

By implementing the forward-backward learning procedure, one can test the kth order MDP assumption for any given $k \in \{1, \dots, K\}$. We detail the testing procedure in Algorithm 1.

7. Numerical Studies

In this section, we will conduct detailed comparisons between our estimator and other state-of-the-art methods for OPE estimation under MDPs via synthetic data (Section 7.1) and real-world data (Section 7.2). The source code is available on github: https://github.com/YangXU63/IVMDP.

7.1. Simulations

In this section, we compare the proposed estimator in Section 5.3 (denoted by IVMDP) against several baseline methods that ignore the unmeasured confounding.

Data generating process. The observed data consists of $N = 1000$ trajectories, each with $T = 100$ time points. We consider a two-dimensional state variable $S_t = (S_{t,1}, S_{t,2})$ whose initial distribution is given by $\mathcal{N}(\mathbf{0}_2, I_2)$ where I_2 denotes a two-dimensional identity matrix. The unmeasured confounders $\{U_t\}_t$ follow i.i.d. Rademacher distributions. Both the IV and the action are binary. At each time, they satisfy $\mathbb{P}(Z_t = 1|S_t) = \text{sigmoid}(S_{t,1} + S_{t,2})$ and $\mathbb{P}(A_t = 1|S_t, Z_t, U_t) = \text{sigmoid}\{S_{t,1} + S_{t,2} + 2Z_t + 0.5U_t\}$, respectively. Finally, the reward and next-state are generated as follows: $R_t = S_{t,1} + S_{t,2} + 2A_t + 2.5U_t$, $S_{t+1,1} = S_{t1} + 0.5U_t + A_t - 0.5$, $S_{t+1,2} = S_{t1} - 0.5U_t - A_t + 0.5$.

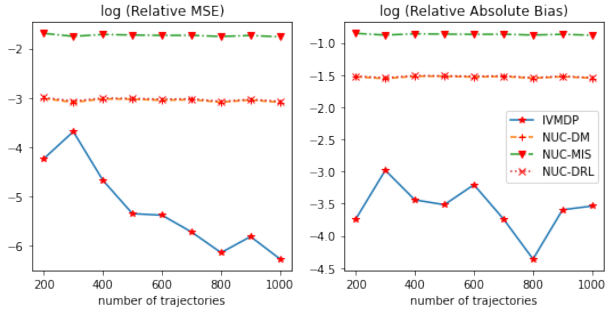

Figure 3. Logarithmic relative MSE (left panel) and logarithmic relative absolute bias (right panel) of various estimators with sample size on the x-axis. Notice that the yellow dashed line and the red dotted line are largely overlapped due to the similar performance between NUC-DM and NUC-DRL.

Competing methods. We consider three baseline methods, corresponding to the DM estimator, the MIS estimator (Liu et al., 2018) and the DRL estimator (Kallus & Uehara, 2022). All the estimators are derived under the NUC assumption without the use of IV, denoted by NUC-DM, NUC-MIS and NUC-DRL, respectively. To ensure a fair comparison, we also incorporate the IV in the state when implementing the three baseline approaches. The detailed estimation procedures are given in Appendix C.2.

Results. The results are shown in Figure 3. We can see that our proposed estimator IVMDP achieves the smallest MSE and bias in all cases. Its MSE generally decays with an increase in the number of trajectories, demonstrating the consistency of our proposal. In contrast, other estimators are severely biased, highlighting the risk of ignoring unobserved confounding. The biases of baseline methods dominate the standard deviations, resulting in the MSEs to be relatively constant despite the increase in the number of trajectories.

7.2. Real Data Analysis

In this section, we apply our method to a real dataset from a world-leading technological company. The company conducts advertising compaigns to attract consumers to download their mobile app products. The advertisements are delivered through multiple media channels, such as search, display, social, mobile and video, provided by ads exchange or mobile application stores. During the compaign, an individual user is typically exposed to various advertisements delivered through these channels. To improve the return on investment, it is crucial for the company to accurately evaluate the long-term effects of different ads exposure policies.

The dataset is collected from a randomized advertising campaign. That is, at each time, the company randomly decided whether to bid against other firms or not to display their ad to a target consumer. As such, the IV independence assumption (see Assumption 1(a)) is automatically satisfied. In addition, if the company chooses to bid, it will largely increase the chance that their ad is indeed displayed to the consumer. This meets the IV relevance assumption (see Assumption 1(b)). Finally, bidding can only affect the conversion rate or the consumer behavior through the ad exposure. This verifies the exclusion restriction assumption (see Assumption 1(c) & (e)). The core IV assumptions are thus satisfied in our example.

Due to privacy considerations, we generate a synthetic data environment based on the real data and report the performance of our proposal applied to this environment. Specifically, we adopt the IV-based MDPUC model to model the data generating process and leverage the IV to estimate the reward and next-state distributions in the presence of unmeasured confounding. See Appendix C.3 for more details.

The numerical results are reported in Figure 4. First, we can see that our proposed estimator achieves the least bias and MSE in all cases. In particular, the bias of IVMDP remains close to zero, and its MSE decreases with an increase in the number of trajectories. In contrast, other estimators significantly underestimate the true value, and their MSEs do not decay with the number of trajectories. This demonstrates that our estimator is able to effectively handle unmeasured confounding factors, while other estimators are considerably biased.

8. Future Work

In this paper, we presented a systematic approach for using instrumental variables to perform off-policy evaluation in infinite-horizon confounded MDPs. To the best of our knowledge, this is the first work to derive the efficient influence function of the value function in an IV-based confounded MDP, high-order MDP and POMDP.

Our numerical results in simulation and real data analysis both demonstrate the effectiveness of our method.

There are several potential avenues for future work that could build on the advances presented in this paper. One possibility is to extend the framework to handle discrete or continuous action spaces, as has been done in previous work such as (Heckman et al., 2008; Cai et al., 2021). Another option is to further explore more efficient estimators, such as IS, DR estimators under confounded POMDP by continuing our discussion in Section 6.

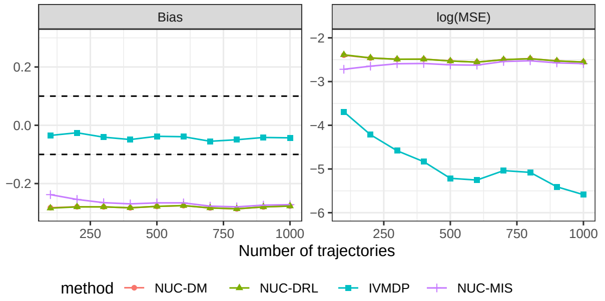

Figure 4. Bias (left panel) and logarithmic MSE (right panel) and as the sample size increases. The two dashed lines correspond to the Monte Carol error bounds for the bias. As we can see from the figure, only our approach is empirically unbiased and achieves the smallest MSE.

Acknowledgements

We would like to thank the anonymous (meta-)reviewers of ICML 2023 for helpful comments. This work was partially supported by an EPSRC grant EP/W014971/1 (to CS). We also appreciate the support from NSF-DMS-2003637.

References

Abadie, A. Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263, 2003.

Abadie, A. and Cattaneo, M. D. Econometric methods for program evaluation. *Annual Review of Economics*, 10(1), 2018.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014.

Angrist, J. and Imbens, G. Identification and estimation of local average treatment effects, 1995.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455, 1996.

Aronow, P. M. and Carnegie, A. Beyond late: Estimation of the average treatment effect with an instrumental variable. *Political Analysis*, 21(4):492–506, 2013.

Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. Reinforcement learning of pomdps using spectral methods. In *Conference on Learning Theory*, pp. 193–256. PMLR, 2016.

Bareinboim, E., Forney, A., and Pearl, J. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28, 2015.

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

Bennett, A., Kallus, N., Li, L., and Mousavi, A. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pp. 1999–2007. PMLR, 2021.

Boots, B., Siddiqi, S. M., and Gordon, G. J. Closing the learning-planning loop with predictive state representations. *The International Journal of Robotics Research*, 30(7):954–966, 2011.

Cai, H., Shi, C., Song, R., and Lu, W. Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34: 15285–15300, 2021.

Chen, S. and Zhang, B. Estimating and improving dynamic treatment regimes with a time-varying instrumental variable. *arXiv preprint arXiv:2104.07822*, 2021.

Chen, X. and Qi, Z. On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 3558–3582. PMLR, 2022.

Chernozhukov, V., Chetverikov, D., and Kato, K. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597, 2014.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1), 2018.

Cui, Y. and Tchetgen Tchetgen, E. A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *Journal of the American Statistical Association*, 116(533):162–173, 2021.

Dai, B., Nachum, O., Chow, Y., Li, L., Szepesvári, C., and Schuurmans, D. Coindice: Off-policy confidence interval estimation. *Advances in neural information processing systems*, 33:9398–9411, 2020.

Ertefaie, A. Constructing dynamic treatment regimes in infinite-horizon settings. *arXiv preprint arXiv:1406.0764*, 2014.

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

Feng, Y., Ren, T., Tang, Z., and Liu, Q. Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pp. 3102–3111. PMLR, 2020.

Fu, Z., Qi, Z., Wang, Z., Yang, Z., Xu, Y., and Kosorok, M. R. Offline reinforcement learning with instrumental variables in confounded markov decision processes. *arXiv preprint arXiv:2209.08666*, 2022.

Guo, Z. D., Doroudi, S., and Brunskill, E. A pac rl algorithm for episodic pomdps. In *Artificial Intelligence and Statistics*, pp. 510–518. PMLR, 2016.

Hallak, A. and Mannor, S. Consistent on-line off-policy evaluation. In *International Conference on Machine Learning*, pp. 1372–1383. PMLR, 2017.

Hanna, J. P., Stone, P., and Niekum, S. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvari, C., and Wang, M. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pp. 4074–4084. PMLR, 2021.

Heckman, J. J., Urzua, S., and Vytlacil, E. Instrumental variables in models with multiple outcomes: The general unordered case. *Annales d'Economie et de Statistique*, pp. 151–174, 2008.

Hu, Y. and Wager, S. Off-Policy Evaluation in Partially Observed Markov Decision Processes. *arXiv e-prints*, art. arXiv:2110.12343, October 2021. doi: 10.48550/arXiv. 2110.12343.

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.

Jin, C., Kakade, S., Krishnamurthy, A., and Liu, Q. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33:18530–18539, 2020.

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

Kallus, N. and Zhou, A. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 33:22293–22304, 2020.

Kallus, N., Mao, X., and Uehara, M. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

Kiyohara, H., Saito, Y., Matsuhiro, T., Narita, Y., Shimizu, N., and Yamamoto, Y. Doubly robust off-policy evaluation for ranking policies under the cascade behavior model. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 487–497, 2022.

Kosorok, M. R. and Laber, E. B. Precision medicine. *Annual review of statistics and its application*, 6:263, 2019.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34: 24523–24534, 2021.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4: 1107–1149, 2003.

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

Li, J., Luo, Y., and Zhang, X. Causal reinforcement learning: An instrumental variable approach. *arXiv preprint arXiv:2103.04021*, 2021.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Liao, L., Fu, Z., Yang, Z., Wang, Y., Kolar, M., and Wang, Z. Instrumental variable value iteration for causal offline reinforcement learning. *arXiv preprint arXiv:2102.09907*, 2021a.

Liao, P., Klasnja, P., and Murphy, S. Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391, 2021b.

Liao, P., Qi, Z., Wan, R., Klasnja, P., and Murphy, S. A. Batch policy learning in average reward markov decision processes. *The Annals of Statistics*, 50(6):3364–3387, 2022.

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*, 115(530):692–706, 2020.

Miao, R., Qi, Z., and Zhang, X. Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. *arXiv preprint arXiv:2209.10064*, 2022.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

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

Nair, Y. and Jiang, N. A spectral approach to off-policy evaluation for pomdps. *arXiv preprint arXiv:2109.10502*, 2021.

Namkoong, H., Keramati, R., Yadlowsky, S., and Brunskill, E. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

Ogburn, E. L., Rotnitzky, A., and Robins, J. M. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):373–396, 2015.

Qiu, H., Carone, M., Sadikova, E., Petukhova, M., Kessler, R. C., and Luedtke, A. Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association*, 116(533):174–191, 2021.

Quillen, D., Jang, E., Nachum, O., Finn, C., Ibarz, J., and Levine, S. Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6284–6291. IEEE, 2018.

Schlegel, M., Chung, W., Graves, D., Qian, J., and White, M. Importance resampling for off-policy prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

Sen, R., Shanmugam, K., Kocaoglu, M., Dimakis, A., and Shakkottai, S. Contextual bandits with latent confounders: An nmf approach. In *Artificial Intelligence and Statistics*, pp. 518–527. PMLR, 2017.

Shi, C., Wan, R., Song, R., Lu, W., and Leng, L. Does the markov decision process fit the data: Testing for the markov property in sequential decision making. In *International Conference on Machine Learning*, pp. 8807–8817. PMLR, 2020a.

Shi, C., Wan, R., Chernozhukov, V., and Song, R. Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pp. 9580–9591. PMLR, 2021.

Shi, C., Uehara, M., Huang, J., and Jiang, N. A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *International Conference on Machine Learning*, pp. 20057–20094. PMLR, 2022a.

Shi, C., Zhang, S., Lu, W., and Song, R. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3): 765–793, 2022b.

Shi, C., Zhu, J., Ye, S., Luo, S., Zhu, H., and Song, R. Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, (just-accepted):1–30, 2022c.

Shi, X., Miao, W., Nelson, J. C., and Tchetgen Tchetgen, E. J. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):521–540, 2020b.

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

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. Doubly robust bias reduction in infinite horizon off-policy estimation. In *International Conference on Learning Representations*, 2020.

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

Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X., and Miao, W. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Tennenholtz, G., Shalit, U., and Mannor, S. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10276–10283, 2020.

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.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

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

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. *Dynamic treatment regimes: Statistical methods for precision medicine*. Chapman and Hall/CRC, 2019.

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., Kiyohara, H., Bennett, A., Chernozhukov, V., Jiang, N., Kallus, N., Shi, C., and Sun, W. Future-dependent value-based off-policy evaluation in pomdps. *arXiv preprint arXiv:2207.13081*, 2022a.

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

Van Der Vaart, A. W., van der Vaart, A., van der Vaart, A. W., and Wellner, J. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

Wang, J., Qi, Z., and Shi, C. Blessing from experts: Super reinforcement learning in confounded environments. *arXiv preprint arXiv:2209.15448*, 2022.

Wang, L. and Tchetgen, E. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):531–550, 2018.

Wang, L., Yang, Z., and Wang, Z. Provably efficient causal reinforcement learning with confounded observational data. *Advances in Neural Information Processing Systems*, 34:21164–21175, 2021.

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.

Xu, L., Kanagawa, H., and Gretton, A. Deep proxy causal learning and its application to confounded bandit policy evaluation. *Advances in Neural Information Processing Systems*, 34:26264–26275, 2021.

Zhang, J. and Bareinboim, E. Markov decision processes with unobserved confounders: A causal approach. Technical report, Technical report, Technical Report R-23, Purdue AI Lab, 2016.

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

APPENDIX

A. Assumptions involved in Theorem 4.1

To introduce the technical assumptions in our identification theorem, we adopt a counterfactual outcome framework that is commonly used in the IV literature. Let $\bar{A}_t = (A_1, \dots, A_t)$ denote the action history up to time t, and $\bar{Z}_t = (Z_1, \dots, Z_t)$ denote the history of IVs up to time t. Define $A_t(\bar{z}_t, \bar{a}_{t-1})$ as the potential action assigned to a subject at time t if they were exposed to $\bar{Z}_t = \{\bar{z}_t\}$ and $\bar{A}_{t-1} = \{\bar{a}_{t-1}\}$, and $R_t(\bar{z}_t, \bar{a}_t)$, $S_{t+1}(\bar{z}_t, \bar{a}_t)$ as the potential reward and next state that would be observed if the subject were to receive $\{\bar{z}_t\}$ and $\{\bar{a}_t\}$ in the past.

