A Reinforcement Learning Framework for Dynamic Mediation Analysis

Lin Ge [1] **Jitao Wang** [2] **Chengchun Shi** [3] **Zhenke Wu** [2] **Rui Song** [1]

Abstract

Mediation analysis learns the causal effect transmitted via mediator variables between treatments and outcomes, and receives increasing attention in various scientific domains to elucidate causal relations. Most existing works focus on point-exposure studies where each subject only receives one treatment at a single time point. However, there are a number of applications (e.g., mobile health) where the treatments are sequentially assigned over time and the dynamic mediation effects are of primary interest. Proposing a reinforcement learning (RL) framework, we are the first to evaluate dynamic mediation effects in settings with infinite horizons. We decompose the average treatment effect into an immediate direct effect, an immediate mediation effect, a delayed direct effect, and a delayed mediation effect. Upon the identification of each effect component, we further develop robust and semi-parametrically efficient estimators under the RL framework to infer these causal effects. The superior performance of the proposed method is demonstrated through extensive numerical studies, theoretical results, and an analysis of a mobile health dataset. A Python implementation of the proposed procedure is available at https://github.com/linlinlin97/MediationRL.

1. Introduction

Mediation analysis aims to understand the causal pathway from an exposure (e.g., treatment or action) to an outcome variable of interest. It is gaining increasing popularity recently and has been frequently employed in a number of domains including epidemiology (Richiardi et al., 2013; Rijnhart et al., 2021), psychology (Rucker et al., 2011), genetics (Chakrabortty et al., 2018; Zeng et al., 2021; Djordjilović et al., 2022), economics (Celli, 2022) and neuroscience (Li et al., 2022; Shi & Li, 2022).

Our paper is motivated by the need to learn the dynamic mediation effects in sequential decision making. One motivating example is given by the Intern Health Study (IHS, NeCamp et al., 2020), which focuses on sequential mobile health interventions to help improve the mental health of medical interns who work in stressful environments. Participants were randomly assigned to receive notifications (e.g., tips and insights) throughout the study. For example, some notifications remind participants to take a break or enjoy a tasty treat, while others summarize the trends of recent physical activity and sleep. All the notifications are designed to improve participants' mood scores (self-reported via a custom-made study App) either directly or indirectly through increased activity or sleep hours. In addition, it is essential to note that participants' recent behavior will not only influence their proximal mood but will also influence their behavior and mood scores in the following days. To design a more effective intervention policy in IHS, it is necessary to understand how mobile prompts impact mood scores. In particular, the mobile prompts may directly impact the mood scores or encourage more physical activity and sleep, which may then impact the mood scores. In addition, an individual's past treatment sequence and behavior trajectory may impact the mood score. Teasing out these distinct sources of causal impacts on mood scores and their relative magnitudes needs new definitions, identification results, and inferential methods.

A fundamental question considered in this paper is how to infer the dynamic mediation effects in the aforementioned applications. Solving this question raises at least three challenges. First, the mediator at a given time affects both the current and future outcomes, inducing temporal carryover effects. As demonstrated in the case study in Section 8, the delayed direct effect (DDE) and the delayed mediator effect (DME) are significant and dominate the average treatment effect for the intervention policy used in the IHS (Sen et al., 2010; NeCamp et al., 2020). In contrast, the immediate direct effect (IDE) and immediate mediator effect (IME) are both insignificant. Nonetheless, most existing mediation analyses focus on estimating the indirect effect on the immediate reward and are hence inappropriate to our application. Second, the horizon (e.g., number of decision

[1]North Carolina State University [2]University of Michigan, Ann Arbor [3]London School of Economics and Political Science. Correspondence to: Rui Song <rsong@ncsu.edu>.

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

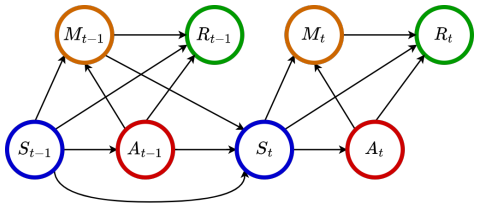

Figure 1. Mediated MDP.

stages) in the aforementioned applications is typically very long or diverges with the sample size. Existing solutions developed in finite horizon settings typically suffer from the curse of horizon in the sense that the variances of the proposed estimators grow exponentially fast with respect to the horizon (Liu et al., 2018) and are hence inapplicable; see Section 2 for details. Third, regardless of how the dynamic effects may change during the sequential treatments (or lack thereof), most works focus on examining the causal effects on the final outcome obtained at the end of the treatment process. However, in the context of behavioral change, the goal is to encourage and maintain small improvements to nudge individuals into generating sustained improvements in outcomes like mood scores. Currently, there is a dearth of methods to analyze causal effects for outcomes measured at every decision point in the sequence.

To address these limitations, we propose formulating the evaluation of dynamic mediation effects as a reinforcement learning (RL) problem. In particular, we use the Markov decision process (MDP) that is commonly employed in RL to model the mediated dynamic decision process over an infinite time horizon. Building upon the standard MDP, we introduce four additional sets of causal relationships, including state-mediator, action-mediator, mediator-state, and mediator-reward, as shown in Figure 1. To evaluate the effects of different treatment policies, we consider using the off-policy evaluation (OPE, Dudík et al., 2014; Uehara et al., 2022), which is widely used to avoid the difficulty of rerunning trials by evaluating treatment policies based on observational data.

Contributions. The main contributions are as follows. Motivated by the mobile health applications, we first construct the mediation analysis within the framework of RL over an infinite time horizon. Second, we propose to decompose the average treatment effect between a target policy and a control policy into IDE, IME, DDE, and DME. While IDE and IME have been extensively studied in single-stage settings, we introduce the DDE and DME to quantify the carryover effects of past actions and mediators. Third, upon the identification result of each effect component, multiply-robust estimators are developed. In particular, each proposed estimator is consistent even when models such as mediator distribution and reward distribution are misspecified (See Section 7.1). Furthermore, we theoretically show the semiparametric efficiency of the proposed estimators and confirm the theoretical prediction using numerical studies. Lastly, we conclude by analyzing the IHS data and providing new insights into guiding future designs of these behavioral interventions.

2. Related Work

Mediation analysis is widely studied in point-exposure studies under the classical structure consisting of a treatment, a mediator, and an outcome (Robins & Greenland, 1992; Pearl, 2022; Petersen et al., 2006; van der Laan & Petersen, 2008; Imai et al., 2010; Tchetgen & Shpitser, 2012; Tchetgen Tchetgen & Shpitser, 2014; VanderWeele, 2015), decomposing the average treatment effect into direct effect and indirect effect. Recently, to address commonly observed intermediate confounders that would be affected by the exposure and then affect both mediator and outcome, multiple methods have been developed to extend the classical mediation analysis (Robins & Richardson, 2010; Tchetgen & VanderWeele, 2014; VanderWeele et al., 2014; Vansteelandt & Daniel, 2017; Díaz et al., 2021; Díaz, 2022), among which the random intervention (RI)-based approach (VanderWeele et al., 2014; Díaz, 2022) further sets the foundation for the recent advancement of longitudinal mediation analysis.

There is a rich literature on longitudinal mediation analysis with no intermediate confounders (Selig & Preacher, 2009; Roth & MacKinnon, 2013). See also Preacher (2015) for a detailed review. However, time-varying intermediate confounders are ubiquitous in longitudinal data contexts. For example, in the IHS, doing exercises may result in a good mood, which may, in turn, increase the likelihood of engaging in more activities the next day and then subsequently affect the mood that follows.

In the presence of time-varying intermediate confounders, there are two major RI-based approaches. VanderWeele & Tchetgen Tchetgen (2017) and Díaz et al. (2022) proposed to intervene in the mediator sequence by randomly drawing mediators from the corresponding *marginal* distribution and defined the longitudinal interventional indirect/direct effect, which is different from the natural effect decomposition. Our work is primarily related to the work of Zheng & van der Laan (2017), which proposed to intervene in the mediator by randomly drawing the mediator from its *conditional* distribution and provided a natural decomposition of the total effect. Using the efficient influence function (EIF), they developed a multiply-robust estimator with less reliance on the correct model specification. However, all the aforementioned methods only focused on the treatment impact on the final outcome in finite horizons and did not consider immediate outcomes or infinite horizon settings. In addition, the estimator developed by Zheng & van der Laan (2017) is based on the product of importance sampling ratios at all time points and suffers from the curse of horizon.

Zheng & van der Laan (2012) also analyzed the longitudinal mediation effect by drawing mediators from *conditional* distribution but with a focus on single-exposure settings.

Using an RL framework for dynamic mediation analysis over an infinite horizon, our work is also connected to the line of research on OPE. Existing OPE-related research evaluates the discounted sum of rewards or average rewards for a target policy using observational data gained by following a different behavior policy. In general, there are three types of estimation procedures. The first is known as the direct method (DM, Le et al., 2019; Feng et al., 2020; Luckett et al., 2020; Hao et al., 2021; Liao et al., 2021; Chen & Qi, 2022; Shi et al., 2022a), which directly learns Q-functions and obtains value estimates based on their estimators. The second category of approaches utilizes importance sampling (IS, Precup, 2000; Thomas et al., 2015; Hallak & Mannor, 2017; Hanna et al., 2017; Liu et al., 2018; Xie et al., 2019; Dai et al., 2020; Zhang et al., 2020), which re-weights the rewards to eliminate the bias due to distributional shift. The third category develops doubly robust (DR) estimators by appropriately integrating DM with IS estimators (Jiang & Li, 2016; Thomas & Brunskill, 2016; Farajtabar et al., 2018; Liao et al., 2020; Tang et al., 2020; Uehara et al., 2020; Kallus & Uehara, 2022). DR estimators are also known to achieve the semiparametric efficiency bound (Bickel et al., 1993). However, none of the above papers studied mediation analysis. Recently, Shi et al. (2022b) proposed a consistent DR estimator for OPE in the presence of unmeasured confounders with the help of a mediator variable, which is used to intercept each directed path from treatments to reward/state. Our paper differs from theirs in that we decompose the off-policy value into the sum of IDE, IME, DDE, and DME and focus on settings without unmeasured confounding.

3. Preliminaries

3.1. Data Generating Process

We consider the observational data generated from a mediated Markov decision process (MMDP), as illustrated in Figure 1. Suppose there exists an agent that tries to learn from the data and interact with a given environment. At each time t, the environment arrives at a state $S_t \in \mathcal{S}$, and the agent selects an action $A_t \in \mathcal{A} = \{0, 1, \cdots, K-1\}$ according to a behavior policy $\pi_b(\bullet|S_t)$. Building upon the usual MDP, to further analyze the mediation effect, we consider an immediate mediator variable $M_t \in \mathcal{M}$ drawn according to $p_m(\bullet|S_t, A_t)$, which mediates the effect of A_t on the environment. Subsequently, the agents would receive an immediate R_t and the the environment transits to a next-state S_{t+1} according to $p_{s',r}(\bullet, \bullet|S_t, A_t, M_t)$. Both $\mathcal{S}$ and $\mathcal{M}$ are finite dimensional vector spaces. To summarize, the observed data sequences consist of the state-action-mediator-reward tuples $(S_t, A_t, M_t, R_t)_{t\geq 0}$ satisfying the following Markov assumption: $(M_t, R_t, S_{t+1}) \perp\!\!\!\perp (S_j, A_j, M_j, R_j)_{j<t}|(S_t, A_t)$ for any t.

3.2. Problem Formulation

Let N denote the number of trajectories. The ith trajectory contains $\{(S_{i,t}, A_{i,t}, M_{i,t}, R_{i,t})\}_{1\leq i\leq N, 0\leq t\leq T}$ where T is the termination time. We assume that all these trajectories are i.i.d. and follow the MMDP. Let π denote a generic (stationary) policy which maps from $\mathcal{S}$ to a probability mass function on $\mathcal{A}$, and $\mathbb{E}^{\pi}[\cdot]$ denote the expectation of a random variable under the policy π. Based on the observed data, our goal is to analyze the average treatment effect (ATE) of a target policy π_e relative to a control policy π_0, given by

$$\text{ATE}(\pi_e, \pi_0) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \text{TE}_t(\pi_e, \pi_0),$$

where $\text{TE}_t(\pi_e, \pi_0) = \mathbb{E}^{\pi_e}[R_t] - \mathbb{E}^{\pi_0}[R_t]$.

To gain a better understanding of the mediated and delayed effects, we consider decomposing $\text{TE}_t(\pi_e, \pi_0)$ into

$$\text{IDE}_t(\pi_e, \pi_0) + \text{IME}_t(\pi_e, \pi_0) + \text{DDE}_t(\pi_e, \pi_0) + \text{DME}_t(\pi_e, \pi_0).$$

Averaging over t for each component, we obtain a four-way decomposition of ATE as IDE + IME + DDE + DME. We formally define each of these effects in the next section.

4. Effect Decomposition

This section begins with a decomposition of $\text{TE}_t(\pi_e, \pi_0)$, from which we define each component in $\text{ATE}(\pi_e, \pi_0)$. We first notice that TE_t can be decomposed into two components: i) the immediate treatment effect (ITE_t) measuring the impact of the current action-mediator pair (A_t, M_t) on the immediate outcome R_t; ii) the delayed treatment effect (DTE_t) that measures the carryover effects of the historical action-mediator sequences $(A_j, M_j)_{j<t}$ on R_t that pass through S_t.

We next consider ITE_t. Let $\pi_{e,0}^t$ denote a nonstationary policy that follows π_e at the first $t-1$ steps and then follows π_0 at t. Mathematically, ITE_t is defined as $\mathbb{E}^{\pi_e}[R_t] - \mathbb{E}^{\pi_{e,0}^t}[R_t]$. Notice that $\pi_{e,0}^t$ differs from the stationary policy π_e only at the current time t, then ITE_t indeed measures the immediate effect. Under the Markov assumption, we obtain

$$\mathbb{E}^{\pi_e}[R_t] = \sum_{s,a,m} p_t^{\pi_e}(s)\pi_e(a|s)p_m(m|s,a)r(s,a,m),$$
$$\mathbb{E}^{\pi_{e,0}^t}[R_t] = \sum_{s,a,m} p_t^{\pi_e}(s)\pi_0(a|s)p_m(m|s,a)r(s,a,m),$$

where $p_t^{\pi}(s)$ denotes the distribution of S_t under a policy π, and $r(\bullet, \bullet, \bullet)$ denotes the conditional expectation of the reward given the state-action-mediator triplet.

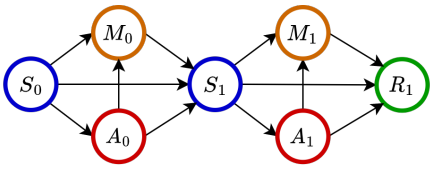

Figure 2. Causal paths from actions to reward received in $t = 1$.

Notice that A_t has both a direct effect and an indirect effect (mediated by M_t) on R_t. This motivates us to further decompose ITE_t into IDE_t and IME_t. Let G_e^t denote the process in which π_e is applied at the first $t-1$ steps to generate S_t, A_t is then generated according to π_0, and M_t is generated as if A_t were assigned according to π_e, i.e.,

$$M_t \sim \sum_a p_m(\bullet|a, S_t)\pi_e(a|S_t). \tag{1}$$

Then, $\mathbb{E}^{G_e^t}[R_t]$ equals

$$\sum_{s,a,m} p_t^{\pi_e}(s)\pi_e(a|s)p_m(m|s,a)\sum_{a'}\pi_0(a'|s)r(s,a',m).$$

It follows that

$$\text{ITE}_t = \underbrace{\mathbb{E}^{\pi_e}[R_t] - \mathbb{E}^{G_e^t}[R_t]}_{\text{IDE}_t(\pi_e,\pi_0)} + \underbrace{\mathbb{E}^{G_e^t}[R_t] - \mathbb{E}^{\pi_{e,0}^t}[R_t]}_{\text{IME}_t(\pi_e,\pi_0)}.$$

By definition, the IDE_t quantifies the direct treatment effect on the proximal outcome R_t whereas the IME_t evaluates the indirect effect mediated by M_t. As an illustration, set $t = 1$ and consider Figure 2. IDE_1 measures the causal effect along the path $A_1 \to R_1$ whereas IME_1 corresponds to the effect along the path $A_1 \to M_1 \to R_1$.