Assumption 1. (IV Assumptions)
For any time $t \in \{1, \dots, T\}$, we assume:
(a) IV Independence: $Z_t \perp\!\!\!\perp U_t | S_t$.
(b) IV Relevance: $Z_t \not\perp\!\!\!\perp A_t | S_t$.
(c) Exclusion Restriction: For any $\bar{z}_t, \bar{a}_t$, $R_t(\bar{z}_t, \bar{a}_t) = R_t(\bar{z}_{t-1}, \bar{a}_t)$.
(d) $R_t(\bar{a}_t) \perp\!\!\!\perp (A_t, Z_t) | (S_t, U_t)$.
(e) Exclusion Restriction: For any $\bar{z}_t, \bar{a}_t$, $S_{t+1}(\bar{z}_t, \bar{a}_t) = S_{t+1}(\bar{z}_{t-1}, \bar{a}_t)$.
(f) $S_{t+1}(\bar{z}_t, \bar{a}_t) \perp\!\!\!\perp (A_t, Z_t) | (S_t, U_t)$.
(g) There is no additive $U - A$ interaction in both $\mathbb{E}[R_t(\bar{z}_t, \bar{a}_t) | S_t, U_t]$ and $\mathbb{E}[S_{t+1}(\bar{z}_t, \bar{a}_t) | S_t, U_t]$. That is,

$$\begin{aligned} &\mathbb{E}[R_t(\bar{z}_t, \bar{a}_{t-1}, a_t = 1) - R_t(\bar{z}_t, \bar{a}_{t-1}, a_t = 0) | S_t, U_t] = \mathbb{E}[R_t(\bar{z}_t, \bar{a}_{t-1}, a_t = 1) - R_t(\bar{z}_t, \bar{a}_{t-1}, a_t = 0) | S_t], \\ \text{and } \; &\mathbb{E}[S_{t+1}(\bar{z}_t, \bar{a}_{t-1}, a_t = 1) - S_{t+1}(\bar{z}_t, \bar{a}_{t-1}, a_t = 0) | S_t, U_t] = \mathbb{E}[S_{t+1}(\bar{z}_t, \bar{a}_{t-1}, a_t = 1) - S_{t+1}(\bar{z}_t, \bar{a}_{t-1}, a_t = 0) | S_t]. \end{aligned}$$

Assumption 1 (a)-(c) ensure the validity of IVs, which are commonly used in the single-stage model setup (Angrist & Imbens, 1995; Abadie, 2003; Qiu et al., 2021). Assumption 1 (d), as discussed in (Wang & Tchetgen, 2018), allows for common causes of Z_t and A_t, and can be interpreted through d-separation. This assumption is mild in real-world settings, as it allows for common causes of Z_t and A_t, A_t and (R_t, S_{t+1}). Assumption 1 (e)-(f) are akin to (c)-(d), which ensures the impact of the IV to be the same for both the current-stage reward and next-stage state variables. As shown in the causal graph in Figure 1, R_t and S_{t+1} have the same causal hierarchy, leading to similar IV-related assumptions.

Assumption 1 (g) guarantees that conditioning on covariates S_t, unmeasured confounders U_t only affect the causal effect of A_t on the mean of current-state reward or next-state covariates in an additive way. This assumption is commonly used in related papers to ensure the indentifiability of the final estimand (Wang & Tchetgen, 2018; Qiu et al., 2021). It is also worth mentioning that instead of requiring the unmeasured confounder U_t to be conditionally independent to the reward function, we only require it to be conditionally independent of the difference of the two reward functions.

Next, let's further impose the conditional independence assumptions that is commonly assumed in Markov decision processes. Define $\bar{W}_t$ as the set of all historical data up to stage t, where

$$\bar{W}_t(\bar{z}_t, \bar{a}_t) = \{S_0, U_0, R_0(z_0, a_0), \dots, S_t(\bar{z}_{t-1}, \bar{a}_{t-1}), U_t, R_t(\bar{z}_t, \bar{a}_t)\}.$$

Assumption 2. (Conditional Independence Assumptions)
(a) (MA) Markov assumption: There exists a Markov transition kernel $\mathcal{P}$ such that for any $t \geq 0$, $\bar{z}_t \in [0,1]^{t+1}$ and $\bar{a}_t \in [0,1]^{t+1}$, we have

$$\mathbb{P}(S_{t+1}(\bar{z}_t, \bar{a}_t) \in \mathcal{S} | \bar{W}_t(\bar{z}_t, \bar{a}_t)) = \mathcal{P}(\mathcal{S}; z_t, a_t, S_t(\bar{z}_{t-1}, \bar{a}_{t-1}), U_t).$$

(b) (CMIA) Conditional mean independence assumption: there exists a function r such that for any $t \geq 0$, and $\bar{a}_t \in [0,1]^{t+1}$, we have

$$\mathbb{E}(R_t(\bar{z}_t, \bar{a}_t) | S_t(\bar{z}_{t-1}, \bar{a}_{t-1}), \bar{W}_{t-1}(\bar{z}_{t-1}, \bar{a}_{t-1})) = r(z_t, a_t, S_t(\bar{z}_{t-1}, \bar{a}_{t-1}), U_t).$$

(c) For any $t \in \{0, \dots, T\}$, the conditional distribution of Z_t, A_t and U_t, given all historical data is only a function of the current state information. Specifically,

$$\begin{aligned} &\mathbb{E}(Z_t | S_t(\bar{z}_{t-1}, \bar{W}_{t-1}(\bar{z}_{t-1}, \bar{a}_{t-1})) = \mathbb{E}(Z_t | S_t(\bar{z}_{t-1})), \\ &\mathbb{P}(U_t | S_t(\bar{z}_{t-1}, \bar{W}_{t-1}(\bar{z}_{t-1}, \bar{a}_{t-1})) = \mathbb{P}(U_t | S_t(\bar{z}_{t-1})), \\ &\mathbb{E}(A_t(\bar{z}_t, \bar{a}_{t-1}) | S_t(\bar{z}_{t-1}, \bar{a}_{t-1}), z_t, U_t, \bar{W}_{t-1}(\bar{z}_{t-1}, \bar{a}_{t-1})) = \mathbb{E}(A_t(\bar{z}_t, \bar{a}_{t-1}) | S_t(\bar{z}_{t-1}, \bar{a}_{t-1}), z_t, U_t). \end{aligned}$$

It is worth mentioning that under Assumption 1 (c) and (e), we can further omit term z_t on the RHS of all equations in Assumption 2. Moreover, when both Assumption 1 and 2 holds, the definition of $\bar{W}_t(\bar{z}_t, \bar{a}_t)$ and the potential outcomes for R_t and S_{t+1} are a function of only $\bar{a}_t$, not $\bar{z}_t$. This result is easy to understand: Assumption 1 (c) restricts the effect of z_t on R_t, making the potential outcome of R_t independent of z_t given the current-state action. Meanwhile, the conditional independence assumption ensures that R_t won't be affected by the historical IVs $\bar{z}_{t-1}$, yielding the potential outcome of R_t to be entirely independent of $\bar{z}_t$ given the action sequence $\bar{a}_t$. As such, one can relax some conditions in Assumption 1 without any loss of information. Details are provided in Proposition 1.

Proposition. Under Assumption 2, the exclusion restriction condition in Assumption 1 (c) is equivalent to assuming that $R_t(\bar{z}_t, \bar{a}_t) = R_t(\bar{a}_t)$ holds for any $\bar{z}_t, \bar{a}_t$. Meanwhile, Assumption 1 (e) is equivalent to assuming that $S_{t+1}(\bar{z}_t, \bar{a}_t) = S_{t+1}(\bar{a}_t)$ holds for any $\bar{z}_t, \bar{a}_t$.

The proof of Proposition 1 is straightforward. Under Assumption 2 (b),

$$R_t(\bar{Z}_t, \bar{A}_t) \perp\!\!\!\perp \bar{Z}_{t-1} | (S_t, Z_t, A_t),$$

which means that $R_t(\bar{z}_t, \bar{a}_t) = R_t(z_t, \bar{a}_t) = R_t(\bar{a}_t)$. The first equality holds by CIMA in Assumption 2 (b), and the second equality holds by the original exclusion restriction in Assumption 1 (c). Similarly, we can prove Assumption 1 (e) by only assuming that $S_{t+1}(\bar{z}_t, \bar{a}_t) = S_{t+1}(\bar{a}_t)$ holds for any $\bar{z}_t, \bar{a}_t$.

B. Estimation Details

In this section, we will summarize the estimation procedures for the models mentioned in Section 5. We will first briefly summarize the estimation of some functions that can be easily modeled, and then discuss the estimation of Q^π, V^π and ω^π in the following two subsections.

Estimating p_z, p_a, and p_r can be treated as standard regression or classification problems, depending on the type of covariates. Any appropriate supervised learning methodology satisfying the convergence rate detailed in Theorem 5.2 can be used to estimate these models. Additionally, since $\rho(s_t, z_t)$, $c(z_t|s_t)$ are both functions of p_z, p_a, p_r and π, we can first estimate these pdfs/pmfs and then use the resulting estimators to construct plug-in estimators for ρ and c.

B.1. The estimation of Q^π and V^π

We first consider the estimation of $Q^\pi(s, z, a)$ and $V^\pi(s)$. According to Formula (4), we can derive the Bellman equation under this confounded MDP as

$$Q^\pi(S_t, Z_t, A_t) = \mathbb{E}\Big\{R_t + \gamma \sum_{z,a} c(z|S_{t+1}) p_a(a|z, S_{t+1}) Q^\pi(S_{t+1}, z, a) \Big| S_t, Z_t, A_t\Big\}.$$

Motivated by (Le et al., 2019), we employ fitted-Q evaluation method to iteratively solve the Q function until convergence. Specifically, at the lth step, we update Q^{l+1} by

$$Q^{\pi,l+1} = \arg\min_{Q^\pi \in \mathcal{Q}} \sum_{i,t} \Big\{R_{i,t} + \gamma \widehat{V}^{\pi,l}(S_{i,t+1}) - Q^\pi(S_{i,t}, Z_{i,t}, A_{i,t})\Big\}^2,$$

where $\mathcal{Q}$ denotes some function class, and $\widehat{V}^{\pi,l}(S_{t+1}) = \sum_{z,a} \widehat{c}(z|S_{t+1}) \widehat{p}_a(a|z, S_{t+1}) \widehat{Q}^{\pi,l}(S_{t+1}, z, a)$ is the value function calculated from the Q function at the previous step. The algorithm terminates when the maximum number of iterations is reached or a convergence criterion is met. We use the Q function and value function from the final iteration as our estimates of Q^π and V^π.

To summarize, the key distinction between this estimation procedure and the original fitted-Q evaluation algorithm described in (Le et al., 2019) is the specific relationship between V^π and Q^π. In the where unmeasured confounders exist, the function $c(z|s)$ plays an important role in iteratively updating the value function, thereby ensuring the consistency of estimators for both V^π and Q^π.

B.2. The estimation of ω^π

Then, let's consider the estimation of $\omega^\pi(s)$. Define

$$L(\omega, f) = \gamma \cdot \mathbb{E}_{(s,a,s') \sim p_t^\pi} \left[\Delta(\omega; s, a, s') \cdot f(s')\right] + (1 - \gamma) \cdot \mathbb{E}_{s_0 \sim \nu_0(s)} \left[(1 - \omega(s)) \cdot f(s)\right],$$

where s' denotes the next-state covariates, $\Delta(\omega; s, a, s') := \omega(s) \cdot \rho(s, z) - \omega(s')$. In confounded MDPs, we can further derive $L(\omega, f)$ as

$$L(\omega, f) = (1-\gamma)\sum_s f(s)\nu(s) - \mathbb{E}\omega(S_{i,t})\Big\{f(S_{i,t}) - \gamma \cdot \rho(S_{i,t}, Z_{i,t}) \cdot f(S_{i,t+1})\Big\}. \tag{12}$$

According to Theorem 4 in (Liu et al., 2018), $\omega^\pi(s)$ is the solution to $L(\omega, f) = 0$ for any discriminator function f. Therefore, ω^π can be learned by solving the mini-max problem for the quadratic form of the loss function $L(\omega, f)$. Specifically, we aim to find the solution to $\arg\min_{\omega\in\Omega}\sup_{f\in\mathcal{F}} L^2(\omega, f)$ for some function class Ω and $\mathcal{F}$.

For the ease of illustrations, let's consider linear bases for Ω and $\mathcal{F}$. Suppose $\omega^\pi(s) = \xi^T(s)\beta$ where $\xi^T(s)$ denotes the basis function. By Formula (12), β can be estimated by

$$\widehat{\beta} = \Big[\sum_{i=1}^{N}\sum_{t=0}^{T-1}\xi(S_{i,t})\Big\{\xi^T(S_{i,t}) - \gamma\widehat{\rho}(S_{i,t}, Z_{i,t})\xi^T(S_{i,t+1})\Big\}\Big]^{-1} \times (1-\gamma)NT \cdot \sum_s \xi(s)\nu(s).$$

Therefore, we can derive the final estimator for ω^π as $\widehat{\omega}^\pi = \xi^T(s) \cdot \widehat{\beta}$.

C. More on Simulation and Real Data Analysis

C.1. The Data Generating Process in Illustrating Double Robustness

In this section, we detail the data generating process used in simulation section to prove the double robustness property of our final estimator $\widehat{\eta}^\pi$. For the sake of computational cost, we let $T = 100$ and the number of data trajectories $N = \{100, 200, \ldots, 1000\}$. The initial state distribution is generated by a Bernoulli distribution with $p = 0.5$, i.e. $S_0 \sim \text{Ber}(1, p)$. We define the unmeasured confounder at each stage as U_t as another Bernoulli random variable with $p = 0.5$. The instrumental variable Z_t, action A_t, reward R_t and next state S_{t+1} all follow Bernoulli distributions with the corresponding success rates $\mathbb{P}(Z_t = 1) = \text{sigmoid}(S_t + \delta_t - 2)$ with $\mathbb{P}(\delta_t = 0.25) = \mathbb{P}(\delta_t = 0) = 0.5$, $\mathbb{P}(A_t = 1) = \text{sigmoid}(S_t + 2Z_t + 0.5U_t - 2)$, and $\mathbb{P}(R_t = 10) = \mathbb{P}(S_{t+1} = 1) = \text{sigmoid}(S_t + A_t + U_t - 2)$. In this simulation, we set $U_t' = 0$ for simplicity, which avoids the confounding between Z_t and A_t. However, the confounder between A_t and (R_t, S_{t+1}) does exists, which is given by U_t.

In order to evaluate the doubly robust property of our estimator, we use Monte Carlo method to approximate the true models for all functions, and then deliberately introduce shifts that can lead to model misspecification. Specifically, to misspecify ω^π, we let $\omega^\pi_{\text{shifted}}(s_0 = 1) = \omega^\pi_{\text{true}}(s_0 = 1)/2$, and $\omega^\pi_{\text{shifted}}(s_0 = 0) = 2\omega^\pi_{\text{true}}(s_0 = 0)$. To misspecify p_z, we define a shift parameter $\alpha \in [0, 1]$, and denote $p_{z,\text{shifted}}(z = 1|s) = \alpha \cdot p_{z,\text{true}}(z = 1|s) + (1-\alpha) \cdot p_{z,\text{true}}(z = 0|s)$. To misspecify the Q function Q^π, we define another shift parameter $\beta \in \mathbb{R}$, and let $Q^\pi_{\text{shifted}}(s, z, a) = Q^\pi_{\text{true}}(s, z, a) + \beta(s, z, a)$. In our simulation setup, we fix $\alpha = 0.55$, and set $\beta(s, z, a) \sim \mathcal{N}(5, 4)$.

C.2. The Estimation Details of Three Baseline Approaches under NUC Assumption

In this section, we detail three baseline approaches used for comparison in Section 7.

The first competing method is a direct estimator (NUC-DM), which is represented by the yellow dashed line in Figure 3. When NUC assumption holds, the Bellman equation becomes

$$\mathbb{E}\Big\{R_t + \gamma \cdot \sum_a \pi(a|S_{t+1}) \cdot Q^\pi(S_{t+1}, a)\Big| S_t, A_t\Big\} = Q^\pi(S_t, A_t).$$

Thus, we can conduct fitted Q evaluation to repeatedly estimate $Q^\pi(s, a)$ and $V^\pi(s)$ until convergence:

$$Q^{\pi,l+1} = \arg\min_{Q^\pi\in\mathcal{Q}}\sum_{i,t}\Big\{R_{i,t} + \gamma\widehat{V}^{\pi,l}(S_{i,t+1}) - Q^\pi(S_{i,t}, A_{i,t})\Big\}^2,$$

where $\mathcal{Q}$ is some function class, and $\widehat{V}^{\pi,l}(S_{t+1}) = \sum_a \pi(a|S_{t+1}) \cdot \widehat{Q}^{\pi,l}(S_{t+1}, a)$ is the value function calculated from the Q function at the lth step. As such, the final NUC-DM estimator is given by

$$\widehat{\eta}^\pi_{\text{NUC-DM}} = \sum_{a,s_0}\pi(a|s_0) \cdot \widehat{Q}^\pi(s_0, a) \cdot \nu(s_0).$$

The second estimator is an MIS estimator (NUC-MIS), which is represented by the green dash-dotted line in Figure 3. According to (Liu et al., 2018), we can calculate the NUC-MIS estimator by

$$\widehat{\eta}^{\pi}_{\text{NUC-MIS}} = (1-\gamma)^{-1}\frac{1}{\sum_i T_i}\sum_{i,t}\hat{\omega}^{\pi}(S_{i,t})\cdot\hat{\beta}_{\pi/\pi_0}(a,s)\cdot R_{i,t},$$

where $\beta_{\pi/\pi_0}(a,s) = \pi(a|s)/\pi_0(a|s)$ and $\hat{\omega}^{\pi}(S_t)$ can be obtained from the method provided in the original paper. Details are omitted here.

The third estimator is the double reinforcement learning estimator (NUC-DRL), which is represented by the red dotted line in Figure 3. DRL combines the NUC-DM and NUC-MIS estimators to provide a more robust estimator under the NUC assumption (Kallus & Uehara, 2022). The final estimator is given by

$$\widehat{\eta}^{\pi}_{\text{NUC-DRL}} = \widehat{\eta}^{\pi}_{\text{NUC-DM}} + \widehat{\phi}_{\text{NUC-aug}} = \sum_{a,s_0}\pi(a|s_0)\cdot\widehat{Q}^{\pi,l}(s_0,a)\cdot\nu(s_0) +$$
$$(1-\gamma)^{-1}\frac{1}{\sum_i T_i}\sum_{i,t}\hat{\omega}^{\pi}(S_{i,t})\hat{\beta}_{\pi/\pi_0}(a,s)\big\{R_{i,t} + \gamma\widehat{V}^{\pi}(S_{i,t+1}) - \widehat{Q}^{\pi}(S_{i,t},A_{i,t})\big\}.$$

C.3. Estimating Transition Functions for Real Data

Specifically, the conditional distribution of reward R_t given S_t, Z_t, A_t is modeled by a logistic regression, i.e. $\hat{p}_r(R_t = 1|S_t, Z_t, A_t) = \sigma((1, S_t^T, Z_t, A_t)^T\beta_r)$ where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function. Similarly, we estimate $\mathbb{E}(S_{t+1}|S_t, Z_t, A_t)$ by fitting a multivariate linear model $T(S_t, Z_t, A_t)$ with response $\{S_{i,t+1}\}$ and covariates $\{(S_{i,t}, Z_{i,t}, A_{i,t})\}$. Then, the transition model of $S_{t+1}|S_t, Z_t, A_t$ is characterized by $N(T(S_t, Z_t, A_t), \text{diag}\{\sigma_1^2, \dots, \sigma_6^2\})$ where $\sigma_i^2(i = 1, \dots, 6)$ are estimated by the residual of the linear model for S_i. For the transition from S_t, Z_t to A_t, we model it with a logistic regression, which is denoted as $\hat{p}_a(A_t = 1|S_t, Z_t) = \sigma((1, S_t^T, Z_t)^T\beta_a)$. Finally, since Z_t is independent to S_t, we simply model it by a binomial random variable with probability p in which p is estimated by empirical frequency.

We depict the procedure for computing the true policy value of target policy: $\pi(1|S_t) = \sigma((1, S_t^T, \mathbb{E}(Z_t))^\top\beta)$. Motivated by our identification result in Theorem 4.1, we can use Monte Carlo method to estimate η^{π}. More precisely, we draw N' trajectories with length T from the fitted MDP, and thus, attain the transition tuple $\{S_{i,t}, Z_{i,t}, A_{i,t}, R_{i,t}\}_{i=1,t=1}^{N',T}$; then we can compute

$$\eta_{\text{MC}} = \frac{1}{N'}\sum_{i=1}^{N'}\sum_{t=j}^{T}R_{i,t}\prod_{j=0}^{t}\frac{c(Z_{i,j}|Z_{i,j})}{p_z(Z_{i,j}|S_{i,j})}$$

that can be treated as the estimator for η^{π}. To guarantees the accuracy of η_{MC}, we set $N' = 10^8$ to overcome the large variance of $\prod_{j=0}^{t}\frac{c(Z_{i,j}|S_{i,j})}{p_z(Z_{i,j}|S_{i,j})}$.

D. More on Simulation Study

To test the performance of IVMDP when either the instrumental variable (IV) or the unmeasured confounder is weak, we conducte ablation studies in five groups of simulation settings.

Same as our main paper, we compare the performance of our approach, IVMDP, with three existing baselines under NUC assumptions. We denote them as NUC-DM, NUC-MIS, and NUC-DRL.

D.1. Data Generating Process

In our paper, Figure 3 was obtained under the following data generating process:

$$A_t \sim \text{Ber}(1, \text{sigmoid}(S_{t1} + S_{t2} + 2Z_t + 0.5U_t))$$
$$R_t \sim \text{Ber}(1, \text{sigmoid}(S_{t1} + S_{t2} + 2A_t + 2.5U_t))$$

To weaken the effect of IV, we adjust the coefficient of Z_t in the expression of A_t as below:

1. **Medium strength IV**: $A_t \sim \text{Ber}(1, \text{sigmoid}(S_{t1} + S_{t2} + Z_t + 0.5U_t))$
2. **Weak IV**: $A_t \sim \text{Ber}(1, \text{sigmoid}(S_{t1} + S_{t2} + 0.5Z_t + 0.5U_t))$

To weaken the effect of unmeasured confounders, we adjust the coefficient of U_t in the expression of R_t as below:

1. **Medium strength confounders**: $R_t \sim \text{Ber}(1, \text{sigmoid}(S_{t1} + S_{t2} + 2A_t + 1.5U_t))$
2. **Weak confounders**: $R_t \sim \text{Ber}(1, \text{sigmoid}(S_{t1} + S_{t2} + 2A_t + 0.5U_t))$

D.2. Result Comparison

Combining the above four cases with the original setting in the paper, we obtained five scenarios in total for comparison. Same as the main paper, all of the figures below are obtained under 100 times of replications.

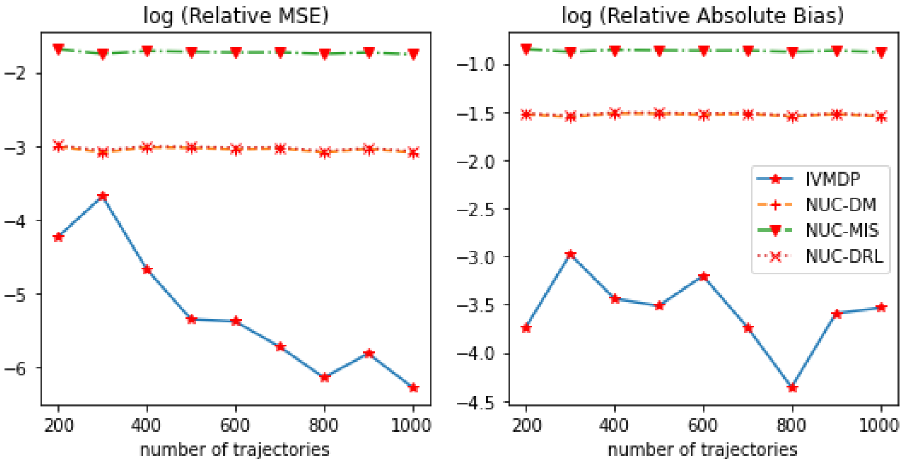

Figure 5. The figure in the main paper

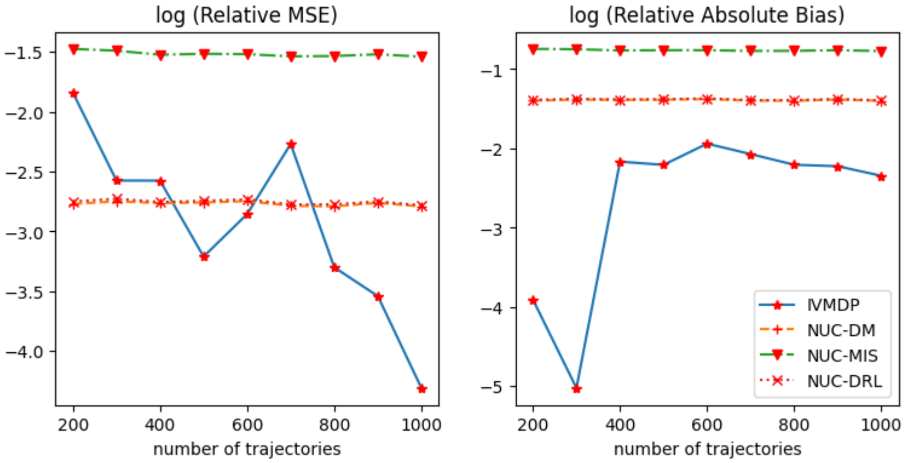

Figure 6. Medium strength IV

As we can see from Figure 5-9, the variance of IVMDP indeed increased when the effect of instrumental variables or confounders is weak. Despite so, the estimation bias of IVMDP is (almost) always the smallest in all settings. Moreover, although the MSE of IVMDP is relatively unstable compared with other approaches under NUC assumption, IVMDP has a potential to perform better than baseline methods when the number of trajectories increases, even in the case when IV or the unmeasured confounder is very weak. However, if the real data indicates either weak signals on the instrumental variable (IV) or unmeasured confounders, we recommend employing IVMDP only when a sufficient number of data points is available.

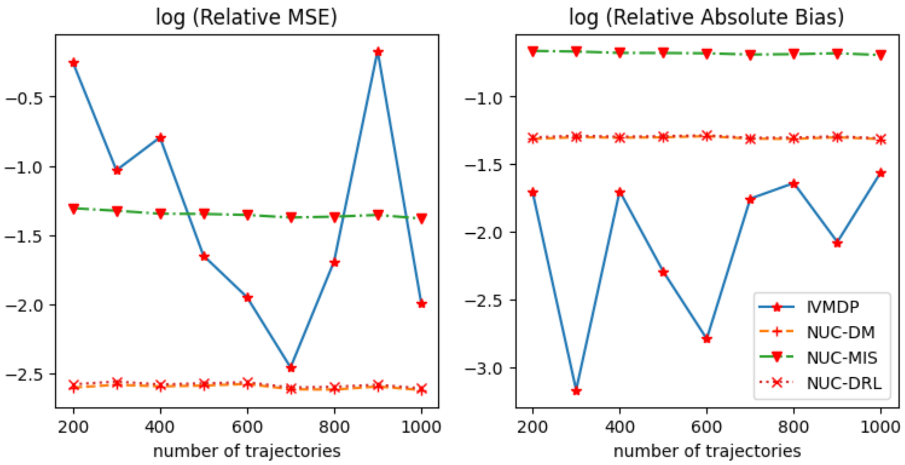

Figure 7. Weak IV

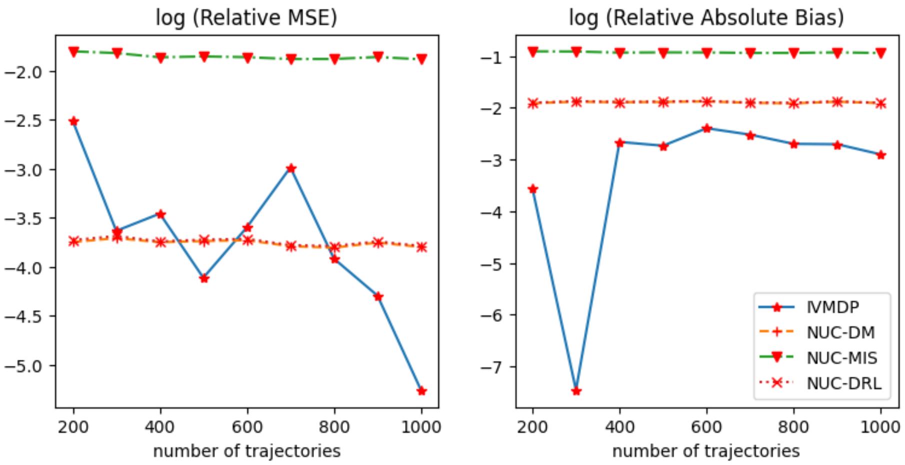

Figure 8. Medium strength confounders

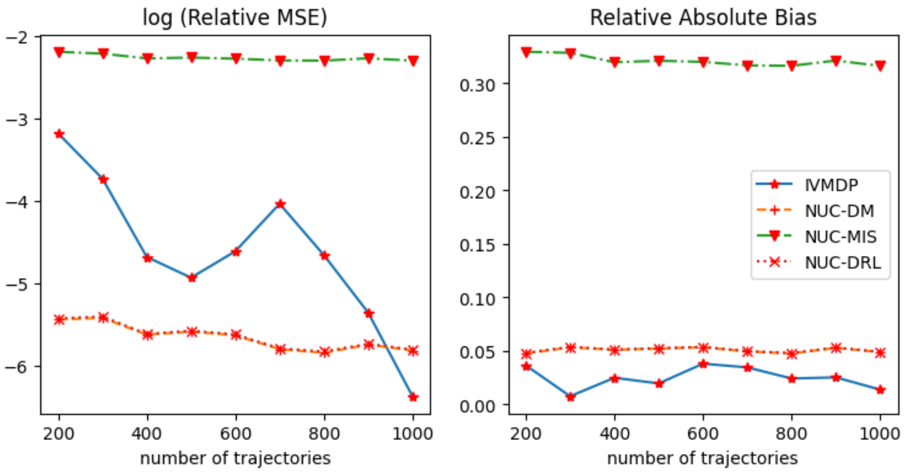

Figure 9. Weak confounders

E. Proof of Theorem 4.1

Theorem 1 states the identifiability of the value function, i.e. $V^\pi(s_0)$ can be entirely estimated from the observed data.

Since the value function is defined as $V^\pi(s_0) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}^\pi\left[R_t | S_0 = s_0\right]$, it suffice to identify $\mathbb{E}^\pi\left[R_t | S_0 = s_0\right]$ under each stage t. This expression, under Assumption 2, can be further decomposed as

$$\mathbb{E}^\pi\left[R_t | S_0 = s_0\right] = \int R_t(\pi) \cdot \Big\{\mathbb{P}(R_t(\pi)|S_t = s_t) \cdot \mathbb{P}(S_t(\pi)|S_{t-1} = s_{t-1}) \cdots \mathbb{P}(S_1(\pi)|S_0 = s_0)\Big\} ds_0 \cdots ds_t. \tag{13}$$

where $R_t(\pi)$ is defined as the potential reward one would observe under target policy π.

Before we proceed, let's define some short-hand notations that will be widely used in the following sections. First, we define

$$p_1^A(S_t) = \mathbb{E}[A_t | Z_t = 1, S_t], \quad p_0^A(S_t) = \mathbb{E}[A_t | Z_t = 0, S_t], \quad \delta^A(S_t) = p_1^A(S_t) - p_0^A(S_t),$$

where the subscript in $p_1^A(S_t)$ and $p_0^A(S_t)$ indicates the value of Z_t that A_t is conditioning on. Let $Y_t = R_t + \gamma V^\pi(S_{t+1})$, which accounts for the reward at current stage and beyond. We'll see in later sections that instead of using R_t, Y_t will be frequently used in our identification and estimation process. Akin to A_t, one can similarly define

$$\begin{gathered} p_1^Y(S_t) = \mathbb{E}[Y_t | Z_t = 1, S_t], \quad p_0^Y(S_t) = \mathbb{E}[Y_t | Z_t = 0, S_t], \quad \delta^Y(S_t) = p_1^Y(S_t) - p_0^Y(S_t). \\ p_1^S(S_t) = \mathbb{E}[S_{t+1} | Z_t = 1, S_t], \quad p_0^S(S_t) = \mathbb{E}[S_{t+1} | Z_t = 0, S_t]. \end{gathered}$$

Finally, we let $p_r(r_t | a_t, S_t) = \mathbb{P}_r(R_t = r_t | A_t = a_t, S_t)$, $p_a(a_t | z_t, S_t) = \mathbb{P}_a(A_t = a_t | Z_t = z_t, S_t)$, and $p_z(z_t | S_t) = \mathbb{P}(Z_t = z_t | S_t)$ denote the probability density function for R_t, A_t and Z_t respectively.

According to (13), the identification procedure of $V^\pi(s_0)$ can be conducted stage-by-stage. In the following steps, we will first focus on the identifiability of each term on the right hand side of Formula (13), and summarize our identification results in Step 3.

Step 1. Identifiability of $\mathbb{E}\left[R_t(\pi) | S_t\right]$.