Next, we consider DTE_t, defined as $\mathbb{E}^{\pi_{e,0}^t}[R_t] - \mathbb{E}^{\pi_0}[R_t]$. By definition, $\pi_{e,0}^t$ differs from π_0 at the first $t-1$ time points. As such, DTE_t characterizes the delayed treatment effects on the current outcome R_t. Similarly, we further decompose DTE_t into the sum of direct and mediation effects. To characterize the delayed mediation effects, we follow the RI-based approach developed by Zheng & van der Laan (2017). Specifically, consider a stochastic process in which at the first $t-1$ time steps, the action is selected according to π_0, and the mediator is drawn assuming the action is assigned according to π_e (see Equation 1), whereas at time t, the system follows π_0. We provide more details on this process in Appendix A. Let G_0^t denote the resulting process and $\mathbb{E}^{G_0^t}[R_t]$ the expected value of R_t generated according to G_0^t. This allows us to decompose DTE_t as follows,

$$\text{DTE}_t = \underbrace{\mathbb{E}^{\pi_{e,0}^t}[R_t] - \mathbb{E}^{G_0^t}[R_t]}_{\text{DDE}_t(\pi_e,\pi_0)} + \underbrace{\mathbb{E}^{G_0^t}[R_t] - \mathbb{E}^{\pi_0}[R_t]}_{\text{DME}_t(\pi_e,\pi_0)}.$$

Notice that in the three processes, the action selection and mediator generation mechanisms at time t are the same. As such, both DDE_t and DME_t characterize the delayed effects. At the first $t-1$ time steps, the action selection mechanism between G_0^t and the process generated by $\pi_{e,0}^t$ are different whereas both processes have the same mediator generation mechanism. As such, DDE_t quantifies how past actions directly impact the current outcome. On the contrary, G_0^t and the process generated by π_0 have the same action selection mechanism. They differ in the way the mediator is generated. Hence, DME_t measures the indirect past treatment effects mediated by $\{M_j\}_{j<t}$. To elaborate, let us revisit Figure 2. DDE_1 captures the causal effect along the path $A_0 \to S_1 \to \{A_1, M_1\} \to R_1$ whereas DME_1 considers the path $A_0 \to M_0 \to S_1 \to \{A_1, M_1\} \to R_1$.

We also remark that the proposed effects are consistent with those in the existing literature. Specifically, when specialized to state-agnostic policies, IDE_0 and IME_0 are reduced to the total direct effect and the pure indirect effect (Robins & Greenland, 1992) in single-stage decision-making. Meanwhile, DDE_t and DME_t are similar to those proposed by Zheng & van der Laan (2017) developed in finite horizons.

Based on these effects, by aggregating IDE_t, IME_t, DDE_t and DME_t over time, we obtain the following four-way decomposition of $\text{ATE}(\pi_e, \pi_0)$,

$$\underbrace{\eta^{\pi_e} - \eta^{G_e}}_{\text{IDE}(\pi_e,\pi_0)} + \underbrace{\eta^{G_e} - \eta^{\pi_{e,0}}}_{\text{IME}(\pi_e,\pi_0)} + \underbrace{\eta^{\pi_{e,0}} - \eta^{G_0}}_{\text{DDE}(\pi_e,\pi_0)} + \underbrace{\eta^{G_0} - \eta^{\pi_0}}_{\text{DME}(\pi_e,\pi_0)},$$

where η^π is the average reward of policy π.

Finally, we remark that to simplify the presentation, we choose not to use the potential outcome framework (Rubin, 2005) to formulate these causal effects of interest in this section. The detailed potential outcome definitions are relegated to Appendix A. In addition, we show that these potential outcomes are identifiable and summarize the results in the following theorem.

Theorem 4.1 (Identification)**.** *Under standard assumptions including consistency, sequential randomization and positivity (Zheng & van der Laan, 2017; Luckett et al., 2019), IDE(π_e, π_0), IME(π_e, π_0), DDE(π_e, π_0), and DME(π_e, π_0) are all identifiable.*

We refer readers to Appendix C for more details.

5. Dynamic Treatment Effects Evaluation

In this section, we first develop DM and IS estimators for each defined dynamic treatment effect, whose consistencies require a given set of nuisance functions to be correctly specified. This motivates us to further develop doubly or triply robust estimators in section 5.3, whose consistencies only require either one of the two or three sets of nuisance functions to be correctly specified. Finally, we discuss the estimation methods for nuisance functions.

5.1. Direct Method Estimators (DM)

The direct estimators are built upon the Q-functions. For $\pi \in \{\pi_e, \pi_0\}$, we first define the conditional relative value function $Q^{\pi}(s, a, m)$ as

$$\sum_{t\geq 0} \mathbb{E}^{\pi}[R_t - \eta^{\pi} | S_0 = s, A_0 = a, M_0 = m], \quad (2)$$

which measures the expected total difference between rewards and the average reward of policy π, given that the initial state-action-mediator triplet equals (s, a, m). Notably, (2) deviates slightly from the standard definition in MDPs (i.e., $\sum_{t\geq 0} \mathbb{E}^{\pi_e}[R_t - \eta^{\pi_e} | S_0 = s, A_0 = a]$) by incorporating the mediator in the conditioning set.

Next, we define $Q^{G_e}(s, a, m)$ as

$$\sum_{t\geq 0} \mathbb{E}^{\pi_e}[r(S_t, \pi_0, M_t) - \eta^{G_e} | S_0 = s, A_0 = a, M_0 = m],$$

where $r(s, \pi_0, m)$ is a shorthand for $\sum_a \pi_0(a|s) r(s, a, m)$. Q^{G_e} aggregates the difference between the expected reward of the interventional process G_e^t starting from a given state-action-mediator triplet and that averaged over different initial conditions. It is crucial to note that Q^{G_e} differs from Q^{π} defined in (2), in that the observed reward R_t in (2) is replaced by the reward function $r(S_t, \pi_0, M_t)$. This is necessary as G_e^t uses different policies for action selection and mediator generation at t.

Following the same logic, we define $Q^{\pi_{e,0}}(s, a, m)$ as

$$\sum_{t\geq 0} \mathbb{E}^{\pi_e}[r(S_t, \pi_0) - \eta^{\pi_{e,0}} | S_0 = s, A_0 = a, M_0 = m],$$

where $r(s, \pi_0) = \sum_{a,m} \pi_0(a|s) p_m(m|a, s) r(s, a, m)$. We similarly define $Q^{G_0}(s, a, m)$ as

$$\sum_{t\geq 0} \mathbb{E}^{G_0^t}[r(S_t, \pi_0) - \eta^{G_0} | S_0 = s, A_0 = a, M_0 = m].$$

We remark that all Q-functions are finite under the assumption of aperiodicity even though the horizon is infinite (Puterman, 2014). This is because aperiodic Markov chains would reach their steady-state exponentially fast. As such, after a few iterations, the differences become very close to zero. More importantly, the ηs and Qs are closely related according to the well-known Bellman equation, which is fundamental to deriving the DM estimator. To elaborate, take the estimation of η^{π_e} as an example. According to the Bellman equation, we have that

$$\eta^{\pi_e} + Q^{\pi_e}(S_t, A_t, M_t) = \mathbb{E}^{\pi_e}[R_t + \sum_{a,m} \pi_e(a|S_{t+1}) \times p_m(m|a, S_{t+1}) Q^{\pi_e}(S_{t+1}, a, m) | S_t, A_t, M_t]. \quad (3)$$

Plugging in $\hat{p}_m$ learned from observed data into (3), we can construct estimation equations to learn Q^{π_e} and η^{π_e} jointly. See Section 5.4 for details. Let η_d^{π} denote the resulting DM estimator for η^{π}. The DM estimator of each effect component is then constructed by plugging in these η_ds, the consistency of which requires correct model specifications of r, the Q-function and p_m.

5.2. Importance Sampling (IS) Estimators

As commented earlier, standard IS estimators suffer from the curse of horizon. In this section, we utilize the marginal importance sampling (MIS) method proposed in Liu et al. (2018) to break the curse of horizon. For a given policy π, we first introduce the MIS ratio, given by

$$\omega^{\pi}(s) = p^{\pi}(s)/p^{\pi_b}(s),$$

where p^{π} and p^{π_b} denote the stationary state distribution under π and π_b, respectively. Using the change of measure theorem, it is immediate to see that, for $\pi \in \{\pi_e, \pi_0\}$,

$$\frac{1}{NT} \sum_{i,t} \omega^{\pi}(S_{i,t}) \frac{\pi(A_{i,t}|S_{i,t})}{\pi_b(A_{i,t}|S_{i,t})} R_{i,t} \quad (4)$$

is unbiased to η^{π}. Similarly, using the change of measure theorem again, it is straightforward to show that

$$\frac{1}{NT} \sum_{i,t} \omega^{\pi_e}(S_{i,t}) \frac{\pi_0(A_{i,t}|S_{i,t})}{\pi_b(A_{i,t}|S_{i,t})} R_{i,t}, \quad (5)$$

$$\frac{1}{NT} \sum_{i,t} \omega^{G_0}(S_{i,t}) \frac{\pi_0(A_{i,t}|S_{i,t})}{\pi_b(A_{i,t}|S_{i,t})} R_{i,t}, \quad (6)$$

are unbiased to $\eta^{\pi_{e,0}}$ and η^{G_0}, respectively. Here, ω^{G_0} is a version of ω^{π} with the numerator equal to the stationary state distribution when the data are generated according to $\{G_0^t\}_t$. These two MIS estimators (5 and 6) differ from (4) in that their state and action ratios are associated with two different interventional policies.

Lastly, we consider η^{G_e}. Recall that at time t, G_e^t selects action according to π_0 and generates the mediator as if π_e were applied to determine A_t. To further account for this distributional shift, we introduce a mediator ratio, $\rho(S, A, M) = p_m^{-1}(M|S, A)[\sum_a \pi_e(a|S) p_m(M|S, a)]$, built upon which the following unbiased estimator, denoted as MIS_1, can be derived,

$$\frac{1}{NT} \sum_{i,t} \omega^{\pi_e}(S_{i,t}) \frac{\pi_0(A_{i,t}|S_{i,t})}{\pi_b(A_{i,t}|S_{i,t})} \rho(S_{i,t}, A_{i,t}, M_{i,t}) R_{i,t}.$$

An alternative way to handle the distributional shift is to use the reward function instead of the observed reward to derive the IS estimator. This motivates the following estimator for η^{G_e},

$$(\text{MIS}_2): \frac{1}{NT} \sum_{i,t} \omega^{\pi_e}(S_{i,t}) \frac{\pi_e(A_{i,t}|S_{i,t})}{\pi_b(A_{i,t}|S_{i,t})} r(S_{i,t}, \pi_0, M_{i,t}),$$

which avoids the use of the mediator ratio.

So far, we have discussed the MIS estimators for those ηs. The subsequent MIS estimators for IDE(π_e, π_0), IME(π_e, π_0), DDE(π_e, π_0), and DME(π_e, π_0) can be similarly defined. Their consistencies require correct specifications of π_b, p_m, r, ω^{π_e}, ω^{π_0}, and ω^{G_0}.

5.3. Multiply Robust (MR) Estimators

This section develops the MR estimators that combine the DM and MIS estimators for efficient and robust OPE. These estimators are derived based on the classical semiparametric theory (see e.g., Tsiatis, 2006). See Appendix F for the detailed derivation. Let O denote a tuple (S, A, M, R, S'). For each η, the proposed MR estimator is built upon the estimating function $\eta_d + I_\eta(O)$, where η_d is the DM estimator of η and $I_\eta(O)$ denotes some augmentation term that involves the MIS ratio. The purpose of introducing these augmentation terms lies in debiasing the bias of the DM estimator, making the resulting estimator more robust against model misspecification. Given the estimating function, its empirical average over the data tuples produces the final MR estimator. We present the detailed forms of these estimating functions below.

First, consider η^{π_e} and η^{π_0}. For a given policy $\pi \in \{\pi_e, \pi_0\}$, $I_{\eta^\pi}(O)$ is given by

$$\omega^\pi(S)\frac{\pi(A|S)}{\pi_b(A|S)}\Big[R + Q^\pi(S', \pi) - Q^\pi(S, A) - \eta_d^\pi\Big],$$

where $Q^\pi(s, \pi) = \sum_{a,m} \pi(a|s)p_m(m|a, s)Q^\pi(s, a, m)$ and $Q^\pi(s, a) = \sum_m p_m(m|a, s)Q^\pi(s, a, m)$. Under the MMDP model, the term in brackets corresponds to a temporal difference error. Therefore, when Q^π, η_d^π, and p_m are correctly specified, it is of mean zero given (A, S). Thus, the resulting estimator is equivalent to DM which is consistent under correct model specification. On the contrary, when ω^π and π_b are correctly specified, the final estimator is equivalent to MIS, which is consistent under these configurations (Liao et al., 2020). As such, the resulting estimator is doubly robust whose consistency relying on the correct specification of (Q^π, η^π, p_m) or (ω^π, π_b).

Next, consider η^{G_e}. Let $I_{\eta^{G_e}}(O)$ denote

$$\omega^{\pi_e}(S)\Big[\frac{\pi_0(A|S)}{\pi_b(A|S)}\rho(S, A, M)\{R - r(S, A, M)\} + \frac{\pi_e(A|S)}{\pi_b(A|S)} \times\Big\{r(S, \pi_0, M) + Q^{G_e}(S', \pi_e) - Q^{G_e}(S, A) - \eta_d^{G_e}\Big\}\Big],$$

where ρ is the mediator ratio defined before. Similarly, the second line is the temporal difference error with a zero mean given (S, A, M) when models in $(r, p_m, Q^{G_e}, \eta_d^{G_e})$ are correctly specified. In addition, when r is correctly specified, conditional on (S, A, M), $\{R - r(S, A, M)\}$ is of zero mean as well. As such, $I_{\eta^{G_e}}(O)$ has a zero mean when $(r, p_m, Q^{G_e}, \eta^{G_e})$ are correctly specified. Further, one can show that the final estimator based on $\eta_d^{G_e} + I_{\eta^{G_e}}(O)$ is unbiased to MIS$_1$ or MIS$_2$ introduced in Section 5.2 when $(p_m, \omega^{G_e}, \pi_b)$ or (r, ω^{G_e}, π_b) are correctly specified. As such, the estimator is triply robust in the sense that its consistency requires $(r, p_m, Q^{G_e}, \eta^{G_e})$, $(p_m, \omega^{G_e}, \pi_b)$ or (r, ω^{G_e}, π_b) to be correct.