First, suppose $a \in [0,1]$ is a constant, and we define $C(z) := 2az + 1 - z - a$ as a function of z. For the simplicity of notations, we will drop $S_t = s_t$ whenever calculating a conditional expectation, and omit $\bar{a}_{t-1}$ in $R_t(1, \bar{a}_{t-1})$ and $S_{t+1}(1, \bar{a}_{t-1})$.

$$\begin{aligned} & a\mathbb{E}[R_t | Z_t = 1] + (1-a)\mathbb{E}[R_t | Z_t = 0] \\ & = \sum_{z_t = 0,1} C(z_t) \cdot \mathbb{E}[R_t | Z_t = z_t] \overset{A1(a)}{=} \sum_{z_t=0,1} C(z_t) \mathbb{E}_{U_t}\left[\mathbb{E}[R_t | Z_t = z_t, U_t]\right] \\ & = \sum_{z=0,1} C(z) \mathbb{E}_{U_t}\Big[\mathbb{E}[R_t A_t | Z_t = z_t, U_t] + \mathbb{E}[R_t(1-A_t) | Z_t = z_t, U_t]\Big] \\ & \overset{A1(d)}{=} \mathbb{E}_{U_t} \sum_{z_t=0,1} C(z)\Big\{\mathbb{E}[R_t(1) | Z_t = z_t, U_t]\mathbb{P}(A_t = 1 | Z_t = z_t, U_t) + \mathbb{E}[R_t(0) | Z_t = z_t, U_t]\mathbb{P}(A_t = 0 | Z_t = z_t, U_t)\Big\} \\ & \overset{A1(d)}{=} \mathbb{E}_{U_t} \sum_{z_t=0,1} C(z)\Big\{\mathbb{E}[R_t(1) | U_t]\mathbb{P}(A_t = 1 | Z_t = z_t, U_t) + \mathbb{E}[R_t(0) | U_t]\mathbb{P}(A_t = 0 | Z_t = z_t, U_t)\Big\} \\ & = \mathbb{E}_{U_t}\Big\{\mathbb{E}[R_t(0) | U_t] + (\mathbb{E}[R_t(1) - R_t(0) | U_t])(a\mathbb{E}[A_t | Z_t = 1, U_t] + (1-a)\mathbb{E}[A_t | Z_t = 0, U_t])\Big\} \\ & \overset{A1(g)}{=} \mathbb{E}[R_t(0)] + \mathbb{E}[R_t(1) - R_t(0)] \cdot \left\{a\mathbb{E}[A_t | Z_t = 1] + (1-a)\mathbb{E}[A_t | Z_t = 0]\right\}. \end{aligned}$$

When $a = 1$, we have

$$\begin{aligned} \mathbb{E}[R_t | Z_t = 1] & = \mathbb{E}_{U_t}\Big\{\mathbb{P}(A_t = 0 | Z_t = 1, U_t) \cdot \mathbb{E}[R_t(0) | U_t] + \mathbb{P}(A_t = 1 | Z_t = 1, U_t) \cdot \mathbb{E}[R_t(1) | U_t]\Big\} \\ & \overset{A1(g)}{=} \mathbb{E}_{U_t}\Big\{\mathbb{P}(A_t = 0 | Z_t = 1) \cdot \mathbb{E}[R_t(0) | U_t] + \mathbb{P}(A_t = 1 | Z_t = 1) \cdot \mathbb{E}[R_t(1) | U_t]\Big\} \\ & = \mathbb{P}(A_t = 0 | Z_t = 1) \cdot \mathbb{E}[R_t(0)] + \mathbb{P}(A_t = 1 | Z_t = 1) \cdot \mathbb{E}[R_t(1)] \end{aligned} \tag{14}$$

Similarly, when $a = 0$,

$$\begin{aligned}\mathbb{E}[R_t|Z_t = 0] &= \mathbb{E}_{U_t}\Big\{\mathbb{P}(A_t = 0|Z_t = 0, U_t) \cdot \mathbb{E}[R_t(0)|U_t] + \mathbb{P}(A_t = 1|Z_t = 0, U_t) \cdot \mathbb{E}[R_t(1)|U_t]\Big\} \\ &\overset{A1(g)}{=} \mathbb{E}_{U_t}\Big\{\mathbb{P}(A_t = 0|Z_t = 0) \cdot \mathbb{E}[R_t(0)|U_t] + \mathbb{P}(A_t = 1|Z_t = 0) \cdot \mathbb{E}[R_t(1)|U_t]\Big\} \\ &= \mathbb{P}(A_t = 0|Z_t = 0) \cdot \mathbb{E}[R_t(0)] + \mathbb{P}(A_t = 1|Z_t = 0) \cdot \mathbb{E}[R_t(1)]\end{aligned} \tag{15}$$

By solving for Equation (14) and (15), one can obtain that

$$\begin{aligned}\mathbb{E}[R_t(0)|S_t] &= \frac{\mathbb{E}[R_t|Z_t = 0, S_t] \cdot \mathbb{P}(A_t = 1|Z_t = 1, S_t) - \mathbb{E}[R_t|Z_t = 1, S_t] \cdot \mathbb{P}(A_t = 1|Z_t = 0, S_t)}{\mathbb{P}(A_t = 1|Z_t = 1, S_t) - \mathbb{P}(A_t = 1|Z_t = 0, S_t)}, \\ \mathbb{E}[R_t(1)|S_t] &= \frac{\mathbb{E}[R_t|Z_t = 1, S_t] \cdot \mathbb{P}(A_t = 0|Z_t = 0, S_t) - \mathbb{E}[R_t|Z_t = 0, S_t] \cdot \mathbb{P}(A_t = 0|Z_t = 1, S_t)}{\mathbb{P}(A_t = 1|Z_t = 1, S_t) - \mathbb{P}(A_t = 1|Z_t = 0, S_t)}.\end{aligned}$$

Notice that all of the terms on the RHS of the expression above can be estimated from the observed data. Specifically, $\mathbb{E}[R_t|Z = 1, S_t]$ can be further written as

$$\mathbb{E}[R_t|Z_t = 1, S_t] = \sum_{a_t, r_t} r_t \cdot \mathbb{P}_r(R_t = r_t|Z_t = 1, A_t = a_t, S_t) \cdot \mathbb{P}_a(A_t = a_t|Z_t = 1, S_t)$$

which is also estimable. To sum up, the identification result of $\mathbb{E}\left[R_t(\pi)|S_t\right]$ is given by

$$\begin{aligned}&\mathbb{E}\left[R_t(\pi)|S_t\right] = \pi(1|S_t) \cdot \mathbb{E}\left[R_t(1)|S_t\right] + (1 - \pi(1|S_t)) \cdot \mathbb{E}\left[R_t(0)|S_t\right] \\ &= \frac{\mathbb{E}[R_t|Z_t = 0, S_t] \cdot \{p_1^A(S_t) - \pi(1|S_t)\} - \mathbb{E}[R_t|Z_t = 1, S_t] \cdot \{p_0^A(S_t) - \pi(1|S_t)\}}{\delta^{A_t}(S_t)}.\end{aligned} \tag{16}$$

Therefore, $\mathbb{E}\left[R_t(\pi_t)|S_t\right]$ is identifiable.

Step 2: Identifiability of $\mathbb{P}[S_t(\pi)|S_{t-1} = s_{t-1}]$.

Under the potential outcome's framework, it holds for $\mathbb{P}(S_t(\pi)|S_{t-1})$ that

$$\mathbb{P}(S_t(\pi)|S_{t-1}) = \pi(S_{t-1}) \cdot \mathbb{P}(S_t(1)|S_{t-1}) + (1 - \pi(S_{t-1})) \cdot \mathbb{P}(S_t(0)|S_{t-1}).$$

In order to indentify $\mathbb{P}(S_t(\pi)|S_{t-1})$, it suffice to derive the identification result for both $\mathbb{P}(S_t(1)|S_{t-1})$ and $\mathbb{P}(S_t(0)|S_{t-1})$. Akin to what we did in Step 1, under Assumption 1 (a), (b), (f) and (g), we can show that

$$\begin{aligned}\mathbb{P}[S_t(0)|S_{t-1}] &= \frac{\mathbb{P}[S_t|Z_{t-1} = 0, S_{t-1}] \cdot p_1^{A_t}(S_{t-1}) - \mathbb{P}[S_t|Z_{t-1} = 1, S_{t-1}] \cdot p_0^{A_t}(S_{t-1})}{\delta^{A_t}(S_{t-1})} \\ \mathbb{P}[S_t(1)|S_{t-1}] &= \frac{\mathbb{P}[S_t|Z_{t-1} = 1, S_{t-1}] \cdot (1 - p_0^{A_t}(S_{t-1})) - \mathbb{P}[S_t|Z_{t-1} = 0, S_{t-1}] \cdot (1 - p_1^{A_t}(S_{t-1}))}{\delta^{A_t}(S_{t-1})}.\end{aligned}$$

Thus, the identification result of $\mathbb{P}(S_t(\pi)|S_{t-1} = s_{t-1})$ is given by

$$\begin{aligned}&\mathbb{P}(S_t(\pi)|S_{t-1} = s_{t-1}) = \pi(S_{t-1}) \cdot \mathbb{P}\left[S_t(1)|S_{t-1}\right] + (1 - \pi(S_{t-1})) \cdot \mathbb{P}\left[S_t(0)|S_{t-1}\right] \\ &= \frac{p_0^S(s_{t-1}) \cdot \{p_1^A(S_{t-1}) - \pi(S_{t-1})\} - p_1^S(s_{t-1}) \cdot \{p_0^A(S_{t-1}) - \pi(S_{t-1})\}}{\delta^A(S_{t-1})}.\end{aligned} \tag{17}$$

Step 3: By repeating the procedure in Step 2, it's easy to show that all of the terms on the RHS of (13) are identifiable.

Define a weighted function $c(z|s)$ as

$$c(z_t|S_t) = \begin{cases} \dfrac{\mathbb{E}[A_t|Z_t = 1, S_t] - \pi(1|S_t)}{\mathbb{E}[A_t|Z_t = 1, S_t] - \mathbb{E}[A_t|Z_t = 0, S_t]}, & \text{when } z_t = 0 \\ \dfrac{\pi(1|S_t) - \mathbb{E}[A_t|Z_t = 0, S_t]}{\mathbb{E}[A_t|Z_t = 1, S_t] - \mathbb{E}[A_t|Z_t = 0, S_t]}, & \text{when } z_t = 1 \end{cases}.$$

Therefore, we can rewrite (16) and (17) as

$$\begin{aligned}\mathbb{E}\left[R_t(\pi)|S_t\right] &= \sum_{a_t,r_t} r_t \cdot \Big\{c(Z_t=0|S_t)\cdot p_r(r_t|a_t,Z_t=0,S_t)\cdot p_a(a_t|Z_t=0,S_t)\\ &\qquad\qquad + c(Z_t=1|S_t)\cdot p_r(r_t|a_t,Z_t=1,S_t)\cdot p_a(a_t|Z_t=1,S_t)\Big\}\\ &= \sum_{z_t,a_t,r_t} r_t\cdot p_r(r_t|a_t,z_t,S_t)\cdot p_a(a_t|z_t,S_t)\cdot c(z_t|S_t)\\ &= \sum_{z_t,a_t,r_t,s_{t+1}} r_t\cdot p_{r,s}(r_t,s_{t+1}|a_t,z_t,S_t)\cdot p_a(a_t|z_t,S_t)\cdot c(z_t|S_t),\end{aligned}$$

and similarly,

$$\begin{aligned}\mathbb{P}\left[S_t(\pi)|S_{t-1}\right] &= \sum_{a_{t-1},s_t} \Big\{c(Z_{t-1}=0|S_{t-1})\cdot p_s(s_t|a_{t-1},Z_{t-1}=0,S_{t-1})\cdot p_a(a_{t-1}|Z_{t-1}=0,S_{t-1})\\ &\qquad + c(Z_{t-1}=1|S_{t-1})\cdot p_s(s_t|a_{t-1},Z_{t-1}=1,S_{t-1})\cdot p_a(a_{t-1}|Z_{t-1}=1,S_{t-1})\Big\}\\ &= \sum_{z_{t-1},a_{t-1},s_t} p_s(s_t|a_{t-1},z_{t-1},S_{t-1})\cdot p_a(a_{t-1}|z_{t-1},S_{t-1})\cdot c(z_{t-1}|S_{t-1})\\ &= \sum_{z_{t-1},a_{t-1},r_{t-1},s_t} p_{r,s}(r_{t-1},s_t|a_{t-1},z_{t-1},S_{t-1})\cdot p_a(a_{t-1}|z_{t-1},S_{t-1})\cdot c(z_{t-1}|S_{t-1}).\end{aligned}$$

Repeating this process until $t=1$, we have

$$\begin{aligned}\mathbb{E}^{\pi}\left[R_t|S_0=s_0\right] &= \int R_t(\pi)\cdot\Big\{\mathbb{P}(R_t(\pi)|S_t=s_t)\cdot\mathbb{P}(S_t(\pi)|S_{t-1}=s_{t-1})\cdots\mathbb{P}(S_1(\pi)|S_0=s_0)\Big\}ds_0\cdots ds_t.\\ &= \sum_{\{z_j,a_j,r_j,s_{j+1}\}_{j=0}^t} r_t\cdot\left\{\prod_{j=0}^t p_{r,s}(r_j,s_{j+1}|a_j,z_j,s_j)\cdot p_a(a_j|z_j,s_j)\cdot c(z_j|s_j)\right\}\end{aligned}$$

Therefore, the value function $V^{\pi}(s_0)$ can be written as

$$\begin{aligned}V^{\pi}(s_0) &= \sum_{t=0}^{\infty}\gamma^t\mathbb{E}^{\pi}\left[R_t|S_0=s_0\right]\\ &= \sum_{t=0}^{T}\sum_{\{z_j,a_j,r_j,s_{j+1}\}_{j=0}^t}\gamma^t r_t\cdot\left\{\prod_{j=0}^t p_{r,s}(r_j,s_{j+1}|a_j,z_j,s_j)\cdot p_a(a_j|z_j,s_j)\cdot c(z_j|s_j)\right\},\end{aligned}$$

where the RHS is purely constructed from the observed data. Furthermore, the identification result of η^{π} can be obtained by taking the expectation of $V^{\pi}(s_0)$ w.r.t. the initial state distribution $\nu(s_0)$, which is given by

$$\eta^{\pi} = \sum_{s_0}\nu(s_0)\cdot\left[\sum_{t=0}^{T}\sum_{\{z_j,a_j,r_j,s_{j+1}\}_{j=0}^t}\gamma^t r_t\cdot\left\{\prod_{j=0}^t p_{r,s}(r_j,s_{j+1}|a_j,z_j,s_j)\cdot p_a(a_j|z_j,s_j)\cdot c(z_j|s_j)\right\}\right]. \tag{18}$$

Therefore, $V^{\pi}(s_0)$ and η^{π} are identifiable.

F. Proof of Theorem 5.1

Define $\mathcal{M}$ as the collection of all models involved in estimating p_s, p_r, p_a, p_z, Q^{π}, ω^{π} and ν. Specifically, we suppose that there exists a parameter θ_0 such that p_{s,θ_0}, p_{r,θ_0}, p_{a,θ_0}, p_{z,θ_0}, $Q^{\pi}_{\theta_0}$, $\omega^{\pi}_{\theta_0}$ and ν_{θ_0} correspond to the true models. As such, the aggregated value η_{π} can also be written as a function of θ, where we denote it as η^{π}_{θ}. Furthermore, we define $p_D(s)=\frac{1}{T}\sum_{t=0}^{T-1}p_t^b(s)$ as the mixture distribution of the observed data. When the stochastic process is stationary, $p_D(s)=p_{\infty}(s)$.

To find the efficient influence function (EIF) for η^π, we first need to derive the canonical gradient in the nonparametric model $\mathcal{M}$. In our proof, we focus on deriving the explicit expression of the Gateaux derivative, which we find the easiest to start with. Specifically, the EIF is calculated as the unique mean zero function $\phi(D_t)$ whose inner product (or covariance) with the score function $S(D_t) := \partial l(\{D_t\};\theta)/\partial\theta_0$ equals the pathwise derivative $\partial\eta_\theta^\pi/\partial\theta_0$.

The proof of this theorem consists of five steps. Step 1 establishes the overall structure of the proof, while Steps 2 to 4 focus on deriving the Gateaux derivative in detail for each subpart identified in Step 1. Lastly, in Step 5, we provide a summary where the final efficient influence function is obtained.

Step 1. By definition, the Cramer-Rao Lower Bound of η_θ^π is given by

$$CR(\mathcal{M}) = \left(\frac{\partial\eta_\theta^\pi}{\partial\theta_0}\right)\left(\mathbb{E}\left[\frac{\partial l(\{D_t\};\theta)}{\partial\theta_0}\frac{\partial^T l(\{D_t\};\theta)}{\partial\theta_0}\right]\right)^{-1}\left(\frac{\partial\eta_\theta^\pi}{\partial\theta_0}\right)^T,$$

where $l(\{D_t\};\theta)$ denotes the log-likelihood function of the observed data, which can be expressed as

$$l(\{D_t\};\theta) = \sum_{t=0}^{T-1}\log p_{s,r,\theta}(S_{t+1},R_j|A_t,Z_t,S_t) + \sum_{t=0}^{T-1}\log p_{a,\theta}(A_t|Z_t,S_t) + \sum_{t=0}^{T-1}\log p_{z,\theta}(Z_t|S_t) + \log\nu_\theta(S_0).$$

To finish the proof of this theorem, it suffice to show that $\sigma_T^2 = \sup_{\mathcal{M}} CR(\mathcal{M})$. If we can show that

$$\frac{\partial\eta_\theta^\pi}{\partial\theta_0} = \mathbb{E}\left[\left\{\sum_{z,a,S_0} c(z|S_0)\cdot p_a(a|z,S_0)\cdot Q^\pi(S_0,z,a) + \frac{1}{T}\sum_{t=0}^{T-1}\phi(D_t)\right\}\frac{\partial l(\{D_t\};\theta)}{\partial\theta_0}\right], \tag{19}$$

where $\phi(D_t)$ is defined in Equation (8), then

$$\frac{\partial\eta_\theta^\pi}{\partial\theta_0} = \mathbb{E}\left[\left\{\sum_{z,a,S_0} c(z|S_0)\cdot p_a(a|z,S_0)\cdot Q^\pi(S_0,z,a) + \frac{1}{T}\sum_{t=0}^{T-1}\phi(D_t) - \eta_\theta^\pi\right\}\frac{\partial l(\{D_t\};\theta)}{\partial\theta_0}\right],$$

since $\eta_\theta^\pi\cdot\mathbb{E}\left[\frac{\partial l(\{D_t\};\theta)}{\partial\theta_0}\right] = 0$ according to the property of score function. Furthermore, by Cauchy-Schwarz Inequality,

$$\sigma_T^2 \le \sup_{\mathcal{M}} CR(\mathcal{M}) \le \mathbb{E}\left\{\sum_{z,a,S_0} c(z|S_0)\cdot p_a(a|z,S_0)\cdot Q^\pi(S_0,z,a) + \frac{1}{T}\sum_{t=0}^{T-1}\phi(D_t) - \eta_\theta^\pi\right\}^2 = \sigma_T^2,$$

which implies $\sigma_T^2 = \sup_{\mathcal{M}} CR(\mathcal{M}) = CR(\mathcal{M}_0)$. (See Lemma 20 of (Kallus & Uehara, 2022) for details.)