Next, we consider $\eta^{\pi_{e,0}}$ and introduce $I_{\eta^{\pi_{e,0}}}(O)$, defined as

$$\omega^{\pi_e}(S)\Big[\frac{\pi_0(A|S)}{\pi_b(A|S)}\{R - r(S, A)\} + \frac{\pi_e(A|S)}{\pi_b(A|S)} \times\Big\{r(S, \pi_0) + Q^{\pi_{e,0}}(S', \pi_e) - Q^{\pi_{e,0}}(S, A) - \eta_d^{\pi_{e,0}}\Big\}\Big],$$

where $r(s, a) = \sum_m p_m(m|a, s)r(s, a, m)$. Following the same logic, we can show that the resulting estimator is doubly robust and requires either models in $(r, p_m, Q^{G_e}, \eta_d^{G_e})$ or those in (ω^{π_e}, π_b) are correctly specified.

Finally, we consider η^{G_0} and introduce $I_{\eta^{G_0}}(O)$ as follows,

$$\omega^{G_0}(S)\frac{\pi_0(A|S)}{\pi_b(A|S)}\Big[\Big\{R - r(S, A)\Big\} + \rho(S, A, M)\Big\{r(S, \pi_0) + Q^{G_0}(S', G_0) - \eta_d^{G_0} - Q^{G_0}(S, A, M)\Big\}\Big] + \omega^{G_0}(S)\frac{\pi_e(A|S)}{\pi_b(A|S)} \times\Big[Q^{G_0}(S, \pi_0, M) - \sum_{a,m}\pi_0(a|S)p_m(m|A, S)Q^{G_0}(S, a, m)\Big],$$

where $Q^{G_0}(s, \pi_0, m)$ is a shorthand of $\sum_a \pi_0(a|s)Q^{G_0}(s, a, m)$ and $Q^{G_0}(s, G_0)$ equals $\sum_{a,a',m} \pi_0(a|s)\pi_e(a'|s)p_m(m|a', s)Q^{G_0}(s, a, m)$. The resulting estimator's doubly robustness property can be similarly established.

So far, we have introduced all the MR estimators for estimating these average rewards ηs. We can plug in these estimators to construct the corresponding MR estimators for those dynamic treatment effects (i.e., MR-IDE(π_e, π_0), MR-IME(π_e, π_0), MR-DDE(π_e, π_0), MR-DME(π_e, π_0)). Their consistencies and robustness can be similarly derived. We summarize and formally prove their robustness properties in Theorem 6.1.

5.4. Learning Nuisance Functions

Recall that the MR estimators require estimation of nuisance functions including π_b, r, p_m, ω, Q, and η. While π_b, r, and p_m can be estimated efficiently using state-of-the-art nonparametric methods (i.e., regression/classification tree (Breiman et al., 2017), random forest (Breiman, 2001), deep learning (Schmidt-Hieber, 2020)) with convergence rates faster than $N^{-\frac{1}{4}}$, we focus on the methods used to learn ω, Q, and η.

We first consider the estimation of ω^π for any stationary policy π. Following the arguments in Liu et al. (2018) and

Uehara et al. (2020), we can show that for any function f

$$\mathbb{E}\left[\omega^{\pi}(S)\{f(S)-\frac{\pi(A|S)}{\pi_b(A|S)}f(S')\}\right]=0,$$

where the expectation is taken over the observed stationary distribution of (S,A,S'). Therefore, estimating ω^{π} is equivalent to solving a mini-max problem such that

$$\min_{\omega^{\pi}\in\Omega}\max_{f\in\mathcal{F}}\mathbb{E}^2\left[\omega^{\pi}(S)\{f(S)-\frac{\pi(A|S)}{\pi_b(A|S)}f(S')\}\right] \quad (7)$$

for some function classes Ω and $\mathcal{F}$. In our implementation, we consider linear function classes Ω and $\mathcal{F}$, which yields closed-form expressions. Specifically, let $\omega^{\pi}(s)=\xi^T(s)\beta$ for some d_{ω}-dimensional $\beta\in\mathbb{R}^{d_{\omega}}$, where $\xi(s)$ is the feature vector generated by RBF sampler (Rahimi & Recht, 2007). Then (7) is equivalent to obtain β by solving the equation

$$\frac{1}{NT}\sum_{i,t}\left[\xi(S_i,t)-\frac{\pi(A_{i,t}|S_{i,t})}{\pi_b(A_{i,t}|S_{i,t})}\xi(S_{i,t+1})\right]\xi^T(S_{i,t})\beta=0.$$

Similarly, considering ω^G, we can show that

$$\mathbb{E}\left[\omega^G(S)\{f(S)-\rho(S,A,M)\frac{\pi_0(A|S)}{\pi_b(A|S)}f(S')\}\right]=0,$$

where the expectation is taken over the distribution of (S,A,M,S'). ω^G can then be estimated following the same steps.

We next consider the estimation of pairs of (Q,η). Taking (Q^{π_e},η^{π_e}) as an example, the estimation procedure is motivated by the Bellman equation model, such that:

$$Q^{\pi_e}(S_t,A_t,M_t)=\mathbb{E}^{\pi_e}[R_t+\mathbb{E}^{\pi_e}_{a,m}Q^{\pi_e}(S_{t+1},a,m)-\eta^{\pi}]. \quad (8)$$

Similar to the work of Shi et al. (2022a), we approximate the Q function using linear sieves. Specifically, we assume that

$$Q^{\pi_e}(s,a,m)\approx\Phi_L^T(s,m)\beta_a,\forall s\in\mathcal{S},a\in\mathcal{A},m\in\mathcal{M},$$

where $\Phi_L^T(s,m)$ is a L-dimensional feature vector derived using L sieve basis functions, such as splines (Huang, 1998). Let $\beta^*=(\beta_0^T,\cdots,\beta_{K-1}^T,\eta^{\pi})^T$. Let $\boldsymbol{U}(s,a,m)$ denotes

$$[\Phi_L^T(s,m)1(a=0),\cdots,\Phi_L^T(s,m)1(a=K-1),1]^T,$$

and $\boldsymbol{V}(s)$ denotes

$$[\mathbb{E}_{m|s,a=0}\Phi_L^T(s,m)\pi_e(0|s),\cdots,\mathbb{E}_{m|s,a=K-1}\Phi_L^T(s,m)\pi_e(K-1|s),0]^T,$$

where $\mathbb{E}_{m|s,a}\Phi_L^T(s,m)=\int_m\Phi_L^T(s,m)p(m|s,a)$ can be approximated by Monte Carlo sampling in practice. Then, the equation (8) can be rewritten as

$$\mathbb{E}U(S,A,M)[R+V(S')^T\beta^*-U(S,A,M)^T\beta^*]=0.$$

Let $U_{i,t}=U(S_{i,t},A_{i,t},M_{i,t})$ and $V_{i,t}=V(S_{i,t})$. Based on the observational data, the closed-form solution of β^* is

$$\left[\frac{1}{NT}\sum_{i,t}U_{i,t}(U_{i,t}-V_{i,t+1})^T\right]^{-1}\frac{1}{NT}\sum_{i,t}U_{i,t}R_{i,t}.$$

In practice, we add ridge penalty to the term within the bracket to prevent overfitting, and let L grow with the sample size to improve the approximation precision.

6. Statistical Guarantees

In this section, we prove the robustness and semi-parametric efficiency of the proposed MR estimator. We begin with some notations. Let $\mathcal{Q}^{(\cdot)},\Omega^{(\cdot)},\mathcal{H}_m,\mathcal{H}_r$, and Π_b respectively denote the function class of $Q^{(\cdot)},\omega^{(\cdot)},p_m,r$, and π_b.

Theorem 6.1. ***Multiply Robustness.*** *Suppose the conditions in Theorem 4.1 holds, the process $\{S_{i,t}\}_{t\geq 0}$ is stationary, π_b, $\hat{\pi}_b$, p_m and $\hat{p}_m$ are uniformly bounded away from 0, and $\mathcal{Q}^{(\cdot)},\Omega^{(\cdot)},\mathcal{H}_m,\mathcal{H}_r$, and Π_b are bounded VC-type classes (Chernozhukov et al., 2014) with VC indices upper bounded by $O(N^k)$ for some $k<1/2$. As $NT\to\infty$,*

1. *MR-IDE(π_e,π_0) is consistent if either the set of models in (ω^{π_e}, π_b, r) or in (ω^{π_e}, π_b, p_m) or in (Q^{π_e}, Q^{G_e}, $\eta_d^{\pi_e}$, $\eta_d^{G_e}$, r, p_m) are consistently estimated;*
2. *MR-IME(π_e,π_0) is consistent if either the set of models in (ω^{π_e}, π_b, r) or in (ω^{π_e}, π_b, p_m) or in (Q^{G_e}, $Q^{\pi_{e,0}}$, $\eta_d^{G_e}$, $\eta_d^{\pi_{e,0}}$, r, p_m) are consistently estimated;*
3. *MR-DDE(π_e,π_0) is consistent if either the set of models in (ω^{π_e}, ω^{G_0}, π_b, p_m) or in ($Q^{\pi_{e,0}}$, Q^{G_0}, $\eta_d^{\pi_{e,0}}$, $\eta_d^{G_0}$, r, p_m) are consistently estimated;*
4. *MR-DME(π_e,π_0) is consistent if either the set of models in (ω^{π_0}, ω^{G_0}, π_b, p_m) or in (Q^{G_0}, Q^{π_0}, $\eta_d^{G_0}$, $\eta_d^{\pi_0}$, r, p_m) are consistently estimated.*

Theorem 6.1 formally establish the triply robustness properties of MR-IDE(π_e,π_0) and MR-IME(π_e,π_0), as well as the doubly robustness properties of MR-DDE(π_e,π_0) and MR-DME(π_e,π_0), respectively. To save space, the proof of this theorem is deferred to the Appendix D.

Theorem 6.2. ***Efficiency.*** *Suppose the conditions in Theorem 6.1 holds, and $\hat{Q}^{(\cdot)},\hat{\omega}^{(\cdot)},\hat{p}_m,\hat{r},\hat{\pi}_b$, and $\hat{\eta}_d^{(\cdot)}$ converges to their oracle value in L_2 norm at a rate of N^{-k^*} for some $k^*>1/4$, respectively. The MR estimators are asymptotically normal with an asymptotic variance achieving the semiparametric efficiency bound.*

To save the space, the proof of this theorem is differed to the Appendix E with a sketch of the proof at the beginning. A Wald-type Confidence Interval (CI) for each MR estimator can be derived from Theorem 6.2.

7. Numerical Examples

In this section, we evaluate the estimation performance of the proposed methods through three simulation studies. Specifically, we demonstrate the robustness of the proposed MR estimator to model misspecification in the first simulation. In the second simulation, we compare the DM, MIS, and MR estimators to the classic direct/indirect estimator (Pearl, 2022) to demonstrate the importance of longitudinal mediation analysis, considering the policy effect on state transition. The final simulation is a semi-synthetic study that simulates the generation process of real data and demonstrates the superiority of the proposed MR estimators. For any effect X, let $\hat{X}$ be an estimator. We define the logbias as $\log|\mathbb{E}(\hat{X}-X)|$ and logMSE as $\mathbb{E}[\log(\hat{X}-X)^2]$.

7.1. Toy Example I

We consider a simplified MMDP setting with binary states, actions, mediators, and rewards. See Appendix G.1 for specific data generation settings. Let $\mathbb{M}_1 = (\omega^{\pi_e}, \pi_b, r)$, $\mathbb{M}_2 = (\omega^{\pi_e}, \omega^{\pi_0}, \omega^{G_0}, \pi_b, p_m)$, $\mathbb{M}_3 = (\{Q^\pi, \eta_d^\pi\}_{\pi\in\{\pi_e, G_e, \pi_{e,0}, G_0, \pi_0\}}, r, p_m)$. To investigate the robustness of the MR estimator, we test its performance in four scenarios: i) $\mathbb{M}_1$, $\mathbb{M}_2$, and $\mathbb{M}_3$ are all correctly specified; ii) only $\mathbb{M}_1$ is correctly specified; iii) only $\mathbb{M}_2$ is correctly specified; iv) only $\mathbb{M}_3$ is correctly specified; and v) all the models in $\mathbb{M}_1$, $\mathbb{M}_2$, and $\mathbb{M}_3$ are incorrectly specified by injecting non-negligible random noises. As shown in Figure 3, MR-IDE(π_e, π_0) and MR-IME(π_e, π_0) are consistent when either $\mathbb{M}_1$, $\mathbb{M}_2$, or $\mathbb{M}_3$ is correctly specified, and MR-DDE(π_e, π_0) and MR-DME(π_e, π_0) are consistent when either $\mathbb{M}_2$ or $\mathbb{M}_3$ is correctly specified.

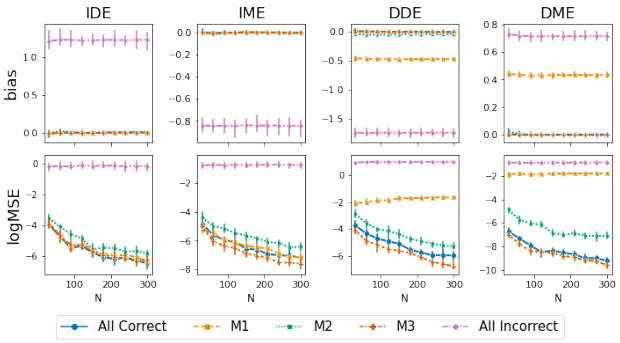

Figure 3. Bias and the logMSE of MR estimators, aggregated over 200 random seeds. The error bars represent the 95% CI.

7.2. Toy Example II

As discussed in Section 2, most existing works focus on a two-way decomposition of immediate treatment effects under the setting with a single stage. In this section, we compare the proposed estimators of IDE and IME to three baseline estimators assuming i.i.d. samples (See Appendix H for details). We first repeat the data generation process from Section 7.1, in which the states are affected by the history observations for each trajectory. Then, by modifying the distribution of the next state, S_{t+1}, as $\Pr(S_{t+1}=1)=.2$, we consider a second scenario in which all observations of states are i.i.d sampled. Note that there are two versions of MIS estimators for IDE and IME. Let MIS2 denote the MIS estimators using the MIS$_2$ to estimate η^{G_e}. According to Figure 4, when states are i.i.d. sampled, all estimators produce consistent estimates. However, when policy-induced state transitions occur, all baseline estimators yield biased estimates, whereas the proposed estimators continue to provide consistent estimates, implying the necessity of accounting for the policy effect on the state transition.

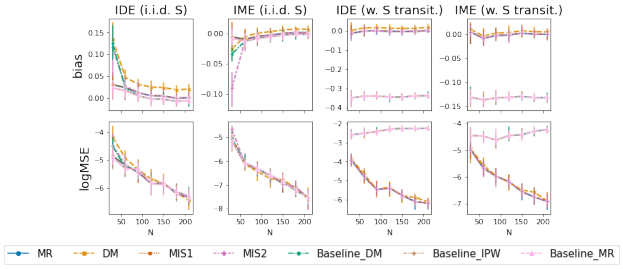

Figure 4. Bias and the logMSE of estimators, under different data generation scenarios. The results are aggregated over 200 random seeds. The error bars represent the 95% CI. Nuisance functions are estimated as discussed in Section 5.4.

7.3. Semi-Synthetic Data

In this section, we evaluate empirical performance of estimators using a semi-synthetic dataset structured similarly to the real dataset analyzed in Section 8. Specifically, we consider an MMDP setting with continuous reward, state, and mediator spaces and a binary action space. See Appendix G.2 for more information on the data-generation process. We compared the MR estimators to the DM estimators, the MIS estimators, and three baseline estimators. As shown in Figure 5 and Figure 6, the MR estimators outperform all other estimators for all components of ATE, especially when the sample size is large. We first focus on IDE(π_e, π_0) and IME(π_e, π_0). On the one hand, the baseline and MIS estimators are all biased, whereas the bias and MSE of the proposed DM and MR estimators decay continuously as N or T increases. On the other hand, the DM estimators yield relatively more significant bias and MSE than MR estimators. Considering the DDE(π_e, π_0) and DME(π_e, π_0), both the DM and MIS estimators are biased with non-decreasing MSE, whereas the MR estimators continue to provide estimates with low bias and low MSE that decrease with N and T. The results are in line with our theoretical findings. To further support the superior performance of the proposed MR estimators, additional simulation studies are conducted in Appendix J under different settings of data-generating

mechanisms, all of which reach the same conclusion as in this section.

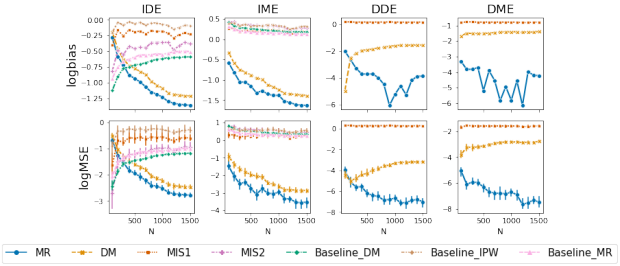

Figure 5. The logbias and logMSE of various estimators, aggregated over 100 random seeds. The error bars represent the 95% CI. Fix $T = 25$.

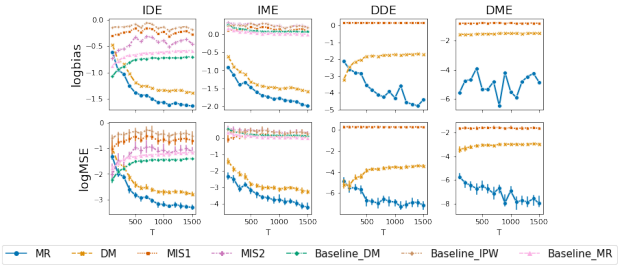

Figure 6. The logbias and logMSE of various estimators, aggregated over 100 random seeds. The error bars represent the 95% CI. Fix $N = 50$.

8. Real Data Application

In this section, we apply the proposed MR estimators to analyze the real dataset from the IHS (NeCamp et al., 2020), which was discussed as a motivating example in Section 1. The study involved 1565 interns and lasted six months. Every day, the participant would either receive a notification ($A_t = 1$) or no notification ($A_t = 0$). Meanwhile, participants' mood score (R_t), step count ($M_{t,1}$), and hours of sleep ($M_{t,2}$) were recorded. At each time step, we consider the previous time step's mood score as the current state (i.e., $S_t = R_{t-1}$).

Using the control policy π_0 of no intervention, we are interested in evaluating the treatment effects of the behavior policy π_b used throughout the study, which sends notifications to individuals randomly with a constant probability of .75. According to NeCamp et al. (2020), pushing notifications has a negative impact on the mood condition when participants are already in a good mood (i.e., $S_t > 6$). Given that the majority of observations in the data have $S_t > 6$, the ATE of π_b is expected to be negative. As summarized in Table 1, the ATE of π_b is significantly negative with an effect size of .1, which is consistent with our expectations. Further investigation of the ATE composition reveals that the immediate effects are all negligible. In contrast, the DDE and DME are both significant and account for the majority of the treatment effect, indicating the importance of learning the delayed effects and mediator effects to understand the entire mechanism from actions to outcomes.

Furthermore, given that the delayed effects are all passing through S_t, rather than simply abandoning the treatment proposal, it is recommended that we consider a state-dependent policy to make more informed decisions based on the S_t and hence to improve the overall treatment effect. To support this claim, we further evaluate an optimal state-dependent policy, $\hat{\pi}_{opt}$, which is estimated by using single-stage policy estimation based on the observed data (See Appendix I for more information). According to Table 1, in contrast to π_b, the estimated ATE of $\hat{\pi}_{opt}$ is .090, with significantly positive direct effects. This further demonstrates the necessity of analyzing dynamic treatment policies as opposed to fixed action sequences, which have been the main focus of most existing literature on mediation analysis.

π_e	IDE	IME	DDE	DME	ATE
π_b	-.007(.007)	-.000(.001)	-.085(.034)	-.008(.004)	-.100 (.041)
$\hat{\pi}_{opt}$	.018(.006)	-.001(.001)	.077(.030)	-.005(.005)	.090 (.037)

Table 1. Estimated treatments effects (standard error) for π_b and $\hat{\pi}_{opt}$, compared to π_0 with no intervention.

9. Conclusion

Motivated by the growing number of applications (e.g., mobile health) with sequential decision-making over an infinite number of decision points, we propose an MMDP framework and a four-way decomposition of ATE of random policies to analyze the dynamic mediation effects. For each effect component, multiply-robust estimators with theoretical and numerical support are provided. The proposed framework can be extended in several aspects. First, the proposed methods are limited to applications with discrete action space. Meanwhile, problems such as dynamic pricing and personalized dose finding typically involve a continuous action space, which is worth studying in future work. Second, the no unmeasured confounder assumption can be violated from data collected from observational studies. Therefore, a confounded MMDP is worth investigating.

Acknowledgements

The research is partially supported by a grant from the NSF (DMS-2003637), a grant from EPSRC (EP/W014971/1), grants from the National Institutes of Health (R01 MH101459 to ZW; R01 NR013658 to JW & ZW), and an investigator grant from Precision Health Initiative at the University of Michigan to ZW. We thank Dr. Srijan Sen for generous support in the IHS data access.

References

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and regression trees*. Routledge, 2017.

Celli, V. Causal mediation analysis in economics: Objectives, assumptions, models. *Journal of Economic Surveys*, 36(1):214–234, 2022.

Chakrabortty, A., Nandy, P., and Li, H. Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models. *arXiv preprint arXiv:1809.10652*, 2018.

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.

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.

Dedecker, J. and Louhichi, S. Maximal inequalities and empirical central limit theorems. In *Empirical process techniques for dependent data*, pp. 137–159. Springer, 2002.

Díaz, I. Causal influence, causal effects, and path analysis in the presence of intermediate confounding. *arXiv preprint arXiv:2205.08000*, 2022.

Díaz, I., Hejazi, N. S., Rudolph, K. E., and van Der Laan, M. J. Nonparametric efficient causal mediation with intermediate confounders. *Biometrika*, 108(3):627–641, 2021.

Díaz, I., Williams, N., and Rudolph, K. E. Efficient and flexible causal mediation with time-varying mediators, treatments, and confounders. *arXiv preprint arXiv:2203.15085*, 2022.

Djordjilović, V., Hemerik, J., and Thoresen, M. On optimal two-stage testing of multiple mediators. *Biometrical Journal*, 2022.

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 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.

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.

Hong, G. et al. Ratio of mediator probability weighting for estimating natural direct and indirect effects. In *Proceedings of the American Statistical Association, biometrics section*, pp. 2401–2415. Alexandria, VA, USA, 2010.

Huang, J. Z. Projection estimation in multiple regression with application to functional anova models. *The annals of statistics*, 26(1):242–272, 1998.

Imai, K., Keele, L., and Tingley, D. A general approach to causal mediation analysis. *Psychological methods*, 15(4): 309, 2010.

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

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

Lange, T., Vansteelandt, S., and Bekaert, M. A simple unified approach for estimating natural direct and indirect effects. *American journal of epidemiology*, 176(3):190–195, 2012.

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

Li, L., Shi, C., Guo, T., and Jagust, W. J. Sequential pathway inference for multimodal neuroimaging analysis. *Stat*, 11 (1):e433, 2022.

Liao, P., Qi, Z., Klasnja, P., and Murphy, S. Batch policy learning in average reward markov decision processes. *arXiv preprint arXiv:2007.11771*, 2020.

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, 2021.

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

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

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.

NeCamp, T., Sen, S., Frank, E., Walton, M. A., Ionides, E. L., Fang, Y., Tewari, A., Wu, Z., et al. Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: micro-randomized trial. *Journal of medical Internet research*, 22(3):e15033, 2020.

Newey, W. K. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135, 1990.

Pearl, J. Direct and indirect effects. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 373–392. 2022.

Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. Estimation of direct causal effects. *Epidemiology*, pp. 276–284, 2006.

Preacher, K. J. Advances in mediation analysis: A survey and synthesis of new developments. *Annual review of psychology*, 66:825–852, 2015.

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

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Richiardi, L., Bellocco, R., and Zugna, D. Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology*, 42(5):1511–1519, 2013.

Rijnhart, J. J., Lamp, S. J., Valente, M. J., MacKinnon, D. P., Twisk, J. W., and Heymans, M. W. Mediation analysis methods used in observational research: a scoping review and recommendations. *BMC medical research methodology*, 21(1):1–17, 2021.

Robins, J. M. and Greenland, S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pp. 143–155, 1992.

Robins, J. M. and Richardson, T. S. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, 84:103–158, 2010.

Roth, D. L. and MacKinnon, D. P. Mediation analysis with longitudinal data. In *Longitudinal data analysis*, pp. 181–216. Routledge, 2013.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Rucker, D. D., Preacher, K. J., Tormala, Z. L., and Petty, R. E. Mediation analysis in social psychology: Current practices and new recommendations. *Social and personality psychology compass*, 5(6):359–371, 2011.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. 2020.

Selig, J. P. and Preacher, K. J. Mediation models for longitudinal data in developmental research. *Research in human development*, 6(2-3):144–164, 2009.

Sen, S., Kranzler, H. R., Krystal, J. H., Speller, H., Chan, G., Gelernter, J., and Guille, C. A prospective cohort study investigating factors associated with depression during medical internship. *Archives of general psychiatry*, 67(6): 557–565, 2010.

Shi, C. and Li, L. Testing mediation effects using logic of boolean matrices. *Journal of the American Statistical Association*, 117(540):2014–2027, 2022.

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, 2022a.

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*, pp. 1–12, 2022b.

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. T. and Shpitser, I. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816, 2012.

Tchetgen, E. J. T. and VanderWeele, T. J. On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology (Cambridge, Mass.)*, 25(2):282, 2014.

Tchetgen Tchetgen, E. J. and Shpitser, I. Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika*, 101(4):849–864, 2014.

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.

Tripathi, G. A matrix extension of the cauchy-schwarz inequality. *Economics Letters*, 63(1):1–3, 1999.

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

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

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

van der Laan, M. J. and Petersen, M. L. Direct effect models. *The international journal of biostatistics*, 4(1), 2008.

Van Der Vaart, A. W. and Wellner, J. A. Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer, 1996.

VanderWeele, T. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

VanderWeele, T. J. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, pp. 224–232, 2013.

VanderWeele, T. J. and Tchetgen Tchetgen, E. J. Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):917–938, 2017.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology (Cambridge, Mass.)*, 25(2):300, 2014.

Vansteelandt, S. and Daniel, R. M. Interventional effects for mediation analysis with multiple mediators. *Epidemiology (Cambridge, Mass.)*, 28(2):258, 2017.

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.

Zeng, P., Shao, Z., and Zhou, X. Statistical methods for mediation analysis in the era of high-throughput genomics: current successes and future challenges. *Computational and structural biotechnology journal*, 19:3209–3224, 2021.

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

Zheng, W. and van der Laan, M. Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *Journal of causal inference*, 5(2), 2017.

Zheng, W. and van der Laan, M. J. Causal mediation in a survival setting with time-dependent mediators. 2012.

A. More Details about Effect Decomposition

A.1. Effect Decomposition in the Framework of Potential Outcomes

Let $\bar{a}_t = (a_0, \cdots, a_t)$ denote a fixed treatment sequence up to time t. Let $W_t^*(\bar{a}_t)$ denote the potential covariates that would be observed at t if $\bar{a}_t$ were taken, and $\bar{W}_t^*(\bar{a}_t) = (W_0^*(\bar{a}_0), \cdots, W_t^*(\bar{a}_t))$ for $W \in \{M, R\}$. We remark that the potential states observed at t are defined analogously, but depend on the action sequence $\bar{a}_{t-1}$ instead of $\bar{a}_t$. Replacing the fixed action sequence by any random policy π, $W_t^*(\pi)$ denotes the potential covariates if the actions were taken under π.

We first focus on the effects of action and mediator on their proximal outcome. Denotes $\pi_{e,0}^t$ a policy where the first $t-1$ steps follow π_e and then follow π_0 at t. For $X \in \{S, R\}$, $X_t^*(\pi_1, \bar{M}_t^*(\pi_2))$ denotes the potential covariate if π_1 were used to determine actions and the mediators were set to levels as if π_2 were used. IDE_t and IME_t are defined as

$$\text{IDE}_t(\pi_e, \pi_0) = \mathbb{E}\left[R_t^*(\pi_e, \bar{M}_t^*(\pi_e)) - R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_e))\right],$$
$$\text{IME}_t(\pi_e, \pi_0) = \mathbb{E}[R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_e)) - R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_{e,0}^t))].$$

Given that both $\bar{A}_{t-1}$ and $\bar{M}_{t-1}$ were set to levels as if π_e were used, $\text{IDE}_t(\pi_e, \pi_0)$ contrasts the impact of A_t generated by π_e and π_0 on the proximal outcome R_t, fixing M_t to $M_t^*(\pi_e)$. $\text{IME}_t(\pi_e, \pi_0)$ compares the effect of M_t at levels $M_t^*(\pi_e)$ and $M_t^*(\pi_{e,0}^t)$ on R_t, when A_t is set by π_0.

Next, we focus on the delayed effects of the historical action sequence $\bar{A}_{t-1}$ and mediator sequence $\bar{M}_{t-1}$ on R_t. Within the MMDP framework, $\bar{A}_{t-1}$ and $\bar{M}_{t-1}$ affect R_t through S_t. Noticing that $\mathbb{E}[R_t^*(\pi_0, \bar{M}_t^*(\pi_{e,0}^t))]$ is unidentifiable due to the presence of intermediate confounders $\bar{S}_t$ (Tchetgen & VanderWeele, 2014), we adopt the RI-based approach proposed in Zheng & van der Laan (2017).

We first define the *conditional* probability density of mediator at t,

$$G_t^{\bar{a}_t'}(\cdot|\bar{s}_t, \bar{m}_{t-1}, \bar{r}_{t-1}) = p_{M_t^*(\bar{a}_t')|\bar{S}_t^*(\bar{a}_t'), \bar{M}_{t-1}^*(\bar{a}_{t-1}'), \bar{R}_{t-1}^*(\bar{a}_{t-1}')}(\cdot|\bar{s}_t, \bar{m}_{t-1}, \bar{r}_{t-1}),$$

if $\bar{a}_t'$ is assigned. At time t, given the historical trajectories $\bar{s}_t$, $\bar{m}_{t-1}$, and $\bar{r}_{t-1}$, we intervene in the mediator by randomly drawing $M_t \sim G_t^{\bar{a}_t'}(\cdot|\bar{s}_t, \bar{m}_{t-1}, \bar{r}_{t-1})$. For brevity, we omit the conditionality and let $\bar{G}_t^{\bar{a}_t'} = (G_0^{\bar{a}_0'}, \cdots, G_t^{\bar{a}_t'})$ denote the process by which the mediator is set to a conditional random draw at each time t. Using a two-stage interventional process as an illustration, we set $\bar{A}_1 = \bar{a}_1$ and $\bar{M}_1 \sim \bar{G}_1^{\bar{a}_1'}$. The generating process of $R_1^*(\bar{a}_1, \bar{G}_1^{\bar{a}_1'})$ is as follows: After observing an initial state s_0, we would first assign a treatment a_0 and set M_0 by randomly drawing $m_0 \sim G_0^{a_0'}(\cdot|s_0)$, and then measure the resulting $R_0^*(a_0, \bar{G}_0^{a_0'}) = r_0$ and $S_1^*(a_0, \bar{G}_0^{a_0'}) = s_1$. At $t = 1$, we then take action a_1 and set M_1 by randomly drawing $m_1 \sim G_1^{\bar{a}_1'}(\cdot|s_0, s_1, m_0, r_0)$, and finally observe $R_1^*(\bar{a}_1, \bar{G}_1^{\bar{a}_1'})$ as the outcome. Analogously, $R_t^*(\pi_1, \bar{G}_t^{\pi_2})$ is the potential reward if π_1 were used to determine $\bar{A}_t$ and $\bar{M}_t$ were set to have the π_2-driven conditional distributions $\bar{G}_t^{\pi_2}$. We then define the delayed effects as

$$\text{DDE}_t(\pi_e, \pi_0) = \mathbb{E}[R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_{e,0}^t)) - R_t^*(\pi_0, \bar{G}_t^{\pi_{e,0}^t})],$$
$$\text{DME}_t(\pi_e, \pi_0) = \mathbb{E}[R_t^*(\pi_0, \bar{G}_t^{\pi_{e,0}^t}) - R_t^*(\pi_0, \bar{M}_t^*(\pi_0))].$$

Setting A_t and M_t to levels as if policy π_0 were used at t, $\text{DDE}_t(\pi_e, \pi_0)$ compares the effects of $\bar{A}_{t-1}$ generated by π_e and π_0 on R_t when $\bar{M}_{t-1}$ is generated by π_e, while $\text{DME}_t(\pi_e, \pi_0)$ contrasts the effects of $\bar{M}_{t-1}$ generated by π_e and π_0 on R_t when $\bar{A}_{t-1}$ is set by π_0. See Appendix A.3 for more discussion about the non-identifiability issue and Appendix A.2 for graphical representations of each component.