Therefore, it remains to show that equation (19) holds. That is,

$$\frac{\partial\eta_\theta^\pi}{\partial\theta_0} = \mathbb{E}\left[\left\{\sum_{z,a,S_0} c(z|S_0)\cdot p_a(a|z,S_0)\cdot Q^\pi(S_0,z,a) + \frac{1}{T}\sum_{t=0}^{T-1}\phi(D_t)\right\}\frac{\partial l(\{D_t\};\theta)}{\partial\theta_0}\right].$$

Notice that in Theorem 1, we've proved that

$$\eta_{\theta_0}^\pi = \sum_{s_0}\left[\sum_{t=0}^{+\infty}\gamma^t\sum_{\tau_t,s_{t+1}} r_t\left\{\prod_{j=0}^{t} p_{\theta_0}(s_{j+1},r_j,a_j,z_j|s_j)\right\}\cdot\nu_{\theta_0}(s_0)\right],$$

where p_θ is the probability of a trajectory that follows target policy π. That is,

$$p_{\theta_0}(s_{j+1},r_j,a_j,z_j|s_j) = c(z_j|s_j)\cdot p_a(a_j|z_j,s_j)\cdot p_{r,s}(s_{j+1},r_j|a_j,z_j,s_j).$$

Since $\nabla_\theta p_\theta = p_\theta\nabla_\theta\log(p_\theta)$, we can decompose $\nabla_\theta\eta_{\theta_0}^\pi$ as

$$\begin{aligned}\nabla_\theta\eta_{\theta_0}^\pi = &\sum_{s_0}\left[\sum_{t=0}^{+\infty}\gamma^t\sum_{\tau_t,s_{t+1}} r_t\left\{\prod_{j=0}^{t} p_{\theta_0}(s_{j+1},r_j,a_j,z_j|s_j)\right\}\cdot\left\{\sum_{j=0}^{t}\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j,a_j,z_j|s_j)\right\}\cdot\nu_{\theta_0}(s_0)\right]\\ &+\sum_{s_0}\left[\sum_{t=0}^{+\infty}\gamma^t\sum_{\tau_t,s_{t+1}} r_t\left\{\prod_{j=0}^{t} p_{\theta_0}(s_{j+1},r_j,a_j,z_j|s_j)\right\}\cdot\nu_{\theta_0}(s_0)\cdot\nabla_\theta\log\nu_{\theta_0}(s_0)\right] := L_1 + L_2.\end{aligned} \tag{20}$$

The first line of the equation above can be further derived as

$$
\begin{aligned}
L_1 &= \sum_{s_0} \sum_{j=0}^{+\infty} \left\{ \sum_{t=j}^{+\infty} \gamma^t \sum_{\tau_t, s_{t+1}} \left[\prod_{k=0}^{t} p_{\theta_0}(s_{k+1}, r_k, a_k, z_k | s_k) \right] \right\} \cdot \nabla_\theta \log p_{\theta_0}(s_{j+1}, r_j, a_j, z_j | s_j) \Big\} \cdot \nu_{\theta_0}(s_0) \\
&= \sum_{s_0} \sum_{j=0}^{+\infty} \gamma^j \cdot \sum_{\tau_t, s_{t+1}} \left[\prod_{k=0}^{j} p_{\theta_0}(s_{k+1}, r_k, a_k, z_k | s_k) \right] \cdot \Bigg\{ r_j + \sum_{t=j+1}^{\infty} \gamma^{t-j} \sum_{\tau_{t/j}, s_{t+1}} r_t \cdot \\
&\qquad \left[\prod_{k=j+1}^{t} p_{\theta_0}(s_{k+1}, r_k, a_k, z_k | s_k) \right] \Bigg\} \cdot \nabla_\theta \log p_{\theta_0}(s_{j+1}, r_j, a_j, z_j | s_j) \cdot \nu_{\theta_0}(s_0) \\
&= \sum_{s_0} \sum_{j=0}^{+\infty} \gamma^j \sum_{\tau_j, s_{j+1}} \Big\{ r_j + \gamma \cdot V^\pi(S_{j+1}) \Big\} \cdot \left[\prod_{k=0}^{j} p_{\theta_0}(s_{k+1}, r_k, a_k, z_k | s_k) \right] \cdot \\
&\qquad \nabla_\theta \log p_{\theta_0}(s_{j+1}, r_j, a_j, z_j | s_j) \cdot \nu_{\theta_0}(s_0).
\end{aligned}
$$

Notice that

$$
\begin{aligned}
&\sum_{s_0} \left[\prod_{k=0}^{j} p_{\theta_0}(s_{k+1}, r_k, a_k, z_k | s_k) \right] \cdot \nu_\theta(s_0) = p_{\theta_0}(s_{j+1}, r_j, a_j, z_j | s_j) \cdot p^\pi_{t,\theta_0}(s_j) \\
&= p_{r,s}(s_{j+1}, r_j | a_j, z_j, s_j) \cdot p_a(a_j | z_j, s_j) \cdot c(z_j | s_j) \cdot p^\pi_{t,\theta_0}(s_j),
\end{aligned}
$$

then by plugging in this expression to L_1, we have

$$
\begin{aligned}
L_1 &= \sum_{j=0}^{+\infty} \gamma^j \sum_{\tau_j, s_{j+1}} \Big\{ r_j + \gamma \cdot V^\pi(S_{j+1}) \Big\} \cdot p_{\theta_0}(s_{j+1}, r_j, a_j, z_j | s_j) \cdot p^\pi_{t,\theta_0}(s_j) \cdot \nabla_\theta \log p_{\theta_0}(s_{j+1}, r_j, a_j, z_j | s_j) \\
&= \sum_{j=0}^{+\infty} \gamma^j \sum_{\tau_j, s_{j+1}} \Big\{ r_j + \gamma \cdot V^\pi(S_{j+1}) \Big\} \cdot p_{\theta_0}(s_{j+1}, r_j, a_j, z_j | s_j) \cdot p^\pi_{t,\theta_0}(s_j) \cdot \nabla_\theta \log p_{s,r,\theta_0}(s_{j+1}, r_j |, a_j, z_j, s_j) \\
&+ \sum_{j=0}^{+\infty} \gamma^j \sum_{\tau_j, s_{j+1}} \Big\{ r_j + \gamma \cdot V^\pi(S_{j+1}) \Big\} \cdot p_{\theta_0}(s_{j+1}, r_j, a_j, z_j | s_j) \cdot p^\pi_{t,\theta_0}(s_j) \cdot \nabla_\theta \log p_{a,\theta_0}(a_j | z_j, s_j) \\
&+ \sum_{j=0}^{+\infty} \gamma^j \sum_{\tau_j, s_{j+1}} \Big\{ r_j + \gamma \cdot V^\pi(S_{j+1}) \Big\} \cdot p_{\theta_0}(s_{j+1}, r_j, a_j, z_j | s_j) \cdot p^\pi_{t,\theta_0}(s_j) \cdot \nabla_\theta \log c(z_j | s_j) \\
&:= (1.1) + (1.2) + (1.3),
\end{aligned}
$$

where the last three lines corresponds to (1.1), (1.2) and (1.3), respectively. We will deal with the detailed expression of these three terms in the next few steps.

Finally, let's focus on the second line of Formula (20). This part, according to our identification result in Theorem 4.1, is equivalent to $\mathbb{E}[V^\pi(S_0) \nabla_\theta \log \nu_{\theta_0}(S_0)]$. Still, by using the fact that the expectation of a score function is 0, we have

$$
\begin{aligned}
L_2 &= \sum_{s_0} \left[\sum_{t=0}^{+\infty} \gamma^t \sum_{\tau_t, s_{t+1}} r_t \left\{ \prod_{j=0}^{t} p_{\theta_0}(s_{j+1}, r_j, a_j, z_j | s_j) \right\} \cdot \nu_{\theta_0}(s_0) \cdot \nabla_\theta \log \nu_{\theta_0}(s_0) \right] \\
&= \mathbb{E}[V^\pi(S_0) \nabla_\theta \log \nu_{\theta_0}(S_0)] = \mathbb{E}\Big[\big(V^\pi(S_0) - \eta^\pi_{\theta_0} \big) \nabla_\theta \log \nu_{\theta_0}(S_0) \Big] \\
&= \mathbb{E}\Big[\big(V^\pi(S_0) - \eta^\pi_{\theta_0} \big) S(\bar{O}_{T-1}) \Big],
\end{aligned}
$$

where $S(s', r, a, z | s)$ denotes the conditional score function of $(S_{t+1}, R_t, A_t, Z_t) | S_t$, and we use $\bar{O}_{T-1}$ to denote all observed data up to stage $T-1$.

The rest of the proof are organized as follows. In Step 2 to Step 4, we will derive the expression for (1.1), (1.2) and (1.3) respectively. In Step 5, we will focus back on Formula (20) to finish the proof of the whole theorem.

Step 2: Derivation of (1.1).
Using the similar trick in deriving L_2, we have

$$\begin{aligned}
(1.1) = &\sum_{j=0}^{+\infty}\gamma^j \sum_{\tau_j, s_{j+1}} \Big\{r_j + \gamma \cdot V^\pi(s_{j+1})\Big\} \cdot p_{r,s}(s_{j+1}, r_j|a_j, z_j, s_j) \cdot p_a(a_j|z_j, s_j) \cdot c(z_j|s_j) \cdot p^\pi_{j,\theta_0}(s_j) \\
&\qquad \cdot \nabla_\theta \log p_{s,r,\theta_0}(s_{j+1}, r_j|, a_j, z_j, s_j) \\
= &\sum_{j=0}^{+\infty}\gamma^j \sum_{\tau_j, s_{j+1}} \Big\{r_j + \gamma \cdot V^\pi(s_{j+1}) - Q^\pi(s_j, z_j, a_j)\Big\} \cdot p_{r,s}(s_{j+1}, r_j|a_j, z_j, s_j) \cdot p_a(a_j|z_j, s_j) \\
&\qquad \cdot c(z_j|s_j) \cdot p^\pi_{j,\theta_0}(s_j) \cdot \nabla_\theta \log p_{s,r,\theta_0}(s_{j+1}, r_j|, a_j, z_j, s_j) \\
= &\sum_{j=0}^{+\infty}\gamma^j \sum_{\tau_j, s_{j+1}} \Big\{r_j + \gamma \cdot V^\pi(s_{j+1}) - Q^\pi(s_j, z_j, a_j)\Big\} \cdot p_{r,s}(s_{j+1}, r_j|a_j, z_j, s_j) \cdot p_a(a_j|z_j, s_j) \\
&\qquad \cdot c(z_j|s_j) \cdot p^\pi_{j,\theta_0}(s_j) \cdot \nabla_\theta \log p_{\theta_0}(s_{j+1}, r_j, a_j, z_j|s_j) \\
= &\sum_{j=0}^{+\infty}\gamma^j \sum_{\tau_j, s_{j+1}} \Big\{r_j + \gamma \cdot V^\pi(s_{j+1}) - Q^\pi(s_j, z_j, a_j)\Big\} \frac{c(z_j|s_j)}{p_z(z_j|s_j)} \cdot p_{r,s}(s_{j+1}, r_j|a_j, z_j, s_j) \\
&\qquad \cdot p_a(a_j|z_j, s_j) \cdot p_z(z_j|s_j) \cdot p^\pi_{j,\theta_0}(s_j) \cdot \nabla_\theta \log p_{\theta_0}(s_{j+1}, r_j, a_j, z_j|s_j)
\end{aligned}$$

$$= \sum_{j=0}^{+\infty}\gamma^j \mathbb{E}\left[\Big\{R_j + \gamma \cdot V^\pi(S_{j+1}) - Q^\pi(S_j, Z_j, A_j)\Big\}\frac{c(Z_j|S_j)}{p_z(Z_j|S_j)} \cdot \frac{p^\pi_{j,\theta_0}(S_j)}{p_D(S_j)} \cdot S(S_{j+1}, R_j, A_j, Z_j)\right].$$

Recall that

$$\omega^\pi(s) = (1-\gamma) \cdot \sum_{t=0}^{+\infty}\gamma^t \frac{p^\pi_t(s)}{p_D(s)}, \quad \rho(z, s) = \frac{c(z|s)}{p_z(z|s)}.$$

Then by Markov property, we have

$$\begin{aligned}
(1.1) &= (1-\gamma)^{-1}\mathbb{E}\left[\Big\{R + \gamma \cdot V^\pi(S') - Q^\pi(S, Z, A)\Big\}\rho(S, Z) \cdot \omega^\pi(S) \cdot S(\bar{O}_{T-1})\right] \\
&:= \mathbb{E}\Big\{\Big[\frac{1}{T}\sum_t \phi_1(D_t)\Big]S(\bar{O}_{T-1})\Big\},
\end{aligned}$$

where we define

$$\phi_1(D_t) = (1-\gamma)^{-1}\Big\{R_t + \gamma \cdot V^\pi(S_{t+1}) - Q^\pi(S_t, Z_t, A_t)\Big\} \cdot \rho(S_t, Z_t) \cdot \omega^\pi(S_t). \tag{21}$$

Step 3: Derivation of (1.2).

$$\begin{aligned}
(1.2) = &\sum_{j=0}^{+\infty}\gamma^j \sum_{\tau_j, s_{j+1}} \Big\{r_j + \gamma \cdot V^\pi(s_{j+1})\Big\} \cdot p_{\theta_0}(s_{j+1}, r_j, a_j, z_j|s_j) \cdot p^\pi_{t,\theta_0}(s_j) \cdot \nabla_\theta \log p_{a,\theta_0}(a_j|z_j, s_j) \\
= &\sum_{j=0}^{+\infty}\gamma^j \sum_{a_j, z_j, s_j} \mathbb{E}\Big\{\big[R_j + \gamma \cdot V^\pi(S_{j+1})\big]|a_j, z_j, s_j\Big\} \cdot p_{a,\theta_0}(a_j|z_j, s_j) \cdot c(z_j|s_j) \cdot p^\pi_{t,\theta_0}(s_j) \cdot \nabla_\theta \log p_{a,\theta_0}(a_j|z_j, s_j) \\
= &\sum_{j=0}^{+\infty}\gamma^j \sum_{a_j, z_j, s_j} Q^\pi(s_j, z_j, a_j) \cdot p_{a,\theta_0}(a_j|z_j, s_j) \cdot c(z_j|s_j) \cdot p^\pi_{t,\theta_0}(s_j) \cdot \nabla_\theta \log p_{a,\theta_0}(a_j|z_j, s_j).
\end{aligned}$$

Notice that

$$
\begin{aligned}
&\sum_{j=0}^{+\infty}\gamma^j \sum_{a_j,z_j,s_j} \Big\{\sum_{a_j^*} Q^\pi(s_j,z_j,a_j^*)\cdot p_{a,\theta_0}(a_j^*|z_j,s_j)\Big\}\cdot p_{a,\theta_0}(a_j|z_j,s_j)\cdot c(z_j|s_j)\cdot p^\pi_{t,\theta_0}(s_j)\cdot \nabla_\theta \log p_{a,\theta_0}(a_j|z_j,s_j)\\
&=\sum_{j=0}^{+\infty}\gamma^j \sum_{z_j,s_j} \Big\{\sum_{a_j^*} Q^\pi(s_j,z_j,a_j^*)\cdot p_{a,\theta_0}(a_j^*|z_j,s_j)\Big\}\cdot c(z_j|s_j)\cdot p^\pi_{t,\theta_0}(s_j)\cdot \sum_{a_j} p_{a,\theta_0}(a_j|z_j,s_j)\cdot \nabla_\theta \log p_{a,\theta_0}(a_j|z_j,s_j)\\
&=0,
\end{aligned}
$$

where the last equality holds since the score function has a zero mean.

Combining the formula above with (1.2), we have

$$
\begin{aligned}
(1.2) &= \sum_{j=0}^{+\infty}\gamma^j \sum_{a_j,z_j,s_j} \Big\{Q^\pi(s_j,z_j,a_j) - \sum_{a_j^*} Q^\pi(s_j,z_j,a_j^*)\cdot p_{a,\theta_0}(a_j^*|z_j,s_j)\Big\}\cdot p_{a,\theta_0}(a_j|z_j,s_j)\cdot c(z_j|s_j)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\cdot p^\pi_{t,\theta_0}(s_j)\cdot \nabla_\theta \log p_{a,\theta_0}(a_j|z_j,s_j)\\
&= \sum_{j=0}^{+\infty}\gamma^j \sum_{a_j,z_j,s_j} \Big\{Q^\pi(s_j,z_j,a_j) - \sum_{a_j^*} Q^\pi(s_j,z_j,a_j^*)\cdot p_{a,\theta_0}(a_j^*|z_j,s_j)\Big\}\\
&\cdot \frac{c(z_j|s_j)}{p_z(z_j|s_j)}\cdot p_{a,\theta_0}(a_j|z_j,s_j)\cdot p_z(z_j|s_j)\cdot p^\pi_{t,\theta_0}(s_j)\cdot \nabla_\theta \log p_{a,\theta_0}(a_j|z_j,s_j)\\
&= \sum_{j=0}^{+\infty}\gamma^j \sum_{s_{j+1},r_j,a_j,z_j,s_j} \Big\{Q^\pi(s_j,z_j,a_j) - \sum_{a_j^*} Q^\pi(s_j,z_j,a_j^*)\cdot p_{a,\theta_0}(a_j^*|z_j,s_j)\Big\}\\
&\cdot \rho(z_j,s_j)\cdot p_{s,r,\theta_0}(s_{j+1},r_j|z_j,s_j)\cdot p_{a,\theta_0}(a_j|z_j,s_j)\cdot p_z(z_j|s_j)\cdot p^\pi_{t,\theta_0}(s_j)\cdot \nabla_\theta \log p_{s,r,a,z,\theta_0}(s_{j+1},r_j,a_j,z_j|s_j)\\
&= \sum_{j=0}^{+\infty}\gamma^j \mathbb{E}\left[\Big\{Q^\pi(S,Z,A) - \sum_{a^*} Q^\pi(S,Z,a^*)\cdot p_{a,\theta_0}(a^*|Z,S)\Big\}\cdot \rho(S,Z)\cdot \frac{p^\pi_{t,\theta_0}(S)}{p_D(S)}\cdot S(S',R,A,Z|S)\right]\\
&= (1-\gamma)^{-1}\cdot \mathbb{E}\left[\omega^\pi(S)\rho(S,Z)\Big\{Q^\pi(S,Z,A) - \sum_{a^*} Q^\pi(S,Z,a^*)\cdot p_{a,\theta_0}(a^*|Z,S)\Big\}\cdot S(S',R,A,Z|S)\right]\\
:&= \mathbb{E}\left[\Big\{\frac{1}{T}\sum_t \phi_2(D_t)\Big\}\cdot S(\bar{O}_{T-1})\right],
\end{aligned}
$$

where we define

$$
\phi_2(O) = (1-\gamma)^{-1}\cdot \omega^\pi(S)\cdot \rho(S,Z)\cdot \Big\{Q^\pi(S,Z,A) - \sum_a Q^\pi(S,Z,a)\cdot p_a(a|Z,S)\Big\}. \tag{22}
$$

Step 4: Derivation of (1.3).

$$
\begin{aligned}
(1.3) &= \sum_{j=0}^{+\infty}\gamma^j \sum_{\tau_j,s_{j+1}} \Big\{r_j + \gamma\cdot V^\pi(S_{j+1})\Big\}\cdot p_{\theta_0}(s_{j+1},r_j,a_j,z_j|s_j)\cdot p^\pi_{t,\theta_0}(s_j)\cdot \nabla_\theta \log c(z_j|s_j)\\
&= \sum_{j=0}^{+\infty}\gamma^j \sum_{a_j,z_j,s_j} \mathbb{E}\Big\{\big[R_j + \gamma\cdot V^\pi(S_{j+1})\big]|a_j,z_j,s_j\Big\}\cdot p_{a,\theta_0}(a_j|z_j,s_j)\cdot c(z_j|s_j)\cdot p^\pi_{t,\theta_0}(s_j)\cdot \nabla_\theta \log c(z_j|s_j)\\
&= \sum_{j=0}^{+\infty}\gamma^j \sum_{a_j,z_j,s_j} Q^\pi(s_j,z_j,a_j)\cdot p_{a,\theta_0}(a_j|z_j,s_j)\cdot c(z_j|s_j)\cdot p^\pi_{t,\theta_0}(s_j)\cdot \nabla_\theta \log c(z_j|s_j).
\end{aligned}
$$

Before we proceed, let's calculate $\nabla_\theta \log c(z_j|s_j)$ first. According to the definition of $c(z|s)$,

$$c(z_t|S_t) = \begin{cases} \dfrac{p_1^A(S_t) - \pi(1|S_t)}{\delta^A(S_t)}, & \text{when } z_t = 0 \\ \dfrac{\pi(1|S_t) - p_0^A(S_t)}{\delta^A(S_t)}, & \text{when } z_t = 1 \end{cases}.$$

Therefore,

$$\nabla_\theta c(z_t|S_t) = \begin{cases} \dfrac{\nabla_\theta p_1^A(S_t)\delta^A(S_t) - (p_1^A(S_t) - \pi(1|S_t))\nabla_\theta \delta^A(S_t)}{[\delta^A(S_t)]^2}, & \text{when } z_t = 0 \\ \dfrac{-\nabla_\theta p_0^A(S_t)\delta^A(S_t) - (\pi(1|S_t) - p_0^A(S_t))\nabla_\theta \delta^A(S_t)}{[\delta^A(S_t)]^2}, & \text{when } z_t = 1 \end{cases}.$$

According to (Wang & Tchetgen, 2018),

$$\nabla_\theta p_z^{A_t}(S_t) = \nabla_\theta \mathbb{E}\left[A_t|Z_t = z, S_t\right] = \mathbb{E}\left[(A_t - p_z^A(S_t)) \cdot S(A_t, Z_t = z|S_t)\right],$$

and thus

$$\begin{aligned} \nabla_\theta \delta^A(S_t) &= \mathbb{E}\Big[\frac{2Z_t - 1}{p_z(Z_t|S_t)}\left(A_t - \mathbb{E}[A_t|Z_t, S_t]\right) \cdot S(A_t, Z_t|S_t)\Big|S_t\Big] \\ &= \mathbb{E}\Big[\Big(A_t - p_1^{A_t}(S_t)\Big) \cdot S(A_t, Z_t = 1|S_t) - \Big(A_t - p_0^{A_t}(S_t)\Big) \cdot S(A_t, Z_t = 0|S_t)\Big|S_t\Big]. \end{aligned}$$

Therefore, when $z_t = 0$, $\nabla_\theta c(z_t|S_t)$ can be further derived as

$$\begin{aligned} &\nabla_\theta c(Z_t = 0|S_t) = \frac{1}{\delta^A(S_t)}\Big\{\nabla_\theta p_1^A(S_t) - \frac{p_1^A(S_t) - \pi(1|S_t)}{\delta^A(S_t)} \cdot \nabla_\theta \delta^A(S_t)\Big\} \\ &= \frac{1}{\delta^A(S_t)}\Big\{\mathbb{E}\Big[(A_t - p_1^A(S_t)) \cdot S(A_t, Z_t = 1|S_t) - c(Z_t = 0|S_t) \cdot (A_t - p_1^A(S_t)) \cdot S(A_t, Z_t = 1|S_t) \\ &\qquad + c(Z_t = 0|S_t)(A_t - p_0^A(S_t)) \cdot S(A_t, Z_t = 0|S_t)\Big]\Big\} \\ &= \frac{1}{\delta^A(S_t)}\Big\{\mathbb{E}\Big[c(Z_t = 1|S_t) \cdot (A_t - p_1^A(S_t)) \cdot S(A_t, Z_t = 1|S_t) + c(Z_t = 0|S_t)(A_t - p_0^A(S_t)) \cdot S(A_t, Z_t = 0|S_t)\Big]\Big\} \\ &= \frac{1}{\delta^A(S_t)}\Big\{\mathbb{E}\Big[\frac{c(Z_t|S_t)}{p_z(Z_t|S_t)} \cdot (A_t - p_{Z_t}^{A_t}(S_t)) \cdot S(A_t, Z_t|S_t)\Big]\Big\} = \frac{1}{\delta^A(S_t)}\Big\{\mathbb{E}\Big[\rho(S_t, Z_t) \cdot (A_t - p_{Z_t}^{A_t}(S_t)) \cdot S(A_t, Z_t|S_t)\Big]\Big\}. \end{aligned}$$

Likewise, when $z_t = 1$,

$$\begin{aligned} &\nabla_\theta c(Z_t = 1|S_t) = \frac{-1}{\delta^A(S_t)}\Big\{\nabla_\theta p_0^A(S_t) + \frac{\pi(1|S_t) - p_0^A(S_t)}{\delta^A(S_t)} \cdot \nabla_\theta \delta^A(S_t)\Big\} \\ &= \frac{-1}{\delta^A(S_t)}\Big\{\mathbb{E}\Big[(A_t - p_0^A(S_t)) \cdot S(A_t, Z_t = 0|S_t) + c(Z_t = 1|S_t) \cdot (A_t - p_1^A(S_t)) \cdot S(A_t, Z_t = 1|S_t) \\ &\qquad - c(Z_t = 1|S_t)(A_t - p_0^A(S_t)) \cdot S(A_t, Z_t = 0|S_t)\Big]\Big\} \\ &= \frac{-1}{\delta^A(S_t)}\Big\{\mathbb{E}\Big[c(Z_t = 1|S_t) \cdot (A_t - p_1^A(S_t)) \cdot S(A_t, Z_t = 1|S_t) + c(Z_t = 0|S_t)(A_t - p_0^A(S_t)) \cdot S(A_t, Z_t = 0|S_t)\Big]\Big\} \\ &= \frac{-1}{\delta^A(S_t)}\Big\{\mathbb{E}\Big[\rho(S_t, Z_t) \cdot (A_t - p_{Z_t}^{A_t}(S_t)) \cdot S(A_t, Z_t|S_t)\Big]\Big\}. \end{aligned}$$

Combining the results for both $z_t = 0$ and $z_t = 1$, we can express $\nabla_\theta c(z_t|S_t)$ as

$$\nabla_\theta c(z_t|S_t) = \frac{(-1)^{z_t}}{\delta^A(S_t)} \cdot \mathbb{E}\Big[\rho(S_t, Z_t) \cdot (A_t - p_{Z_t}^{A_t}(S_t)) \cdot S(A_t, Z_t|S_t)\Big].$$

Therefore,

$$\begin{aligned}(1.3) &= \sum_{j=0}^{+\infty}\gamma^j \sum_{a_j,z_j,s_j} Q^\pi(s_j,z_j,a_j)\cdot p_{a,\theta_0}(a_j|z_j,s_j)\cdot c(z_j|s_j)\cdot p^\pi_{t,\theta_0}(s_j)\cdot\nabla_\theta \log c(z_j|s_j)\\ &= \sum_{j=0}^{+\infty}\gamma^j \sum_{a_j,z_j,s_j} Q^\pi(s_j,z_j,a_j)\cdot p_{a,\theta_0}(a_j|z_j,s_j)\cdot p^\pi_{t,\theta_0}(s_j)\cdot\nabla_\theta c(z_j|s_j)\\ &= \sum_{j=0}^{+\infty}\gamma^j \sum_{a_j,z_j,s_j} Q^\pi(s_j,z_j,a_j)\cdot p_{a,\theta_0}(a_j|z_j,s_j)\cdot p^\pi_{t,\theta_0}(s_j)\cdot\frac{(-1)^{z_j}}{\delta^A(s_j)}\cdot\mathbb{E}\Big[\rho(S,Z)\cdot(A-p^A_Z(S))\cdot S(A,Z|S)\Big|S=s_j\Big].\end{aligned}$$

Notice that

$$\sum_{a_j,z_j} Q^\pi(s_j,z_j,a_j)\cdot p_{a,\theta_0}(a_j|z_j,s_j)\cdot\frac{(-1)^{z_j}}{\delta^A(s_j)} = -\Delta(s_j).$$

Thus, by replacing this part in the expression of (1.3), we can obtain

$$\begin{aligned}(1.3) &= \sum_{j=0}^{+\infty}\gamma^j \sum_{a_j,z_j,s_j} Q^\pi(s_j,z_j,a_j)\cdot p_{a,\theta_0}(a_j|z_j,s_j)\cdot p^\pi_{t,\theta_0}(s_j)\cdot\frac{(-1)^{z_j}}{\delta^A(s_j)}\cdot\mathbb{E}\Big[\rho(S,Z)\cdot(A-p^A_Z(S))\cdot S(A,Z|S)\Big|S=s_j\Big]\\ &= -\sum_{j=0}^{+\infty}\gamma^j\sum_{s_j} p^\pi_{t,\theta_0}(s_j)\cdot\Delta(s_j)\cdot\mathbb{E}\Big[\rho(S,Z)\cdot(A-p^A_Z(S))\cdot S(A,Z|S)\Big|S=s_j\Big]\\ &= -(1-\gamma)^{-1}\mathbb{E}\Big[\omega^\pi(S)\cdot\rho(S,Z)\cdot(A-p^A_Z(S))\cdot\Delta(S)\cdot S(A,Z|S)\Big]\\ &= -(1-\gamma)^{-1}\mathbb{E}\Big[\omega^\pi(S)\cdot\rho(S,Z)\cdot(A-p^A_Z(S))\cdot\Delta(S)\cdot S(S',R,A,Z|S)\Big]\\ :&= \mathbb{E}\left[\left\{\frac{1}{T}\sum_t\phi_3(D_t)\right\}\cdot S(\bar{O}_{T-1})\right],\end{aligned}$$

where

$$\phi_3(D_t) = -(1-\gamma)^{-1}\cdot\omega^\pi(S_t)\cdot\rho(S_t,Z_t)\cdot(A_t-\mathbb{E}[A_t|S_t,Z_t])\cdot\Delta(S_t). \tag{23}$$

Step 5: Summary.
By change of measure theorem, we can obtain that

$$\begin{aligned}V^\pi(S_0) &= \sum_{t=0}^{\infty}\gamma^t\mathbb{E}^\pi[R_t|S_0] = \mathbb{E}_{S_t\sim p^\pi_t(s)}\Big[\sum_{t=0}^{\infty}\gamma^t\cdot\mathbb{E}^\pi[R_t|S_t]\Big]\\ &= \mathbb{E}_{S_t\sim p_\infty(s)}\Big[\sum_{t=0}^{\infty}\gamma^t\cdot\frac{p^\pi_t(S_t)}{p_\infty(S_t)}\cdot\mathbb{E}^\pi[R_t|S_t]\Big] = (1-\gamma)^{-1}\mathbb{E}_{S_t\sim p_\infty(s)}\Big[\omega^\pi(S_t)\cdot\mathbb{E}^\pi[R_t|S_t]\Big]\\ &= (1-\gamma)^{-1}\mathbb{E}_{S_t\sim p_\infty(s)}\left[\sum_{z_t}\omega^\pi(S_t)\cdot c(z_t|S_t)\cdot\mathbb{E}[R_t|Z_t=z_t,S_t]\right].\end{aligned} \tag{24}$$

Therefore, we can further write L_2 as

$$L_2 = \mathbb{E}\Big[\big(V^\pi(S_0)-\eta^\pi_{\theta_0}\big)S(\bar{O}_{T-1})\Big] = \mathbb{E}\Big[\Big\{(1-\gamma)^{-1}\sum_{z_t}\omega^\pi(S_t)c(z_t|S_t)\mathbb{E}[R_t|Z_t=z_t,S_t]-\eta^\pi\Big\}\cdot S(\bar{O}_{T-1})\Big]. \tag{25}$$

Then, by combining the results in Formula (21), (22), (23) and (25), we have

$$
\begin{aligned}
\nabla_\theta \eta^\pi_{\theta_0} &= \mathbb{E}\Big[\Big\{\sum_{j=1}^{3}\phi(D_t) + \sum_{z_t}\omega^\pi(S_t)c(z_t|S_t)\mathbb{E}[R_t|Z_t = z_t, S_t] - \eta^\pi\Big\}S(\bar{O}_{T-1})\Big] \\
&= \mathbb{E}\Bigg[\Big\{(1-\gamma)^{-1}\cdot\omega^\pi(S_t)\cdot\Big[\rho(S_t, Z_t)\cdot\Big\{R_t + \gamma\cdot V^\pi(S_{t+1}) - \mathbb{E}\big[R_t + \gamma\cdot V^\pi(S_{t+1})\big|Z_t, S_t\big] \\
&\quad - \big(A_t - \mathbb{E}\left[A_t\middle|Z_t, S_t\right]\big)\cdot\Delta(S_t)\Big\} + \sum_{z_t} c(z_t|S_t)\cdot\mathbb{E}[R_t|z_t, S_t]\Big] - \eta^\pi\Big\}S(\bar{O}_{T-1})\Bigg].
\end{aligned}
$$

It is not surprising to see that augmentation term $\phi(D_t)$ defined in Equation (8) is just the summation of ϕ_1, ϕ_2 and ϕ_3 derived in Formula (21), (22), and (23). The canonical gradient contributed by Formula (25) serves as a direct estimator in the final formula of the efficient influence function.

Therefore, the EIF of η^π is given by

$$
\begin{aligned}
\text{EIF}_{\eta^\pi} =& (1-\gamma)^{-1}\cdot\omega^\pi(S_t)\cdot\Big[\rho(S_t, Z_t)\cdot\Big\{R_t + \gamma\cdot V^\pi(S_{t+1}) - \mathbb{E}\big[R_t + \gamma\cdot V^\pi(S_{t+1})\big|Z_t, S_t\big] \\
& - \big(A_t - \mathbb{E}\left[A_t\middle|Z_t, S_t\right]\big)\cdot\Delta(S_t)\Big\} + \sum_{z_t} c(z_t|S_t)\cdot\mathbb{E}[R_t|z_t, S_t]\Big] - \eta^\pi.
\end{aligned}
$$

The proof of Theorem 2 is thus complete.

G. Proof of Theorem 5.2

We divide the proof of Theorem 3 into two parts. In Section G.1, we mainly discuss the consistency of $\widehat{\eta}_{\text{DR}}$ under model $\mathcal{M}_1$ or $\mathcal{M}_2$. In Section G.2, we will prove that $\widehat{\eta}_{\text{DR}}$ is asymptotically normal with variance σ_T^2, which attains the semiparametric lower bound.

G.1. Consistency

Scenario 1: Suppose the models in $\mathcal{M}_1$ and p_a are correctly specified.