Remark A.1. As suggested in Robins & Greenland (1992), there are two ways to decompose the total effect. The above definitions of direct effects and mediator effects are analogous to the Total Direct Effect (TDE) and the Pure Indirect Effect (PIE) (Robins & Greenland, 1992), while an alternative decomposition is provided in Appendix B. By replacing π_e and π_0 with $\bar{a}_t'$ and $\bar{a}_t$, IDE and IME are equivalent to TDE and PIE. Let $\tilde{\bar{a}}_t = \{\bar{a}_{t-1}', a_t\}$, we further replace $\pi_{e,0}^t$ with $\tilde{\bar{a}}_t$ to define DDE and DME. When $t > 0$, if we set $\tilde{\bar{a}}_t = \bar{a}_t'$, DDE and DME are analogous to the effect components defined in Zheng & van der Laan (2017).

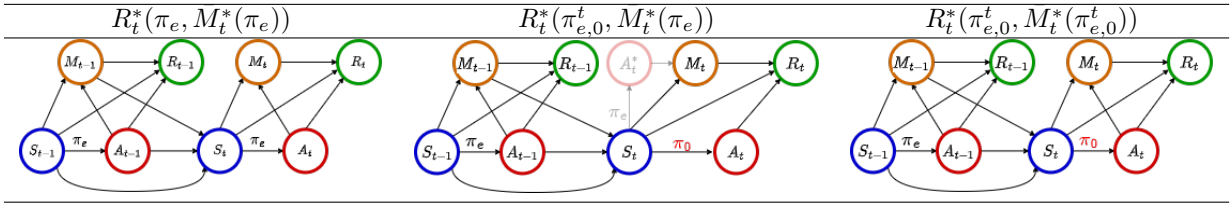

Table 2. Potential Outcomes Related to Immediate Effects.

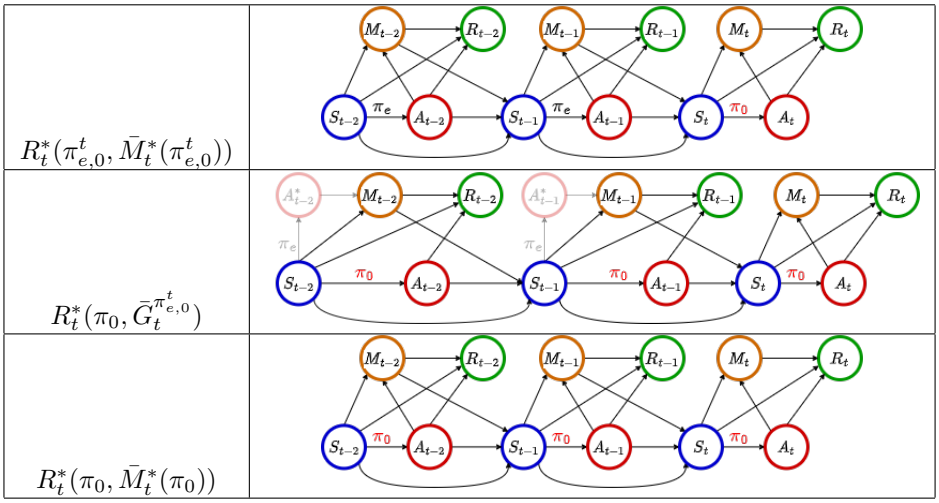

Table 3. Potential Outcomes Related to Delayed Effects.

A.2. Graphical Representation of Potential Outcomes

In Table 2 and Table 3, using causal graphs, we explicitly depict the process generating the potential reward terms involved in the effect decomposition. Specifically, $R_t^*(\pi_e, \bar{M}_t^*(\pi_e))$ is the potential reward that would be observed if π_e were used to determine $\bar{A}_t$ and $\bar{M}_t$; $R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_e))$ is the potential reward that would be observed if π_e were used to determine the historical sequences $\bar{A}_{t-1}$ and $\bar{M}_{t-1}$, while A_t were determined by π_0 and M_t were set to $M_t^*(\pi_e)$; $R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_{e,0}^t))$ is the potential reward if π_e were used to determine $\bar{A}_{t-1}$ and $\bar{M}_{t-1}$, while A_t and M_t are generated by π_0; $R_t^*(\pi_0, \bar{G}_t^{\pi_{e,0}^t})$ is the potential reward if π_0 were used to determine A_t and M_t, while the historical sequences $\bar{A}_{t-1}$ and $\bar{M}_{t-1}$ were determined by π_0 and π_e respectively; and $R_t^*(\pi_0, \bar{M}_t^*(\pi_0))$ is the potential reward that would be observed if π_0 were used to determine $\bar{A}_t$ and $\bar{M}_t$.

By definition, $\text{IDE}(\pi_e, \pi_0)$ is the contrast between causal structures of $R_t^*(\pi_e, \bar{M}_t^*(\pi_e))$ and $R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_e))$; $\text{IME}(\pi_e, \pi_0)$ is the contrast between causal structures of $R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_e))$ and $R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_{e,0}^t))$; $\text{DDE}(\pi_e, \pi_0)$ is the contrast between causal structures of $R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_{e,0}^t))$ and $R_t^*(\pi_0, \bar{G}_t^{\pi_{e,0}^t})$; and $\text{DME}(\pi_e, \pi_0)$ is the contrast between causal structures of $R_t^*(\pi_0, \bar{G}_t^{\pi_{e,0}^t})$ and $R_t^*(\pi_0, \bar{M}_t^*(\pi_0))$.

A.3. Non-identifiability Issue

To understand the non-identifiability issue, we consider two fixed action sequences $\bar{a}_t$ and $\bar{a}_t'$ and discuss the identification of $\mathbb{E}[R_t^*(\bar{a}_t, \bar{M}_t^*(\tilde{\bar{a}}_t))]$. For simplicity, let the mediator and state be discrete values. Based on the definition, let $\tilde{\bar{a}}_t = (\bar{a}_{t-1}^*, a_t)$ and $t = 1$, we have that

$$\mathbb{E}[R_1^*(\bar{a}_1, \bar{M}_1^*(\tilde{\bar{a}}_1))] = \sum_{a_0, a_1, a_0^*, m_0, m_1, s_0, s_1, s_1^*} \mathbb{E}(R|a_0, a_1, m_0, m_1, s_0, s_1)\Pr(M_1^* = m_1|a_0^*, a_1, m_0, s_0, s_1^*)$$
$$\times \Pr(S_1^*(a_0) = s_1, S_1^*(a_0^*) = s_1^*|m_0, s_0)\Pr(M_0^* = m_0|a_0^*, s_0)\Pr(S_0 = s_0).$$

While $\mathbb{E}(R|a_0, a_1, m_0, m_1, s_0, s_1)$, $\Pr(M_1^* = m_1|a_0^*, a_1, m_0, s_0, s_1^*)$, $\Pr(M_0^* = m_0|a_0^*, s_0)$, and $\Pr(S_0 = s_0)$ are identifiable from the observational data, the joint distribution of $\Pr(S_1^*(a_0) = s_1, S_1^*(a_0^*) = s_1^*|m_0, s_0)$ is not identified, leading to the non-identifiability of $\mathbb{E}[R_1^*(\bar{a}_1, \bar{M}_1^*(\tilde{\bar{a}}_1))]$. The non-identifiability of $\mathbb{E}[R_t^*(\pi_0, \bar{M}_t^*(\pi_{e,0}^t))]$ is then followed.

B. Alternative Decomposition of ATE(π_e, π_0)

In this section, we provide an alternative decomposition of ATE(π_e, π_0). Let $\tilde{G}$ denote the stochastic process selecting actions according to π_e and drawing mediators assuming π_0 was applied. Adopting the notations used in the main text, we have that

$$\text{ATE}(\pi_e, \pi_0) = \underbrace{\eta^{\pi_e} - \eta^{\tilde{G}_e}}_{\text{DME}^{(2)}(\pi_e,\pi_0)} + \underbrace{\eta^{\tilde{G}_e} - \eta^{\pi_{0,e}}}_{\text{DDE}^{(2)}(\pi_e,\pi_0)} + \underbrace{\eta^{\pi_{0,e}} - \eta^{\tilde{G}_0}}_{\text{IME}^{(2)}(\pi_e,\pi_0)} + \underbrace{\eta^{\tilde{G}_0} - \eta^{\pi_0}}_{\text{IDE}^{(2)}(\pi_e,\pi_0)}.$$

In the following subsections, we further written the alternative decomposition in the framework of potential outcomes along with the corresponding MR estimators.

B.1. Decomposition in the Framework of Potential Outcomes

We follow the notations used in the Appendix A. Another classic decomposition of the total effect is well-known as natural effect decomposition, which divides the total effect into Natural Direct Effect (NDE) (also named as Pure Direct Effect) and Natural Indirect Effect (NIE) (also named as Total Indirect Effect) (Robins & Greenland, 1992; Pearl, 2022; VanderWeele, 2013). Denotes $\pi_{0,e}^t$ a policy where the first $t-1$ steps follow π_0 and then follow π_e at t. Following the natural effect decomposition , we alternatively decompose the $\text{TE}_t(\pi_e, \pi_0)$ as follows:

$$\text{TE}_t(\pi_e, \pi_0) = \text{DME}_t^{(2)}(\pi_e, \pi_0) + \text{DDE}_t^{(2)}(\pi_e, \pi_0) + \text{IME}_t^{(2)}(\pi_e, \pi_0) + \text{IDE}_t^{(2)}(\pi_e, \pi_0),$$

where

$$\begin{aligned}
\text{DME}_t^{(2)}(\pi_e, \pi_0) &= \mathbb{E}[R_t^*(\pi_e, \bar{M}_t^*(\pi_e)) - R_t^*(\pi_e, \bar{G}_t^{\pi_{0,e}^t})],\\
\text{DDE}_t^{(2)}(\pi_e, \pi_0) &= \mathbb{E}[R_t^*(\pi_e, \bar{G}_t^{\pi_{0,e}^t}) - R_t^*(\pi_{0,e}^t, \bar{M}_t^*(\pi_{0,e}^t))],\\
\text{IME}_t^{(2)}(\pi_e, \pi_0) &= \mathbb{E}[R_t^*(\pi_{0,e}^t, \bar{M}_t^*(\pi_{0,e}^t)) - R_t^*(\pi_{0,e}^t, \bar{M}_t^*(\pi_0))],\\
\text{IDE}_t^{(2)}(\pi_e, \pi_0) &= \mathbb{E}[R_t^*(\pi_{0,e}^t, \bar{M}_t^*(\pi_0)) - R_t^*(\pi_0, \bar{M}_t^*(\pi_0))].
\end{aligned}$$

Then, for $X \in \{\text{IDE}^{(2)}, \text{IME}^{(2)}, \text{DDE}^{(2)}, \text{DME}^{(2)}\}$, we have that

$$X = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} X_t. \tag{9}$$

By replacing π_e and π_0 with $\bar{a}_t'$ and $\bar{a}_t$, $\text{IDE}^{(2)}$ and $\text{IME}^{(2)}$ are equivalent to NDE and NIE derived in Pearl (2022). Let $\tilde{\bar{a}}_t = \{\bar{a}_{t-1}, a_t'\}$, we further replace $\pi_{0,e}^t$ with $\tilde{\bar{a}}_t$ to define $\text{DDE}^{(2)}$ and $\text{DME}^{(2)}$ for fixed action sequnces. When $t > 0$, if we set $\tilde{\bar{a}}_t = \bar{a}_t$, $\text{DDE}^{(2)}$ and $\text{DME}^{(2)}$ are equivalent to NDE/NIE defined in Zheng & van der Laan (2017).

B.2. MR Estimators of the Alternative Decomposition

Similar to Section 5.3, we first define three additional Q functions:

$$\begin{aligned}
Q^{\tilde{G}_0}(s, a, m) &= \sum_{t\geq 0} \mathbb{E}^{\pi_0}[\mathbb{E}_{a^*}^{\pi_e} r(S_t, a^*, M_t) - \eta^{\tilde{G}_0}|S_0 = s, A_0 = a, M_0 = m],\\
Q^{\pi_{0,e}}(s, a, m) &= \sum_{t\geq 0} \mathbb{E}^{\pi_0}[\mathbb{E}_{a^*,m^*}^{\pi_e} r(S_t, a^*, m^*) - \eta^{\pi_{0,e}}|S_0 = s, A_0 = a, M_0 = m],\\
Q^{\tilde{G}_e}(s, a, m) &= \sum_{t\geq 0} \mathbb{E}^{\tilde{G}}[\mathbb{E}_{a^*,m^*}^{\pi_e} r(S_t, a^*, m^*) - \eta^{\tilde{G}_e}|S_0 = s, A_0 = a, M_0 = m],
\end{aligned}$$

where $\eta^{\tilde{G}_0}$ is the expected value of $\mathbb{E}^{\pi_e}_{a^*} r(S_t, a^*, M_t)$ under policy π_0, $\eta^{\pi_{0,e}}$ is the expectation of $\mathbb{E}^{\pi_e}_{a^*,m^*} r(S_t, a^*, m^*)$ under π_0, and $\eta^{\tilde{G}_e}$ is the expectation of $\mathbb{E}^{\pi_e}_{a^*,m^*} r(S_t, a^*, m^*)$ under the treatment process and the intervened mediator process of $\tilde{G}$.