In this case, $c(z|s)$, $V^\pi(S_{t+1})$ and $\Delta(S_t)$ are all correctly specified, since they can all be estimated from $Q^\pi(s, z, a)$ and $p_a(a|s, z)$. Therefore,

$$
\mathbb{E}\left[\widehat{\eta}^*\right] = \mathbb{E}\left[\widehat{\eta}^\pi_{\text{DM}} + (NT)^{-1}\sum_{i,t}\widehat{\phi}(D_{i,t})\right] = \widehat{\eta}^\pi_{\text{DM}} + \mathbb{E}\left[\widehat{\phi}(D_t)\right]
$$

Now we only need to prove that $\widehat{\eta}^\pi_{\text{DM}}$ is consistent to η^π. Since $\widehat{c}(z|s)$, $\widehat{p}_a(a|z, s)$ and $\widehat{Q}^\pi$ are correctly specified,

$$
\mathbb{E}_{s\sim\widehat{p}_0(s)}\left[\widehat{c}(z|s)\cdot\widehat{p}_a(a|z, s)\cdot\widehat{Q}^\pi(s, z, a)\right] - \mathbb{E}_{s\sim\widehat{p}_0(s)}\left[c(z|s)\cdot p_a(a|z, s)\cdot Q^\pi(s, z, a)\right] = o_p(1)
$$

Also, notice that $\widehat{p}_0(s)$ was set as the estimated initial state distribution. As the sample size goes to infinity, according to WLLN,

$$
\mathbb{E}_{s\sim\widehat{p}_0(s)}\left[c(z|s)\cdot p_a(a|z, s)\cdot Q^\pi(s, z, a)\right] - \mathbb{E}_{s\sim p_0(s)}\left[c(z|s)\cdot p_a(a|z, s)\cdot Q^\pi(s, z, a)\right] = o_p(1).
$$

Therefore,

$$
\widehat{\eta}^\pi_{\text{DM}} \xrightarrow{P} \mathbb{E}_{s\sim p_0(s)}\left[c(z|s)\cdot p_a(a|z, s)\cdot Q^\pi(s, z, a)\right] = \mathbb{E}_{s\sim p_0(s)}\Big[V^\pi(S_t)\Big] = \eta^\pi.
$$

Now it remains to show that $\mathbb{E}\left[\widehat{\phi}(D_t)\right] = 0$. Since $Q^\pi(s,z,a), p_a(a|s,z)$ are correctly specified,

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\phi}(D_t)\right] &= (1-\gamma)^{-1} \cdot \mathbb{E}\Big[\widehat{\omega}^\pi(S_t) \cdot \widehat{\rho}(S_t,Z_t) \cdot \Big\{R_t + \gamma \cdot V^\pi(S_{t+1}) - \mathbb{E}\left[R_t + \gamma \cdot V^\pi(S_{t+1})\big|Z_t,S_t\right] \\
&\qquad - \big(A_t - \mathbb{E}\left[A_t\big|Z_t,S_t\right]\big) \cdot \Delta(S_t)\Big\}\Big] \\
&= (1-\gamma)^{-1} \cdot \mathbb{E}\Big[\widehat{\omega}^\pi(S_t) \cdot \widehat{\rho}(S_t,Z_t) \cdot \mathbb{E}\Big\{R_t + \gamma \cdot V^\pi(S_{t+1}) - \mathbb{E}\left[R_t + \gamma \cdot V^\pi(S_{t+1})\big|Z_t,S_t\right] \\
&\qquad - \big(A_t - \mathbb{E}\left[A_t\big|Z_t,S_t\right]\big) \cdot \Delta(S_t)\big|Z_t,S_t\Big\}\Big|S_t,Z_t\Big] \\
&= (1-\gamma)^{-1} \cdot \mathbb{E}\Big[\widehat{\omega}^\pi(S_t) \cdot \widehat{\rho}(S_t,Z_t) \cdot \Big\{\mathbb{E}\left[R_t + \gamma \cdot V^\pi(S_{t+1})\big|Z_t,S_t\right] \\
&\qquad -\mathbb{E}\left[R_t + \gamma \cdot V^\pi(S_{t+1})\big|Z_t,S_t\right] - \big(\mathbb{E}\left[A_t\big|Z_t,S_t\right] - \mathbb{E}\left[A_t\big|Z_t,S_t\right]\big) \cdot \Delta(S_t)\Big\}\Big] \\
&= (1-\gamma)^{-1} \cdot \mathbb{E}\Big[\widehat{\omega}^\pi(S_t) \cdot \widehat{\rho}(S_t,Z_t) \cdot \Big\{0 - 0 \cdot \Delta(S_t)\Big\}\Big] = 0.
\end{aligned}
$$

Scenario 2: Suppose the models in $\mathcal{M}_2$ and p_a are correctly specified.

In this case, $c(z_t|S_t)$ is also correctly specified as a function of p_a and π. Then

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\eta}^\pi_{\text{DM}}\right] &= \mathbb{E}_{s\sim p_0(s)}\left[\sum_{z,a} c(z|s) \cdot p_a(a|s,z) \cdot \widehat{Q}^\pi(s,z,a)\right] \\
&= \mathbb{E}_{S\sim p^\pi_\infty(s), Z\sim p_z}\left[\frac{c(Z|S)}{p_z(Z|S)} \cdot \frac{p^\pi_0(S)}{p^\pi_\infty(S)} \cdot \sum_a p_a(a|Z,S) \cdot \widehat{Q}^\pi(S,Z,a)\right] \\
&= \mathbb{E}_{S\sim p^\pi_\infty(s)}\left[\frac{p^\pi_0(S)}{p^\pi_\infty(S)} \cdot \widehat{V}^\pi(S)\right].
\end{aligned}
$$

The augmentation terms satisfies

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\phi}(D_t)\right] &= (1-\gamma)^{-1} \cdot \mathbb{E}\Bigg[\omega^\pi(S_t) \cdot \rho(S_t,Z_t) \cdot \Big\{R_t + \gamma \cdot \widehat{V}^\pi(S_{t+1}) - \sum_{a_t} p_a(a_t|Z_t,S_t) \cdot \widehat{Q}^\pi(S_t,Z_t,a_t) \\
&\qquad - \big(A_t - \mathbb{E}\left[A_t\big|Z_t,S_t\right]\big) \cdot \widehat{\Delta}(S_t)\Big\}\Bigg] \\
&= (1-\gamma)^{-1} \cdot \mathbb{E}\Bigg[\omega^\pi(S_t) \cdot \frac{c(Z_t|S_t)}{p_z(Z_t|S_t)} \cdot \Big\{R_t + \gamma \cdot \widehat{V}^\pi(S_{t+1}) - \sum_{a_t} p_a(a_t|Z_t,S_t) \cdot \widehat{Q}^\pi(S_t,Z_t,a_t) \\
&\qquad - \big(\mathbb{E}\left[A_t\big|Z_t,S_t\right] - \mathbb{E}\left[A_t\big|Z_t,S_t\right]\big) \cdot \widehat{\Delta}(S_t)\Big\}\Bigg] \\
&= (1-\gamma)^{-1} \cdot \mathbb{E}\Bigg[\omega^\pi(S_t) \cdot \frac{c(Z_t|S_t)}{p_z(Z_t|S_t)} \cdot \Big\{R_t + \gamma \cdot \widehat{V}^\pi(S_{t+1}) - \sum_{a_t} p_a(a_t|Z_t,S_t) \cdot \widehat{Q}^\pi(S_t,Z_t,a_t)\Big\}\Bigg] \\
&= (1-\gamma)^{-1} \cdot \mathbb{E}\left[\omega^\pi(S_t) \cdot \frac{c(Z_t|S_t)}{p_z(Z_t|S_t)} \cdot R_t\right] \\
&\quad + (1-\gamma)^{-1} \cdot \mathbb{E}\left[\omega^\pi(S_t) \cdot \frac{c(Z_t|S_t)}{p_z(Z_t|S_t)} \cdot \Big\{\gamma \cdot \widehat{V}^\pi(S_{t+1}) - \sum_{a_t} p_a(a_t|Z_t,S_t) \cdot \widehat{Q}^\pi(S_t,Z_t,a_t)\Big\}\right].
\end{aligned}
$$

Now it remains to show that

$$
\mathbb{E}\left[\widehat{\eta}^\pi_{\text{DM}}\right] + (1-\gamma)^{-1} \cdot \mathbb{E}\left[\omega^\pi(S_t) \cdot \frac{c(Z_t|S_t)}{p_z(Z_t|S_t)} \cdot \Big\{\gamma \cdot \widehat{V}^\pi(S_{t+1}) - \sum_{a_t} p_a(a_t|Z_t,S_t) \cdot \widehat{Q}^\pi(S_t,Z_t,a_t)\Big\}\right] = 0. \tag{26}
$$

Since

$$
\begin{aligned}
&(1-\gamma)^{-1}\mathbb{E}\left[\omega^\pi(S_t)\frac{c(Z_t|S_t)}{p_z(Z_t|S_t)}\Big\{\gamma\widehat{V}^\pi(S_{t+1})-\sum_{a_t}p_a(a_t|Z_t,S_t)\widehat{Q}^\pi(S_t,Z_t,a_t)\Big\}\right]\\
=&(1-\gamma)^{-1}\mathbb{E}\left[\omega^\pi(S_t)\frac{c(Z_t|S_t)}{p_z(Z_t|S_t)}\Big\{\gamma\widehat{V}^\pi(S_{t+1})-\sum_{a_t}p_a(a_t|Z_t,S_t)\widehat{Q}^\pi(S_t,Z_t,a_t)\Big\}\right]\\
=&(1-\gamma)^{-1}\mathbb{E}_{S\sim p_\infty}\left[\omega^\pi(S)\frac{c(Z|S)}{p_z(Z|S)}\gamma\widehat{V}^\pi(S^*)\right]-(1-\gamma)^{-1}\mathbb{E}_{S\sim p_\infty}\left[\omega^\pi(S)\frac{c(Z|S)}{p_z(Z|S)}\sum_a p_a(a|Z,S)\widehat{Q}^\pi(S,Z,a)\right]\\
=&(1-\gamma)^{-1}\cdot\mathbb{E}_{S\sim p_\infty}\left[\omega^\pi(S)\frac{c(Z|S)}{p_z(Z|S)}\gamma\widehat{V}^\pi(S^*)\right]-(1-\gamma)^{-1}\mathbb{E}_{S\sim p_\infty}\left[\omega^\pi(S)\sum_{a,z}c(z|S)p_a(a|z,S)\widehat{Q}^\pi(S,z,a)\right]\\
=&(1-\gamma)^{-1}\cdot\mathbb{E}_{S\sim p_\infty}\left[\omega^\pi(S)\cdot\frac{c(Z|S)}{p_z(Z|S)}\cdot\gamma\widehat{V}^\pi(S^*)\right]-(1-\gamma)^{-1}\cdot\mathbb{E}_{S\sim p_\infty}\left[\omega^\pi(S)\cdot\widehat{V}^\pi(S)\right],
\end{aligned}
$$

by plugging in the definition for $\omega^\pi(s)$, we have

$$
\begin{aligned}
&(1-\gamma)^{-1}\cdot\mathbb{E}_{S\sim p_\infty}\left[\omega^\pi(S)\cdot\frac{c(Z|S)}{p_z(Z|S)}\cdot\gamma\cdot\widehat{V}^\pi(S^*)\right]-(1-\gamma)^{-1}\cdot\mathbb{E}_{S\sim p_\infty}\left[\omega^\pi(S)\cdot\widehat{V}^\pi(S)\right]\\
=&\mathbb{E}_{S\sim p_\infty}\left[\sum_{t=0}^\infty\gamma^t\frac{p_t^\pi(S)}{p_\infty(S)}\frac{c(Z|S)}{p_z(Z|S)}\gamma\widehat{V}^\pi(S^*)\right]-(1-\gamma)^{-1}\mathbb{E}_{S^*\sim p_\infty}\left[\omega^\pi(S^*)\widehat{V}^\pi(S^*)\right]\\
=&\mathbb{E}_{S\sim p_\infty}\left[\sum_{t=0}^\infty\gamma^{t+1}\frac{p_t^\pi(S)}{p_\infty(S)}\frac{c(Z|S)}{p_z(Z|S)}\widehat{V}^\pi(S^*)\right]-\mathbb{E}_{S^*\sim p_\infty}\left[\sum_{t=0}^\infty\gamma^t\frac{p_t^\pi(S^*)}{p_\infty(S^*)}\widehat{V}^\pi(S^*)\right]\\
=&\mathbb{E}_{S\sim p_\infty}\left[\sum_{t=0}^\infty\gamma^{t+1}\frac{p_t^\pi(S)}{p_\infty(S)}\frac{c(Z|S)}{p_z(Z|S)}\widehat{V}^\pi(S^*)\right]-\mathbb{E}_{S^*\sim p_\infty}\left[\sum_{t=1}^\infty\gamma^t\frac{p_t^\pi(S^*)}{p_\infty(S^*)}\widehat{V}^\pi(S^*)\right]-\mathbb{E}_{S^*\sim p_\infty}\left[\frac{p_0^\pi(S^*)}{p_\infty(S^*)}\widehat{V}^\pi(S^*)\right]\\
=&\mathbb{E}_{S\sim p_\infty}\left[\sum_{t=0}^\infty\gamma^{t+1}\cdot\left\{\frac{p_t^\pi(S)}{p_\infty(S)}\cdot\frac{c(Z|S)}{p_z(Z|S)}-\frac{p_{t+1}^\pi(S^*)}{p_\infty(S^*)}\right\}\cdot\widehat{V}^\pi(S^*)\right]-\mathbb{E}_{S^*\sim p_\infty}\left[\frac{p_0^\pi(S^*)}{p_\infty(S^*)}\cdot\widehat{V}^\pi(S^*)\right].
\end{aligned}
$$

Since $\mathbb{E}\left[\widehat{\eta}^\pi_{\text{DM}}\right]=\mathbb{E}_{S\sim p^\pi_\infty(s),Z\sim p_z}\left[\frac{p_0^\pi(S)}{p^\pi_\infty(S)}\cdot\widehat{V}^\pi(S)\right]$can be cancelled with the second term of the last expression, all we need to prove is that

$$
\mathbb{E}_{S\sim p_\infty}\left[\left\{\frac{p_t^\pi(S)}{p_\infty(S)}\cdot\frac{c(Z|S)}{p_z(Z|S)}-\frac{p_{t+1}^\pi(S^*)}{p_\infty(S^*)}\right\}\cdot\widehat{V}^\pi(S^*)\right]=0,\quad\text{for any t}\in\mathbb{N}.
$$

This statement holds naturally because, for any t,

$$
\begin{aligned}
&\mathbb{E}_{S\sim p_\infty}\left[\frac{p_t^\pi(S)}{p_\infty(S)}\cdot\frac{c(Z|S)}{p_z(Z|S)}\cdot\widehat{V}^\pi(S^*)\right]=\mathbb{E}_{S\sim p_t^\pi(S)}\left[\frac{c(Z|S)}{p_z(Z|S)}\cdot\widehat{V}^\pi(S^*)\right]\\
=&\ \mathbb{E}_{S\sim p_t^\pi(S)}\left[\sum_z c(z|S)\cdot\widehat{V}^\pi(S^*)\right]=\mathbb{E}_{S\sim p_t^\pi(S)}\left[(c(z=1|S)+c(z=0|S))\cdot\widehat{V}^\pi(S^*)\right]\\
=&\ \mathbb{E}_{S\sim p_t^\pi(S)}\left[\widehat{V}^\pi(S^*)\right]=\mathbb{E}_{S^*\sim p_{t+1}^\pi(S^*)}\left[\widehat{V}^\pi(S^*)\right]=\mathbb{E}_{S\sim p_\infty}\left[\frac{p_{t+1}^\pi(S^*)}{p_\infty(S^*)}\cdot\widehat{V}^\pi(S^*)\right].
\end{aligned}
$$

Therefore, under $\mathcal{M}_2$, i.e. when p_a, p_z, ω^π are correctly specified, $\widehat{\eta}_{\text{DR}}$ is consistent to η^π.

Based on the results in both scenarios, the double robustness of our estimator $\widehat{\eta}_{\text{DR}}$ thus holds.

G.2. Asymptotic Normality

For the simplicity of notations, we drop the subscript and denote our final DR estimator as $\widehat{\eta}$. First, let's define the oracle estimator $\widehat{\eta}^*$ as

$$
\widehat{\eta}^*:=\eta^*_{\text{DM}}+(NT)^{-1}\sum_{i,t}\phi^*(D_{i,t}),
$$

where we use the star sign in superscript to indicate that all of the models in η_{DM} and ϕ are substituted by their ground truth. To be more clear, we define $\eta^*_{\text{DM}} := \eta^\pi_{\text{DM}}(p_a, Q^\pi)$, and $\phi^*(D_{i,t}) := \phi(D_{i,t}, p_a, Q^\pi, p_z, \omega^\pi)$. Similarly, to emphasize the dependence of $\widehat{\eta}^\pi_{\text{DM}}$ and $\widehat{\phi}$ on nuisance functions, we denote $\widehat{\eta}^\pi_{\text{DM}} := \eta^\pi_{\text{DM}}(\widehat{p}_a, \widehat{Q}^\pi)$, and $\widehat{\phi}(D_{i,t}) := \phi(D_{i,t}, \widehat{p}_a, \widehat{Q}^\pi, \widehat{p}_z, \widehat{\omega}^\pi)$.

The proof of Section G.2 is decomposed to two steps. In Step 1, we will show that our DR estimator, $\widehat{\eta}$, is asymptotically equivalent to the oracle estimator $\widehat{\eta}^*$ by proving $\|\widehat{\eta} - \widehat{\eta}^*\| = o_p(N^{-\frac{1}{2}})$. In Step 2, we will illustrate the asymptotic normality of our oracle estimator $\widehat{\eta}^*$, and prove that the asymptotic variance indeed reaches the semiparametric efficiency bound.

Step 1: show that $\|\widehat{\eta} - \widehat{\eta}^*\| = o_p(N^{-\frac{1}{2}})$.

According to the assumptions in Theorem 2, all of the nuisance functions, i.e. $\widehat{p}_z$, $\widehat{p}_a$, $\widehat{Q}^\pi$ and $\widehat{\omega}^\pi$, converge in L_2-norm to their oracle values with rates no smaller than $1/4$. That is, there exist a constant $\alpha \geq 1/4$, such that

$$\sqrt{\mathbb{E}_{s\sim p_\infty}\mathbb{E}\big|\widehat{p}_z(z|s) - p_z(z|s)\big|^2} = O_p(N^{-\alpha}), \quad \sqrt{\mathbb{E}_{s\sim p_\infty}\mathbb{E}\big|\widehat{p}_a(a|z,s) - p_a(a|z,s)\big|^2} = O_p(N^{-\alpha}),$$
$$\sqrt{\mathbb{E}_{s\sim p_\infty}\mathbb{E}\big|\widehat{Q}^\pi(s,z,a) - Q^\pi(s,z,a)\big|^2} = O_p(N^{-\alpha}), \quad \sqrt{\mathbb{E}_{s\sim p_\infty}\mathbb{E}\big|\widehat{\omega}^\pi(s) - \omega^\pi(s)\big|^2} = O_p(N^{-\alpha}).$$

Based on definition of oracle estimator, we can decompose the difference between $\widehat{\eta}$ and $\widehat{\eta}^*$ as

$$\begin{aligned}\widehat{\eta} - \widehat{\eta}^* &= \Big[\eta^\pi_{\text{DM}}(\widehat{p}_a, \widehat{Q}^\pi) - \eta^\pi_{\text{DM}}(p_a, Q^\pi)\Big] + \frac{1}{NT}\sum_{i,t}\Big[\phi(D_{i,t}, \widehat{p}_a, \widehat{Q}^\pi, \widehat{p}_z, \widehat{\omega}^\pi) - \phi(D_{i,t}, p_a, Q^\pi, p_z, \omega^\pi)\Big]\\ &:= \Delta^{(1)} + \Delta^{(2)} + \Delta^{(3)},\end{aligned}$$

where we define

$$\begin{aligned}\Delta^{(1)} =& \Big[\eta^\pi_{\text{DM}}(p_a, \widehat{Q}^\pi) - \eta^\pi_{\text{DM}}(p_a, Q^\pi)\Big] + \frac{1}{NT}\sum_{i,t}\Big[\phi(D_{i,t}, p_a, \widehat{Q}^\pi, p_z, \omega^\pi) - \phi(D_{i,t}, p_a, Q^\pi, p_z, \omega^\pi)\Big],\\ \Delta^{(2)} =& \Big[\eta^\pi_{\text{DM}}(\widehat{p}_a, \widehat{Q}^\pi) - \eta^\pi_{\text{DM}}(p_a, \widehat{Q}^\pi)\Big] + \frac{1}{NT}\sum_{i,t}\Big[\phi(D_{i,t}, \widehat{p}_a, \widehat{Q}^\pi, p_z, \omega^\pi) - \phi(D_{i,t}, p_a, \widehat{Q}^\pi, p_z, \omega^\pi)\Big],\\ \Delta^{(3)} =& \frac{1}{NT}\sum_{i,t}\Big[\phi(D_{i,t}, \widehat{p}_a, \widehat{Q}^\pi, \widehat{p}_z, \widehat{\omega}^\pi) - \phi(D_{i,t}, \widehat{p}_a, \widehat{Q}^\pi, p_z, \omega^\pi)\Big].\end{aligned} \tag{27}$$

To finish the proof of this step, it suffice to show that the three terms in (27) are all $o_p(N^{-1/2})$ for any $\widehat{p}_z$, $\widehat{p}_a$, $\widehat{Q}^\pi$ and $\widehat{\omega}^\pi$ in a neighborhood of their true values.

For the brevity of content, we will only give a sketch of the proof and detail on just one term for illustration purpose.

Let's first consider $\Delta^{(1)}$. According to the double robustness property that will be proved in Theorem 5.2, for any $\widehat{Q}^\pi$ that may be misspecified in $\Delta^{(1)}$, we have $\mathbb{E}[\Delta^{(1)}] = 0$. Define $\mathcal{Q}^\pi$ as a neighborhood of Q^π. We construct an empirical operator $\mathbb{G}_N$ as

$$\mathbb{G}_N(\widehat{Q}^\pi) = \sqrt{N}\Big\{\Delta^{(1)}(p_a, \widehat{Q}^\pi, p_z, \omega^\pi) - \mathbb{E}\big[\Delta^{(1)}(p_a, \widehat{Q}^\pi, p_z, \omega^\pi)\big]\Big\},$$

Define a function class $\mathcal{F}_Q = \big\{\eta_{\text{DR}}(p_a, \widehat{Q}^\pi, p_z, \omega^\pi) - \eta_{\text{DR}}(p_a, Q^\pi, p_z, \omega^\pi) : \widehat{Q}^\pi \in \mathcal{Q}^\pi\big\}$ with $\eta_{\text{DR}}(p_a, Q^\pi, p_z, \omega^\pi) = \sum_{z_0,a_0} c(z_0|S_0)p_a(a_0|z_0, S_0)Q^\pi(S_0, z_0, a_0) + \frac{1}{T}\sum_{t=1}^T \phi(D_t, p_a, Q^\pi, p_z, \omega^\pi)$. It can be proved that $\mathcal{F}_Q$ is a VC-type class. By applying the Maximal Inequality specialized to VC type classes (Chernozhukov et al., 2014), one can show that $\sup_{\widehat{Q}^\pi\in\mathcal{Q}^\pi}\|\mathbb{G}_N(\widehat{Q}^\pi)\| = o_p(1)$, which further indicates that

$$\sup_{\widehat{Q}^\pi\in\mathcal{Q}^\pi}\Big\|\Delta^{(1)}(p_a, \widehat{Q}^\pi, p_z, \omega^\pi)\Big\| = \sup_{\widehat{Q}^\pi\in\mathcal{Q}^\pi}\Big\|\Delta^{(1)}(p_a, \widehat{Q}^\pi, p_z, \omega^\pi) - \mathbb{E}\big[\Delta^{(1)}(p_a, \widehat{Q}^\pi, p_z, \omega^\pi)\big]\Big\| = o_p(N^{-1/2}).$$

Therefore, as long as $\widehat{Q}^\pi$ converges in L_2 norm to its oracle value with rate $\alpha > 1/4$, we have $\Delta^{(1)} = o_p(N^{-1/2})$. [See (Van Der Vaart et al., 1996) for details.]

Next, let's consider $\Delta^{(2)}$. This term can be further decomposed as

$$\begin{aligned}
\Delta^{(2)} &= \Big[\eta_{\text{DM}}^{\pi}(\widehat{p}_a, \widehat{Q}^{\pi}) - \eta_{\text{DM}}^{\pi}(p_a, \widehat{Q}^{\pi}) - \eta_{\text{DM}}^{\pi}(\widehat{p}_a, Q^{\pi}) + \widehat{\eta}_{\text{DM}}^{\pi}(p_a, Q^{\pi})\Big] \\
&+ \frac{1}{NT}\sum_{i,t}\Big[\phi(D_{i,t}, \widehat{p}_a, \widehat{Q}^{\pi}, p_z, \omega^{\pi}) - \phi(D_{i,t}, p_a, \widehat{Q}^{\pi}, p_z, \omega^{\pi}) - \phi(D_{i,t}, \widehat{p}_a, Q^{\pi}, p_z, \omega^{\pi}) + \phi(D_{i,t}, p_a, Q^{\pi}, p_z, \omega^{\pi})\Big] \\
&+ \Big[\eta_{\text{DM}}^{\pi}(\widehat{p}_a, Q^{\pi}) - \widehat{\eta}_{\text{DM}}^{\pi}(p_a, Q^{\pi})\Big] + \frac{1}{NT}\sum_{i,t}\Big[\phi(D_{i,t}, \widehat{p}_a, Q^{\pi}, p_z, \omega^{\pi}) - \widehat{\phi}(D_{i,t}, p_a, Q^{\pi}, p_z, \omega^{\pi})\Big].
\end{aligned}$$

The last line of the equation above, akin to $\Delta^{(1)}$, can be proved to be $o_p(N^{-1/2})$ under the L_2 convergence assumption of $\widehat{Q}^{\pi}$. The first two lines, after some manipulations, can be bounded by the product of two nuisance function error terms.

To be more specific, let's consider the first line and prove that it is bound with order $o_p(N^{-1/2})$.

$$\begin{aligned}
&\Big|\eta_{\text{DM}}^{\pi}(\widehat{p}_a, \widehat{Q}^{\pi}) - \eta_{\text{DM}}^{\pi}(p_a, \widehat{Q}^{\pi}) - \eta_{\text{DM}}^{\pi}(\widehat{p}_a, Q^{\pi}) + \widehat{\eta}_{\text{DM}}^{\pi}(p_a, Q^{\pi})\Big| \\
&= \Big|\frac{1}{N}\sum_{i=1}^{N}\sum_{z,a}\Big\{\widehat{c}(z|S_{i,0})\widehat{p}_a(a|z,S_{i,0})\widehat{Q}^{\pi}(S_{i,0},z,a) - c(z|S_{i,0})p_a(a|z,S_{i,0})\widehat{Q}^{\pi}(S_{i,0},z,a)\Big\} \\
&\quad - \frac{1}{N}\sum_{i=1}^{N}\sum_{z,a}\Big\{\widehat{c}(z|S_{i,0})\widehat{p}_a(a|z,S_{i,0})Q^{\pi}(S_{i,0},z,a) - c(z|S_{i,0})p_a(a|z,S_{i,0})Q^{\pi}(S_{i,0},z,a)\Big\}\Big|
\end{aligned}$$

$$\begin{aligned}
&= \Big|\frac{1}{N}\sum_{i=1}^{N}\sum_{z,a}\Big\{\Big[\widehat{c}(z|S_{i,0})\widehat{p}_a(a|z,S_{i,0}) - c(z|S_{i,0})p_a(a|z,S_{i,0})\Big]\cdot\Big[\widehat{Q}^{\pi}(S_{i,0},z,a) - Q^{\pi}(S_{i,0},z,a)\Big]\Big\}\Big| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\sum_{z,a}\Big|\widehat{c}(z|S_{i,0})\widehat{p}_a(a|z,S_{i,0}) - c(z|S_{i,0})p_a(a|z,S_{i,0})\Big|\cdot\Big|\widehat{Q}^{\pi}(S_{i,0},z,a) - Q^{\pi}(S_{i,0},z,a)\Big| \\
&= o_p(N^{-1/4})\cdot o_p(N^{-1/4}) = o_p(N^{-1/2}),
\end{aligned}$$

where the second last equality is derived from the L_2 convergence of $\widehat{p}_a$ and $\widehat{Q}^{\pi}$ with rate $\alpha \geq 1/4$.

The proof of $\Delta^{(3)}$ is also similar to other terms, which is thus omitted. A similar proof can be found in (Shi et al., 2022c). Therefore, $\|\widehat{\eta} - \widehat{\eta}^*\| = o_p(N^{-\frac{1}{2}})$.

Step 2: show that $\sqrt{N}(\widehat{\eta}^* - \eta^{\pi}) \xrightarrow{d} \mathcal{N}(0, \sigma_T^2)$, where σ_T^2 is the efficiency bound.
Since $\widehat{\eta}^*$ is the average of N i.i.d. data trajectories, by standard CLT, the oracle estimator $\widehat{\eta}^*$ is asymptotically normal with variance

$$\begin{aligned}
\sigma_T^2 &= \text{Var}\Big\{\sum_{z,a} c(z|S_0)\cdot p_a(a|z,S_0)\cdot Q^{\pi}(S_0,z,a) + \frac{1}{T}\sum_{t=1}^{T}\phi(D_t)\Big\} \\
&= \text{Var}\Big\{V^{\pi}(S_0) + \frac{1}{T}\sum_{t=1}^{T}\phi(D_t)\Big\} = \text{Var}\Big\{V^{\pi}(S_0)\Big\} + \frac{1}{T^2}\sum_{t=1}^{T}\text{Var}\Big\{\phi(D_t)\Big\},
\end{aligned}$$

where the second equality is due to (4) and the last equality is due to the Markov and conditional mean independence assumptions. It follows directly from the proof of Theorem 2 that $\sigma_T^2 = CR(\mathcal{M})$.

Therefore, by combining the claims in Step 1 and Step 2, we have

$$\sqrt{N}(\widehat{\eta}_{\text{DR}} - \eta^{\pi}) \xrightarrow{d} \mathcal{N}(0, \sigma_T^2).$$

The proof of Theorem 3 is thus complete.

H. Proof of Theorem 6.1

In this section, we detail the identification proof under confounded POMDP. For any $g_Q \in \mathcal{G}_Q$,

$$\begin{aligned} 0 &= \mathbb{E}\Big\{R + \gamma \sum_{z,a} c(z|O') \cdot p_a(a|z,O') \cdot g_Q(F',z,a) - g_Q(F,z,a)\Big| H, Z, A\Big\} \\ &= \mathbb{E}\Big\{\mathbb{E}\big[R + \gamma \sum_{z,a} c(z|O') \cdot p_a(a|z,O') \cdot g_Q(F',z,a) - g_Q(F,z,a)\big| S, H, Z, A\big]\Big| H, Z, A\Big\}, \end{aligned}$$

where the first equality holds by the definition of learnable future-dependent Q function, and the second equality holds by the law of total expectation.

Since $H \perp\!\!\!\perp F'|(S,Z,A)$, we have

$$0 = \mathbb{E}\Big\{\mathbb{E}\big[R + \gamma \sum_{z,a} c(z|O') \cdot p_a(a|z,O') \cdot g_Q(F',z,a) - g_Q(F,z,a)\big| S, Z, A\big]\Big| H, Z, A\Big\}.$$

Then by the invertibility assumption, it holds almost surely that

$$\mathbb{E}\Big[R + \gamma \sum_{z,a} c(z|O') \cdot p_a(a|z,O') \cdot g_Q(F',z,a) - g_Q(F,z,a)\big| S, Z, A\Big] = 0.$$

According to the derivation details of importance sampling estimator in Formula (24), the aggregated value η^π can be written as

$$\eta^\pi = (1-\gamma)^{-1}\mathbb{E}_{S\sim p_\infty}\Big[\omega^\pi(S) \cdot \mathbb{E}\big\{\rho(Z,O) \cdot R|S\big\}\Big].$$

In Formula (26), since $\widehat{Q}$ is misspecified, one can substitute $\widehat{Q}$ to g_Q and the equality still holds. That is,

$$\begin{aligned} 0 = \mathbb{E}_{f\sim\nu_F}\Big[\sum_{z,a} c(z|o)p_a(a|z,o)g_Q(f,z,a)\Big] + (1-\gamma)^{-1}\mathbb{E}\Big[\omega^\pi(S) \cdot \rho(Z,O) \cdot \\ \Big\{\gamma \cdot \sum_{z,a} c(z|O')p_a(a|z,O')g_Q(F',z,a) - \sum_a p_a(a|Z,O) \cdot g_Q(F,Z,a)\Big\}\Big]. \end{aligned} \tag{28}$$

Therefore, by plugging in the expression in (28), we have

$$\begin{aligned} &\eta^\pi - \mathbb{E}_{f\sim\nu_F}\Big[\sum_{z,a} c(z|o)p_a(a|z,o)g_Q(f,z,a)\Big] \\ =&(1-\gamma)^{-1}\mathbb{E}_{s\sim p_\infty}\Big[\omega^\pi(s) \cdot \mathbb{E}\big\{\rho(Z,O) \cdot R|S=s\big\}\Big] - \mathbb{E}_{f\sim\nu_F}\Big[\sum_{z,a} c(z|o)p_a(a|z,o)g_Q(f,z,a)\Big] \\ =&(1-\gamma)^{-1}\mathbb{E}\Big[\omega^\pi(S)\rho(Z,O)\Big\{R + \gamma\sum_{z,a} c(z|O')p_a(a|z,O')g_Q(F',z,a) - \sum_a p_a(a|Z,O) \cdot g_Q(F,Z,a)\Big\}\Big] \\ =&(1-\gamma)^{-1}\mathbb{E}\Big[\omega^\pi(S)\rho(Z,O)\Big\{R + \gamma\sum_{z,a} c(z|O')p_a(a|z,O')g_Q(F',z,a) - g_Q(F,Z,A)\Big\}\Big] = 0, \end{aligned}$$

where the last equality holds by the overlap condition and the definition of learnable future-dependent Q function. Therefore, η^π can be written as a function of the observed data, which is indeed identifiable. The identification expression for η^π is given by

$$\eta^\pi = \mathbb{E}_{f\sim\nu_F}\Big[\sum_{z,a} c(z|o)p_a(a|z,o)g_Q(f,z,a)\Big].$$

The proof of this theorem is thus complete.