Next, we construct three additional augmentation terms similar to the augmentation terms defined in the main text. Let $\rho^{(2)}(S, A, M) = \frac{\sum_a \pi_0(a|S)p(M|S,a)}{p(M|S,A)}$. We define that

$$
\begin{aligned}
I_6(O) &= \omega^{\pi_0}(S)\frac{\pi_0(A|S)}{\pi_b(A|S)}\Big\{\mathbb{E}^{\pi_e}_{a'} r(S, a', M) + \mathbb{E}^{\pi_0}_{a,m} Q^{\tilde{G}_0}(S', a, m) - \mathbb{E}_m Q^{\tilde{G}_0}(S, A, m) - \eta^{\tilde{G}_0}\Big\} \\
&\quad + \omega^{\pi_0}(S)\frac{\pi_e(A|S)}{\pi_b(A|S)}\rho^{(2)}(S, A, M)\{R - r(S, A, M)\}, \\
I_7(O) &= \omega^{\pi_0}(S)\frac{\pi_0(A|S)}{\pi_b(A|S)}\Big\{\mathbb{E}^{\pi_e}_{a',m} r(S, a', m) + \mathbb{E}^{\pi_0}_{a,m} Q^{\pi_{0,e}}(S', a, m) - \mathbb{E}_m Q^{\pi_{0,e}}(S, A, m) - \eta^{\pi_{0,e}}\Big\} \\
&\quad + \omega^{\pi_0}(S)\frac{\pi_e(A|S)}{\pi_b(A|S)}\{R - \mathbb{E}_m r(S, A, m)\}, \\
I_8(O) &= \omega^{\tilde{G}}(S)\frac{\pi_e(A|S)}{\pi_b(A|S)}\rho^{(2)}(S, A, M)\Big\{\mathbb{E}^{\pi_e}_{a',m} r(S, a', m) + \mathbb{E}^{\tilde{G}}_{a,m} Q^{\tilde{G}_e}(S', a, m) - Q^{\tilde{G}_e}(S, A, M) - \eta^{\tilde{G}_e}\Big\} \\
&\quad + \omega^{\tilde{G}}(S)\frac{\pi_e(A|S)}{\pi_b(A|S)}\Big\{R - \mathbb{E}_m r(S, A, m)\Big\} \\
&\quad + \omega^{\tilde{G}}(S)\frac{\pi_0(A|S)}{\pi_b(A|S)} \times \sum_a \pi_e(a|S)\Big[Q^{\tilde{G}_e}(S, a, M) - \sum_m p(m|A, S)Q^{\tilde{G}_e}(S, a, m)\Big]
\end{aligned}
$$

Then the MR estimator of $\text{IDE}^{(2)}(\pi_e, \pi_0)$ is

$$
\text{MR-IDE}^{(2)}(\pi_e, \pi_0) = \frac{1}{NT}\sum_{i,t} \eta^{\tilde{G}_0} - \eta^{\pi_0} + I_6(O_{i,t}) - I_5(O_{i,t}).
$$

The MR estimator of $\text{IME}^{(2)}(\pi_e, \pi_0)$ is

$$
\text{MR-IME}^{(2)}(\pi_e, \pi_0) = \frac{1}{NT}\sum_{i,t} \eta^{\pi_{0,e}} - \eta^{\tilde{G}_0} + I_7(O_{i,t}) - I_6(O_{i,t}).
$$

The MR estimator of $\text{DDE}^{(2)}(\pi_e, \pi_0)$ is

$$
\text{MR-DDE}^{(2)}(\pi_e, \pi_0) = \frac{1}{NT}\sum_{i,t} \eta^{\tilde{G}_e} - \eta^{\pi_{0,e}} + I_8(O_{i,t}) - I_7(O_{i,t}).
$$

The MR estimator of $\text{DME}^{(2)}(\pi_e, \pi_0)$ is

$$
\text{MR-IDE}^{(2)}(\pi_e, \pi_0) = \frac{1}{NT}\sum_{i,t} \eta^{\pi_e} - \eta^{\tilde{G}_e} + I_1(O_{i,t}) - I_8(O_{i,t}).
$$

Following Theorem 6.1 and Theorem 6.2, we can show that $\text{MR-IDE}^{(2)}$, $\text{MR-IME}^{(2)}$, $\text{MR-DDE}^{(2)}$, and $\text{MR-DME}^{(2)}$ are multiply robust and achieve the semi-parametric efficiency bound.

C. Proof of Theorem 4.1

This proof adheres strictly to the definitions of potential outcomes discussed in Appendix A.

We first clarify three standard assumptions, and then identify the potential rewards $\mathbb{E}\left[R^*_t(\pi, \bar{M}^*_t(\pi))\right]$ for any arbitrary policy π and $\mathbb{E}\left[R^*_t(\pi_0, \bar{G}^{\pi^t_{e,0}}_t)\right]$ using the observed data distribution, followed by the identification function for each of the $\text{IDE}(\pi_e, \pi_0)$, $\text{IME}(\pi_e, \pi_0)$, $\text{DDE}(\pi_e, \pi_0)$, and $\text{DME}(\pi_e, \pi_0)$.

C.1. Standard Assumptions

The decomposed effects are identifiable under three standard assumptions (Zheng & van der Laan, 2017; Luckett et al., 2019):

Assumption 1 (Consistency). $S_t = S_t^*(\bar{A}_{t-1})$, $M_t = M_t^*(\bar{A}_t)$, $R_t = R_t^*(\bar{A}_t)$, $S_t = S_t^*(\bar{A}_{t-1}, \bar{M}_{t-1})$, and $R_t = R_t^*(\bar{A}_t, \bar{M}_t)$ for all t.

Assumption 2 (Sequential Randomization). For all $j \geq t$: i) $(S_{j+1}^*(\bar{a}_j), M_j^*(\bar{a}_j), R_j^*(\bar{a}_j)) \perp\!\!\!\perp A_t | \bar{A}_{t-1}, \bar{M}_{t-1}, \bar{S}_t$; ii) $(S_{j+1}^*(\bar{a}_j, \bar{m}_j), R_j^*(\bar{a}_j, \bar{m}_j)) \perp\!\!\!\perp A_t | \bar{A}_{t-1}, \bar{M}_{t-1}, \bar{S}_t$; and iii) $(S_{j+1}^*(\bar{a}_j, \bar{m}_j), R_j^*(\bar{a}_j, \bar{m}_j)) \perp\!\!\!\perp M_t | \bar{A}_{t-1}, \bar{M}_{t-1}, \bar{S}_t$

Assumption 3 (Positivity). Let $h_t = (\bar{m}_t, \bar{r}_t, \bar{s}_{t+1})$. For all $t \geq 0$ and all $(h_t, \bar{a}_t, \bar{a}'_t)$: i) if $p^{\pi_b}(\bar{a}_t, h_t) > 0$, then $p^{\pi_b}(a_{t+1}|\bar{a}_t, h_t) > 0$; ii) if $p^{\pi_b}(\bar{a}'_t, h_t) > 0$, then $p^{\pi_b}(a'_{t+1}|\bar{a}'_t, h_t) > 0$; iii) if $p^{\pi_b}(r_t, s_{t+1}|\bar{a}_t, h_{t-1}, m_t) > 0$, then $p^{\pi_b}(r_t, s_{t+1}|\bar{a}'_t, h_{t-1}, m_t) > 0$; and iv) if $p^{\pi_b}(m_t|\bar{a}'_t, h_{t-1}) > 0$, then $p^{\pi_b}(m_t|\bar{a}_t, h_{t-1}) > 0$.

Assumption 1 states that the observed mediator, state, and reward are equivalent to their counterfactuals, which would be observed if the observed actions were carried out, and that the observed reward and state are consistent with the potential reward and state if the observed sequences of actions and mediators were taken. Assumption 2 requires that there are no unmeasured confounders between A_t and all of its subsequent covariates and between M_t and all of its subsequent covariates. Lastly, assumption 3 ensures that treatments and covariates are not exclusive to a specific stratum of covariates. The identification result is summarized as follows.

C.2. Identification of $\mathbb{E}\left[R_t^*(\pi, \bar{M}_t^*(\pi))\right]$

Without loss of generality, we first consider the states and mediators in discrete values. By definition, we have that

$$\begin{aligned}\mathbb{E}\left[R_t^*(\pi, \bar{M}_t^*(\pi))\right] = \sum_{\bar{a}_t, \bar{m}_t, \bar{s}_{t+1}, \bar{r}_t} & r_t \Pr(S_0 = s_0) \prod_{j=0}^{t} \pi(a_j | S_j^*(\bar{a}_{j-1}, \bar{M}_{j-1}^*(\bar{a}_{j-1})) = s_j) \\ & \times \Pr[M_j^*(\bar{a}_j) = m_j | \bar{S}_j^*(\bar{a}_{j-1}, \bar{M}_{j-1}^*(\bar{a}_{j-1})) = \bar{s}_j, \bar{M}_{j-1}^*(\bar{a}_{j-1}) = \bar{m}_{j-1}] \\ & \times \Pr[S_{j+1}^*(\bar{a}_j, \bar{M}_j^*(\bar{a}_j)) = s_{j+1}, R_j^*(\bar{a}_j, \bar{M}_j^*(\bar{a}_j)) = r_j | \bar{S}_j^*(\bar{a}_{j-1}, \bar{M}_{j-1}^*(\bar{a}_{j-1})) = \bar{s}_j, \bar{M}_j^*(\bar{a}_j) = \bar{m}_j].\end{aligned}$$

To identify the potential reward, we first consider $t = 0$ and observe that

$$\pi(a_t | S_t^*(\bar{a}_{t-1}, \bar{M}_{t-1}^*(\bar{a}_{t-1})) = s_t) = \pi(a_0 | S_0 = s_0).$$

Next, we show that

$$\begin{aligned}\Pr[M_0^*(a_0) = m_0 | S_0 = s_0] &= \Pr[M_0^*(a_0) = m_0 | A_0 = a_0, S_0 = s_0] \\ &= \Pr[M_0 = m_0 | A_0 = a_0, S_0 = s_0],\end{aligned}$$

where the first equality holds by Assumption 2 and the second equality follows from the Assumption 1. Similarly, using the same arguments, we can show that

$$\begin{aligned}&\Pr[S_1^*(a_0, M_0^*(a_0)) = s_1, R_0^*(a_0, M_0^*(a_0)) = r_0 | S_0 = s_0, M_0^*(a_0) = m_0] \\ &= \Pr[S_1^*(a_0, M_0^*(a_0)) = s_1, R_0^*(a_0, M_0^*(a_0)) = r_0 | A_0 = a_0, S_0 = s_0, M_0^*(a_0) = m_0] \\ &= \Pr[S_1 = s_1, R_0 = r_0 | A_0 = a_0, S_0 = s_0, M_0 = m_0].\end{aligned}$$

Applying the same arguments for the subsequent potential covariates repeatedly, we can show that

$$\begin{aligned}\mathbb{E}\left[R_t^*(\pi, \bar{M}_t^*(\pi))\right] = \sum_{\bar{a}_t, \bar{m}_t, \bar{s}_{t+1}, r_t} & r_t \Pr(S_0 = s_0) \prod_{j=0}^{t} \pi(a_j | S_j = s_j) \Pr[M_j = m_j | \bar{A}_j = \bar{a}_j, \bar{S}_j = \bar{s}_j, \bar{M}_{j-1} = \bar{m}_{j-1}] \\ & \times \Pr[S_{j+1} = s_{j+1}, R_j = r_j | \bar{A}_j = \bar{a}_j, \bar{S}_j = \bar{s}_j, \bar{M}_j = \bar{m}_j].\end{aligned}$$

Finally, under the assumption that the data generating process satisfied the Markov property, such that i) the distribution of A_t is independent of all the past history observations given S_t, ii) the distribution of M_t is independent of all the past history

observations given (S_t, A_t), and iii) the distributions of R_t and S_{t+1} are independent of all the past history observations given (S_t, A_t, M_t), we have that

$$\mathbb{E}\left[R_t^*(\pi, \bar{M}_t^*(\pi))\right] = \sum_{\bar{a}_t, \bar{m}_t, \bar{s}_{t+1}, r_t} r_t \Pr(S_0 = s_0) \prod_{j=0}^{t} \pi(a_j|S_j = s_j) \Pr[M_j = m_j|A_j = a_j, S_j = s_j] \times \Pr[S_{j+1} = s_{j+1}, R_j = r_j|A_j = a_j, S_j = s_j, M_j = m_j].$$

Let τ_t denote the data trajectory $\{(s_j, a_j, m_j, r_j, s_{j+1})\}_{0 \le j \le t}$. Replacing the probability mass functions by probability density functions, we have that

$$\begin{aligned}\mathbb{E}\left[R_t^*(\pi, \bar{M}_t^*(\pi))\right] &= \sum_{\tau_t} r_t \prod_{j=0}^{t} p(s_{j+1}, r_j|s_j, a_j, m_j) p(m_j|s_j, a_j) \pi(a_j|s_j) \nu(s_0) \\ &= \sum_{\tau_t} r_t p(s_{t+1}, r_t|s_t, a_t, m_t) p(m_t|s_t, a_t) \pi(a_t|s_t) \prod_{j=0}^{t-1} p^{\pi}(s_{j+1}, r_j, m_j, a_j|s_j) \nu(s_0),\end{aligned}$$

the identifiability of which is guaranteed by Assumption 3.

When $\pi = \pi_e$,

$$\mathbb{E}\left[R_t^*(\pi_e, \bar{M}_t^*(\pi_e))\right] = \sum_{\tau_t} r_t p(s_{t+1}, r_t|s_t, a_t, m_t) p(m_t|s_t, a_t) \pi_e(a_t|s_t) \prod_{j=0}^{t-1} p^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j) \nu(s_0).$$

When $\pi = \pi_0$,

$$\mathbb{E}\left[R_t^*(\pi_0, \bar{M}_t^*(\pi_0))\right] = \sum_{\tau_t} r_t p(s_{t+1}, r_t|s_t, a_t, m_t) p(m_t|s_t, a_t) \pi_0(a_t|s_t) \prod_{j=0}^{t-1} p^{\pi_0}(s_{j+1}, r_j, m_j, a_j|s_j) \nu(s_0).$$

When $\pi = \pi_{e,0}^t$,

$$\mathbb{E}\left[R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_{e,0}^t))\right] = \sum_{\tau_t} r_t p(s_{t+1}, r_t|s_t, a_t, m_t) p(m_t|s_t, a_t) \pi_0(a_t|s_t) \prod_{j=0}^{t-1} p^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j) \nu(s_0).$$

Following the same arguments, we can show that

$$\mathbb{E}\left[R_t^*(\pi_{e,0}^t, \bar{M}_t^*(\pi_e))\right] = \sum_{\tau_t} \sum_{s^*, r^*, a'} r^* p(s^*, r^*|s_t, a', m_t) \pi_0(a'|s_t) p(m_t|s_t, a_t) \pi_e(a_t|s_t) \prod_{j=0}^{t-1} p^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j) \nu(s_0).$$

C.3. Identification of $\mathbb{E}[R_t^*(\pi_0, \bar{G}_t^{\pi_{e,0}^t})]$

Without loss of generality, we first consider the states and mediators in discrete values. Let $\tilde{\bar{a}}_t = (\bar{a}'_{t-1}, a_t)$. By definition, we have that

$$\mathbb{E}[R_t^*(\pi_0, \bar{G}_t^{\tilde{\bar{a}}_t})] = \sum_{\bar{a}_t, \bar{a}'_{t-1}, \bar{m}_t, \bar{s}_{t+1}, \bar{r}_t} r_t \Pr(S_0 = s_0) \prod_{j=0}^{t-1} \pi(a_j|S_j^*(\bar{a}_{j-1}, \bar{G}_{j-1}^{\tilde{\bar{a}}_t})) = s_j) \pi(a'_j|S_j^*(\bar{a}_{j-1}, \bar{G}_{j-1}^{\tilde{\bar{a}}_t})) = s_j) \quad (10)$$

$$\times \Pr[G_j^{\tilde{\bar{a}}_t} = m_j|\bar{S}_j^*(\bar{a}_{j-1}, \bar{G}_{j-1}^{\tilde{\bar{a}}_t}) = \bar{s}_j, \bar{G}_{j-1}^{\tilde{\bar{a}}_t} = \bar{m}_{j-1}] \quad (11)$$

$$\times \Pr[S_{j+1}^*(\bar{a}_j, \bar{G}_j^{\tilde{\bar{a}}_t}) = s_{j+1}, R_j^*(\bar{a}_j, \bar{G}_j^{\tilde{\bar{a}}_t}) = r_j|\bar{S}_j^*(\bar{a}_{j-1}, \bar{G}_{j-1}^{\tilde{\bar{a}}_t}) = \bar{s}_j, \bar{G}_j^{\tilde{\bar{a}}_t} = \bar{m}_j] \quad (12)$$

$$\times \pi(a_t|S_t^*(\bar{a}_{t-1}, \bar{G}_{t-1}^{\tilde{\bar{a}}_t})) = s_t) \Pr[G_t^{\tilde{\bar{a}}_t} = m_t|\bar{S}_t^*(\bar{a}_{t-1}, \bar{G}_{t-1}^{\tilde{\bar{a}}_t}) = \bar{s}_t, \bar{G}_{t-1}^{\tilde{\bar{a}}_t} = \bar{m}_{t-1}] \quad (13)$$

$$\times \Pr[S_{t+1}^*(\bar{a}_t, \bar{G}_t^{\tilde{\bar{a}}_t}) = s_{t+1}, R_t^*(\bar{a}_t, \bar{G}_t^{\tilde{\bar{a}}_t}) = r_t|\bar{S}_t^*(\bar{a}_{t-1}, \bar{G}_{t-1}^{\tilde{\bar{a}}_t}) = \bar{s}_t, \bar{G}_t^{\tilde{\bar{a}}_t} = \bar{m}_t]. \quad (14)$$

For $j < t$, By the definition of $\bar{G}_j^{\tilde{\bar{a}}_t}$, we have that

$$\Pr[G_j^{\tilde{\bar{a}}_t} = m_j | \bar{S}_j^*(\bar{a}_{j-1}, \bar{G}_{j-1}^{\tilde{\bar{a}}_t}) = \bar{s}_j, \bar{G}_{j-1}^{\tilde{\bar{a}}_t} = \bar{m}_{j-1}] = \Pr[M_j^*(\bar{a}'_j) = m_j | \bar{S}_j^*(\bar{a}'_j) = \bar{s}_j, \bar{M}_{j-1}^*(\bar{a}'_{j-1}) = \bar{m}_{j-1}]. \quad (15)$$

Using the same arguments in C.2, we can show that equation (15) equals

$$\Pr[M_j = m_j | \bar{A}_j = \bar{a}'_j, \bar{S}_j = \bar{s}_j, \bar{M}_{j-1} = \bar{m}_{j-1}],$$

which is identifiable under Assumption 3.

Further, to show the identification of equation (12), we prove it at $j = 0$ as follows:

$$\begin{aligned}
&\Pr[S_1^*(a_0, G_0^{\tilde{\bar{a}}_t}) = s_1, R_0^*(a_0, G_0^{\tilde{\bar{a}}_t}) = r_0 | S_0 = s_0, G_0^{\tilde{\bar{a}}_t} = m_0] \\
&= \Pr[S_1^*(a_0, m_0) = s_1, R_0^*(a_0, m_0) = r_0 | S_0 = s_0, G_0^{\tilde{\bar{a}}_t} = m_0] \\
&= \Pr[S_1^*(a_0, m_0) = s_1, R_0^*(a_0, m_0) = r_0 | S_0 = s_0] \\
&= \Pr[S_1 = s_1, R_0 = r_0 | A_0 = a_0, S_0 = s_0, M_0 = m_0].
\end{aligned}$$

The second equality holds by the definition of the process $G_t^{\tilde{\bar{a}}_t}$, in which we randomly draw M_0 from $G_0^{\tilde{\bar{a}}_t}$. Specifically, given $S_0 = s_0$, $G_0^{\tilde{\bar{a}}_t}$ is independent of $S_1^*(a_0, m_0)$ and $R_1^*(a_0, m_0)$. The last equality follows from Assumption 1 and 2. A similar proof can be found in Zheng & van der Laan (2017).

Then, following the steps in C.2, we can show that

$$\mathbb{E}\left[R_t^*(\pi_0, \bar{G}_t^{\pi_{e,0}^t})\right] = \sum_{\tau_t, \bar{a}_{t-1}^*} r_t p(s_{t+1}, r_t, m_t | s_t, a_t) \pi_0(a_t | s_t) \prod_{j=0}^{t-1} p(s_{j+1}, r_j | s_j, a_j, m_j) \pi_0(a_j | s_j) p(m_j | s_j, a_j^*) \pi_e(a_j^* | s_j) \nu(s_0),$$

the identifiability of which is guaranteed by Assumption 3.

C.4. Identification of IDE(π_e, π_0), IME(π_e, π_0), DDE(π_e, π_0), DME(π_e, π_0)

Using the above identification results, the identification functions of IDE(π_e, π_0), IME(π_e, π_0), DDE(π_e, π_0), DME(π_e, π_0) are directly induced. Specifically,

$$\begin{aligned}
\text{IDE}(\pi_e, \pi_0) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\tau_t} \Big\{ r_t p(s_{t+1}, r_t | s_t, a_t, m_t) - \sum_{s^*, r^*, a'} r^* p(s^*, r^* | s_t, a', m_t) \pi_0(a' | s_t) \Big\} \\
\times p(m_t | s_t, a_t) \pi_e(a_t | s_t) \prod_{j=0}^{t-1} p^{\pi_e}(s_{j+1}, r_j, m_j, a_j | s_j) \nu(s_0),
\end{aligned}$$

$$\begin{aligned}
\text{IME}(\pi_e, \pi_0) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\tau_t} r_t p(s_{t+1}, r_t | s_t, a_t, m_t) \pi_0(a_t | s_t) [\sum_{a'} p(m_t | s_t, a') \pi_e(a' | s_t) - p(m_t | a_t, s_t)] \\
\times \prod_{j=0}^{t-1} \left[p^{\pi_e}(s_{j+1}, r_j, m_j, a_j | s_j) \right] \nu(s_0),
\end{aligned}$$

$$\begin{aligned}
\text{DDE}(\pi_e, \pi_0) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\tau_t} r_t p(s_{t+1}, r_t | s_t, a_t, m_t) p(m_t | s_t, a_t) \pi_0(a_t | s_t) \\
\times \Big\{ \prod_{j=0}^{t-1} p^{\pi_e}(s_{j+1}, r_j, m_j, a_j | s_j) - \sum_{\bar{a}_{t-1}^*} \prod_{j=0}^{t-1} p(s_{j+1}, r_j | s_j, a_j, m_j) \pi_0(a_j | s_j) p(m_j | s_j, a_j^*) \pi_e(a_j^* | s_j) \Big\} \nu(s_0),
\end{aligned}$$

and

$$\text{DME}(\pi_e,\pi_0)=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t}r_t p(s_{t+1},r_t|s_t,a_t,m_t)p(m_t|s_t,a_t)\pi_0(a_t|s_t)$$
$$\times\Big\{\sum_{\bar{a}^*_{t-1}}\prod_{j=0}^{t-1}p(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p(m_j|s_j,a_j^*)\pi_e(a_j^*|s_j)-\prod_{j=0}^{t-1}p^{\pi_0}(s_{j+1},r_j,m_j,a_j|s_j)\Big\}\nu(s_0).$$

The proof of Theorem 4.1 is thus completed.

D. Proof of Theorem 6.1

The proof of the triply robustness property of the proposed estimator is similar for IDE(π_e,π_0), IME(π_e,π_0), DDE(π_e,π_0) and DME(π_e,π_0). Here, we take the estimator of IDE as an example. Let O denote a data tuple (S,A,M,R,S'), $\rho(S,A,M)=\frac{\sum_a\pi_e(a|S)p(M|S,a)}{p(M|S,A)}$, and $\delta^\pi(S,A)=\omega^\pi(S)\frac{\pi(A|S)}{\pi_b(A|S)}$ for any policy π. Without loss of generality, we let $T_i=T,\forall i=1,\cdots,N$. We first reorganize the estimator of IDE into four parts. Recall that $\eta_d=\eta$. Let

$$\phi_1(O)=\eta^{\pi_e}-\eta^{G_e},$$
$$\phi_2(O)=\delta^{\pi_e}(S,A)\left[R+\mathbb{E}_{\substack{a\sim\pi_e(\bullet|S')\\ m\sim p(\bullet|a,S')}}Q^{\pi_e}(S',a,m)-\mathbb{E}_{m\sim p(\bullet|A,S)}Q^{\pi_e}(S,A,m)-\eta^{\pi_e}\right],$$
$$\phi_3(O)=\delta^{\pi_e}(S,A)\rho(S,A,M)\frac{\pi_0(A|S)}{\pi_e(A|S)}\{R-r(S,A,M)\},$$
$$\phi_4(O)=\delta^{\pi_e}(S,A)\left[\mathbb{E}_{a\sim\pi_0(\bullet|S)}r(S,a',M)+\mathbb{E}_{\substack{a\sim\pi_e(\bullet|S')\\ m\sim p(\bullet|S',a)}}Q^{G_e}(S',a,m)-\mathbb{E}_{m\sim p(\bullet|S,A)}Q^{G_e}(S,A,m)-\eta^{G_e}\right].$$

Then the proposed MR estimator of IDE is

$$\text{MR-IDE}(\pi_e,\pi_0)=\frac{1}{NT}\sum_{i,t}[\hat{\phi}_1(O_{i,t})+\hat{\phi}_2(O_{i,t})-\hat{\phi}_3(O_{i,t})-\hat{\phi}_4(O_{i,t})].$$

The proof of robustness can be divided into four parts. In **part I**, we show that when $\hat{\pi}_b$ and $\hat{\omega}^{\pi_e}$ are consistent, the sum of terms involving Q^{π_e}, Q^{G_e}, η^{π_e}, and η^{G_e} converges to zero by the stationary property. Then, the remaining part of MR-IDE(π_e,π_0) is

$$\frac{1}{NT}\sum_{i,t}\underbrace{\hat{\delta}^{\pi_e}(S_{i,t},A_{i,t})\left[R_{i,t}-\mathbb{E}_{a\sim\pi_0(\bullet|S)}\hat{r}(S_{i,t},a',M_{i,t})\right]}_{\hat{\phi}_5(O_{i,t})}-\hat{\phi}_3(O_{i,t}). \tag{16}$$

In **part II**, we consider the condition $\mathbb{M}_1$, where $\hat{\pi}_b$, $\hat{\omega}^{\pi_e}$, and $\hat{r}$ are consistent. We show that $\frac{1}{NT}\sum_{i,t}\hat{\phi}_3(O_{i,t})$ converged to 0, and $\frac{1}{NT}\sum_{i,t}\hat{\phi}_5(O_{i,t})$ is unbiased to the IS estimator with correctly specified π_b, ω^{π_e}, and r and thus unbiased and consistent to IDE(π_e,π_0), using the arguments used in part I. Together with the results from part I, the consistency of our estimator is proved.

In **part III**, we focus on the condition $\mathbb{M}_2$, where $\hat{\pi}_b$, $\hat{\omega}^{\pi_e}$, and $\hat{p}_m$ are consistent. We show that (16) is consistent to IDE(π_e,π_0). The consistency is then completed, together with part I.

Finally, in **part IV**, applying similar arguments in part I, we observe that $\frac{1}{NT}\sum_{i,t}\hat{\phi}_2(O_{i,t})$, $\frac{1}{NT}\sum_{i,t}\hat{\phi}_3(O_{i,t})$, and $\frac{1}{NT}\sum_{i,t}\hat{\phi}_4(O_{i,t})$ converge to 0 respectively, when $\hat{Q}^{\pi_e}$, $\hat{Q}^{G_e}$, $\hat{\eta}^{\pi_e}$, $\hat{\eta}^{G_e}$, $\hat{r}$, and $\hat{p}_m$ are consistent. Then, we show that MR-IDE$(\pi_e,\pi_0)=\hat{\phi}_1$ is consistent to IDE(π_e,π_0), with consistent $\hat{\eta}^{\pi_e}$ and $\hat{\eta}^{G_e}$. The consistency of the proposed estimator is thus proved, and the proof of triply-robustness is thus completed.

We next detail the proof for each part.

Part I. Condition: $\hat{\pi}_b$ and $\hat{\omega}^{\pi_e}$ are consistent.

First, we focus on the terms involving Q^{π_e}. Let $f_1(O;\omega^{\pi_e},\pi_b,p_m,Q^{\pi_e})$ denotes

$$\delta^{\pi_e}(A|S)\left[\mathbb{E}_{\substack{a\sim\pi_e(\bullet|S')\\ m\sim p(\bullet|a,S')}}Q^{\pi_e}(S',a,m)-\mathbb{E}_{m\sim p(\bullet|A,S)}Q^{\pi_e}(S,A,m)\right].$$

To show that $\frac{1}{NT}\sum_{i,t}f_1(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{Q}^{\pi_e})$ converges to 0, when π_b and ω^{π_e} are consistent, we decompose it into

$$\underbrace{\frac{1}{NT}\sum_{i,t}f_1(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{Q}^{\pi_e})-\frac{1}{NT}\sum_{i,t}f_1(O_{i,t};\omega^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{Q}^{\pi_e})}_{\Gamma_1}$$
$$+\underbrace{\frac{1}{NT}\sum_{i,t}f_1(O_{i,t};\omega^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{Q}^{\pi_e})-\frac{1}{NT}\sum_{i,t}f_1(O_{i,t};\omega^{\pi_e},\pi_b,\hat{p}_m,\hat{Q}^{\pi_e})}_{\Gamma_2}$$
$$+\underbrace{\frac{1}{NT}\sum_{i,t}f_1(O_{i,t};\omega^{\pi_e},\pi_b,\hat{p}_m,\hat{Q}^{\pi_e})}_{\Gamma_3}.$$

It suffices to show that Γ_1, Γ_2, and Γ_3 all converge to zero in probability.

Let us focus on Γ_1 first. Under the assumptions that Ω^{π_e}, $\mathcal{Q}^{\pi_e}$, $\mathcal{H}_m$, and Π_b are all bounded function classes and $\hat{\pi}_b(A_{i,t}|S_{i,t})$ is uniformly bounded away from zero, $|\Gamma_1|$ is upper bounded by

$$\frac{O(1)}{NT}\sum_{i,t}|\hat{\omega}^{\pi_e}(S_{i,t})-\omega^{\pi_e}(S_{i,t})|, \tag{17}$$

where $O(1)$ is some positive constant. By Markov's inequality, to prove (17) converges to zero in probability, it suffices to show that

$$\frac{1}{NT}\mathbb{E}\sum_{i,t}|\hat{\omega}^{\pi_e}(S_{i,t})-\omega^{\pi_e}(S_{i,t})|=o(1). \tag{18}$$

For any sufficient small constant $\epsilon>0$, let $\Omega^{\pi_e}(\epsilon)$ defines a set of function ω, such that,

$$\mathbb{E}_{s\sim p_\infty}|\omega(s)-\omega^{\pi_e}(s)|^2\le\epsilon^2, \tag{19}$$

where p_∞ denotes the limiting distribution of state under behavior policy. Since $\hat{\omega}^{\pi_e}$ is consistent and converge to ω^{π_e} in L_2-norm, we can show that $\hat{\omega}^{\pi_e}\in\Omega^{\pi_e}(\epsilon)$ with probability approaching to 1 (wpa1) for large NT, by Markov's inequality. Therefore, the right-hand side (RHS) of (18) is upper bounded by

$$\frac{1}{NT}\mathbb{E}\sup_{\omega\in\Omega^{\pi_e}(\epsilon)}\sum_{i,t}|\omega(S_{i,t})-\omega^{\pi_e}(S_{i,t})|, \tag{20}$$

wpa1. Then, it suffices to show that (20) is $o_p(1)$.

Implementing the empirical process theory (Van Der Vaart & Wellner, 1996), we first decompose (20) into

$$\underbrace{\frac{1}{NT}\mathbb{E}\sup_{\omega\in\Omega^{\pi_e}(\epsilon)}\left\{\sum_{i,t}|\omega(S_{i,t})-\omega^{\pi_e}(S_{i,t})|-\mathbb{E}\sum_{i,t}|\omega(S_{i,t})-\omega^{\pi_e}(S_{i,t})|\right\}}_{\Gamma_4}+\underbrace{\frac{1}{NT}\sup_{\omega\in\Omega^{\pi_e}(\epsilon)}\left\{\mathbb{E}\sum_{i,t}|\omega(S_{i,t})-\omega^{\pi_e}(S_{i,t})|\right\}}_{\Gamma_5}.$$

By the definition of $\Omega^{\pi_e}(\epsilon)$ and the Cauchy Schwartz inequality, $\mathbb{E}|\omega(S_{i,t})-\omega^{\pi_e}(S_{i,t})|\le\epsilon$ for any $\omega\in\Omega^{\pi_e}(\epsilon)$. Thus, Γ_5 is upper bounded by ϵ and converges to zero when $\epsilon\to 0$ (i.e., $\Gamma_5=o(1)$).

Next, we show that Γ_4 converges to zero as well. Under the assumption that $\Omega^{\pi_e}(\epsilon)$ is a VC-type classes with VC indices upper bounded by $O(N^k)$ for $k < \frac{1}{2}$ and ϵ is sufficiently small, using the maximal inequality (See Section 4.2 in Dedecker & Louhichi (2002) and Corollary 5.1 in Chernozhukov et al. (2014)), we can show that $\sqrt{NT}\Gamma_4$ converges to zero (i.e., $\sqrt{NT}\Gamma_4 = o_p(1)$). Therefore, we have that, $\Gamma_4 = o_p(\frac{1}{\sqrt{NT}})$. The proof of $\Gamma_1 = o_p(1)$ is then completed.

Similarly, following the steps to prove $\Gamma_1 = o_p(1)$, we can show that $\Gamma_2 = o_p(1)$. Then, it remains to show that $\Gamma_3 = o_p(1)$. By Markov's inequality, it suffices to show that $\mathbb{E}(\Gamma_3) = o(1)$. By the definition of Γ_3, $\mathbb{E}(\Gamma_3)$ is upper bounded by

$$\frac{1}{NT}\mathbb{E}\sup_{\tilde{p}\in\mathcal{H}_m, Q\in\mathcal{Q}}\sum_{i,t} f_1(O_{i,t};\omega^{\pi_e},\pi_b,\tilde{p},Q). \tag{21}$$

We first observe that, for any $Q \in \mathcal{Q}^{\pi_e}$ and $\tilde{p} \in \mathcal{H}_m$, the expectation of Γ_3 is zero. Specifically,

$$\begin{aligned}
&\mathbb{E}\Big[\omega^{\pi_e}(S)\frac{\pi_e(A|S)}{\pi_b(A|S)}\mathbb{E}_{\substack{a\sim\pi_e(\bullet|S')\\ m\sim\tilde{p}(\bullet|a,S')}} Q(S',a,m) - \mathbb{E}_{m\sim\tilde{p}(\bullet|A,S)}Q(S,A,m)\Big]\\
=&\sum_a\int_{s,m,s'} p(m,s'|a,s)p^{\pi_e}(s)\pi_e(a|s)\sum_{a'}\int_{m'} Q(s',a',m')\tilde{p}(m'|a',s')\pi_e(a'|s')\\
&-\sum_a\int_{s,m,s'} p(m,s'|a,s)p^{\pi_e}(s)\pi_e(a|s)\int_{m'} Q(s,a,m')\tilde{p}(m'|a,s)\\
=&\sum_{a'}\int_{s',m'} p^{\pi_e}(s')\pi_e(a'|s')Q(s',a',m')\tilde{p}(m'|a',s')\\
&-\sum_a\int_{s,m'} p^{\pi_e}(s)\pi_e(a|s)Q(s,a,m')\tilde{p}(m'|a,s)\\
=&0.
\end{aligned}$$

Then, following the same steps we used to bound (20), we can show that (21) is $o(1)$. Thus, $\Gamma_3 = o_p(1)$. Together with $\Gamma_1 = o_p(1)$ and $\Gamma_2 = o_p(1)$, we finish the proof of $\frac{1}{NT}\sum_{i,t} f_1(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{Q}^{\pi_e}) = o_p(1)$.

Then we focus on the terms involving Q^{G_e}. Let $f_2(O;\omega^{\pi_e},\pi_b,p_m,Q^{G_e})$ denotes

$$\delta^{\pi_e}(A|S)\left[\mathbb{E}_{\substack{a\sim\pi_e(\bullet|S')\\ m\sim p(\bullet|a,S')}} Q^{G_e}(S',a,m) - \mathbb{E}_{m\sim p(\bullet|A,S)}Q^{G_e}(S,A,m)\right].$$

Replacing $Q^{\pi_e}(S,A,m)$ with $Q^{G_e}(S,A,m)$ in the proof of $\frac{1}{NT}\sum_{i,t} f_1(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{Q}^{\pi_e}) = o_p(1)$, we can directly show that $\frac{1}{NT}\sum_{i,t} f_2(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{Q}^{G_e}) = o_p(1)$ as well.

Finally, we need to show that the sum of terms involving η^{π_e} and η^{G_e} converges to zero. Let $f_3(O;\omega^{\pi_e},\pi_b,\eta^{\pi_e},\eta^{G_e})$ denotes

$$\left[1-\omega^{\pi_e}(S)\frac{\pi_e(A|S)}{\pi_b(A|S)}\right](\eta^{\pi_e}-\eta^{G_e}).$$

For any $\eta_1 \in \mathbb{R}$ and $\eta_2 \in \mathbb{R}$, $\frac{1}{NT}\sum_{i,t} f_3(O_{i,t};\omega^{\pi_e},\pi_b,\eta_1,\eta_2)$ has mean zero. Specifically,

$$\begin{aligned}
&\mathbb{E}[\eta_1-\eta_2-\omega^{\pi_e}(S)\frac{\pi_e(A|S)}{\pi_b(A|S)}(\eta_1-\eta_2)]\\
=&\Big\{1-\sum_a\int_s p^{\pi_b}(a,s)\omega^{\pi_e}(s)\frac{\pi_e(a|s)}{\pi_b(a|s)}\Big\}(\eta_1-\eta_2)\\
=&0\times(\eta_1-\eta_2)\\
=&0.
\end{aligned}$$

Applying the same arguments in showing that $\Gamma_3 = o_p(1)$, we can show that $\frac{1}{NT}\sum_{i,t} f_3(O_{i,t};\omega^{\pi_e},\pi_b,\hat{\eta}^{\pi_e},\hat{\eta}^{G_e}) = o_p(1)$. Then, following the same steps proving that $\Gamma_1 = o_p(1)$, we can show that

$$\frac{1}{NT}\sum_{i,t}\left\{f_3(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{\eta}^{\pi_e},\hat{\eta}^{G_e}) - f_3(O_{i,t};\omega^{\pi_e},\hat{\pi}_b,\hat{\eta}^{\pi_e},\hat{\eta}^{G_e})\right\} = o_p(1),$$

and

$$\frac{1}{NT}\sum_{i,t}\left\{f_3(O_{i,t};\omega^{\pi_e},\hat{\pi}_b,\hat{\eta}^{\pi_e},\hat{\eta}^{G_e}) - f_3(O_{i,t};\omega^{\pi_e},\pi_b,\hat{\eta}^{\pi_e},\hat{\eta}^{G_e})\right\} = o_p(1).$$

Therefore, $\frac{1}{NT}\sum_{i,t} f_3(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{\eta}^{\pi_e},\hat{\eta}^{G_e}) = o_p(1)$. The proof of part I is thus completed.

Part II. Condition: $\hat{\pi}_b(A|S)$, $\hat{\omega}^{\pi_e}(S)$, and $\hat{r}$ are consistent.

With true r, ω^{π_e}, and π_b, we can show that $\mathbb{E}\phi_3(O_{i,t};\omega^{\pi_e},\pi_b,\hat{p}_m,r)$ has a mean of zero, as $\mathbb{E}[R - r(s,a,m)|S = s, A = a, M = m] = 0$. Then, using the same arguments in showing that $\Gamma_3 = o_p(1)$ in part I, we can show that $\frac{1}{NT}\sum_{i,t}\hat{\phi}_3(O_{i,t};\omega^{\pi_e},\pi_b,\hat{p}_m,r) = o_p(1)$. Next, following the same steps proving that $\Gamma_1 = o_p(1)$, we can show that

$$\frac{1}{NT}\sum_{i,t}\hat{\phi}_3(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{r}) - \frac{1}{NT}\sum_{i,t}\hat{\phi}_3(O_{i,t};\omega^{\pi_e},\pi_b,\hat{p}_m,r) = o_p(1).$$

Therefore, we finish the proof showing that $\frac{1}{NT}\sum_{i,t}\hat{\phi}_3(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{r}) = o_p(1)$. Then, it remains to show that $\frac{1}{NT}\sum_{i,t}\hat{\phi}_5(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{r})$ is consistent to IDE(π_e,π_0). Again, applying the arguments used in showing that $\Gamma_1 = o_p(1)$, we can show that

$$\frac{1}{NT}\sum_{i,t}\hat{\phi}_5(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{r}) - \frac{1}{NT}\sum_{i,t}\hat{\phi}_5(O_{i,t};\omega^{\pi_e},\pi_b,\hat{p}_m,r) = o_p(1).$$

Then, it suffices to show that $\frac{1}{NT}\sum_{i,t}\hat{\phi}_5(O_{i,t};\omega^{\pi_e},\pi_b,\hat{p}_m,r)$ is consistent to IDE(π_e,π_0). Specifically,

$$\frac{1}{NT}\sum_{i,t}\hat{\phi}_5(O_{i,t};\omega^{\pi_e},\pi_b,\hat{p}_m,r) = \frac{1}{NT}\sum_{i,t}\omega^{\pi_e}\frac{\pi_e(A_{i,t}|S_{i,t})}{\pi_b(A_{i,t}|S_{i,t})}\left[R_{i,t} - \sum_a \pi_0(a|S_{i,t})r(S_{i,t},a,M_{i,t})\right]. \tag{22}$$

Under the assumption of stationary state process, since the action space is finite, it suffices to show that,

$$\mathbb{E}_{\substack{s\sim\hat{p}^{\pi_e}\\ m\sim\hat{p}_m}}\omega^{\pi_e}\frac{\pi_e(a|s)}{\pi_b(a|s)}\left[r - \sum_{a'}\pi_0(a'|s)r(s,a',m)\right] \xrightarrow{P} \mathbb{E}_{\substack{s\sim p^{\pi_e}\\ m\sim p_m}}\omega^{\pi_e}\frac{\pi_e(a|s)}{\pi_b(a|s)}\left[r - \sum_{a'}\pi_0(a'|s)r(s,a',m)\right] \tag{23}$$

for any a. By the weak law of large number, we can show that (23) holds when NT is sufficiently large. Together with the results in part I, we thus complete the proof of Part II.

Part III. Condition: $\hat{\pi}_b(A|S)$, $\hat{\omega}^{\pi_e}(S)$, and $\hat{p}_m$ are consistent.

Applying the same arguments used in showing that $\Gamma_1 = o_p(1)$, we can show that

$$\frac{1}{NT}\sum_{i,t}\hat{\phi}_5(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{r}) - \frac{1}{NT}\sum_{i,t}\hat{\phi}_5(O_{i,t};\omega^{\pi_e},\pi_b,p_m,\hat{r}) = o_p(1),$$

and

$$\frac{1}{NT}\sum_{i,t}\hat{\phi}_3(O_{i,t};\hat{\omega}^{\pi_e},\hat{\pi}_b,\hat{p}_m,\hat{r}) - \frac{1}{NT}\sum_{i,t}\hat{\phi}_3(O_{i,t};\omega^{\pi_e},\pi_b,p_m,\hat{r}) = o_p(1).$$

Then, it suffices to show that

$$\frac{1}{NT}\sum_{i,t}\hat{\phi}_5(O_{i,t};\omega^{\pi_e},\pi_b,p_m,\hat{r}) - \frac{1}{NT}\sum_{i,t}\hat{\phi}_3(O_{i,t};\omega^{\pi_e},\pi_b,p_m,\hat{r}) \xrightarrow{p} \text{IDE}(\pi_e,\pi_0). \tag{24}$$

The LHS of (24) can be decomposed into two parts. Specifically, it suffices to show that

$$\frac{1}{NT}\sum_{i,t}\delta^{\pi_e}(S_{i,t},A_{i,t})\left\{\mathbb{E}_{a'\sim\pi_0(\bullet|S_{i,t})}\hat{r}(S_{i,t},a',M_{i,t})-\rho(S_{i,t},A_{i,t},M_{i,t})\frac{\pi_0(A_{i,t}|S_{i,t})}{\pi_e(A_{i,t}|S_{i,t})}\hat{r}(S_{i,t},A_{i,t},M_{i,t})\right\}=o_p(1), \tag{25}$$

and

$$\frac{1}{NT}\sum_{i,t}\delta^{\pi_e}(S_{i,t},A_{i,t})\left\{R_{i,t}-\rho(S_{i,t},A_{i,t},M_{i,t})\frac{\pi_0(A_{i,t}|S_{i,t})}{\pi_e(A_{i,t}|S_{i,t})}R_{i,t}\right\}\xrightarrow{P}\text{IDE}(\pi_e,\pi_0). \tag{26}$$

Following the steps showing that $\Gamma_3=o_p(1)$ in part I, since the expectation of the LHS of (25) is 0, we can show that (25) holds. Furthermore, applying the arguments used in showing (23) in part II, we can show that (26) holds. Together with the results in part I, we thus complete the proof of Part III.

Part IV. Condition: $\hat{Q}^{\pi_e}$, $\hat{Q}^{G_e}$, $\hat{\eta}^{\pi_e}$, $\hat{\eta}^{G_e}$, $\hat{r}$, and $\hat{p}_m$ are consistent.

As we discussed in the main context, with true Q^{π_e}, Q^{G_e}, η^{π_e}, η^{G_e}, r, and p_m, we can show that $\mathbb{E}\hat{\phi}_j(O_{i,t};Q^{\pi_e},Q^{G_e},\eta^{\pi_e},\eta^{G_e},r,p_m,\hat{\omega}^{\pi_e},\hat{\pi}_b)=0$ for $j=2,3,4$. Then, using the same arguments in showing that $\Gamma_3=o_p(1)$ in part 1, we can show that

$$\frac{1}{NT}\sum_{i,t}\hat{\phi}_j(O_{i,t};Q^{\pi_e},Q^{G_e},\eta^{\pi_e},\eta^{G_e},r,p_m,\hat{\omega}^{\pi_e},\hat{\pi}_b)=o_p(1),\ \text{for } j=2,3,4.$$

Then, applying the arguments used in showing that $\Gamma_1=o_p(1)$, we can further show that

$$\frac{1}{NT}\sum_{i,t}\left\{\hat{\phi}_j(O_{i,t};\hat{Q}^{\pi_e},\hat{Q}^{G_e},\hat{\eta}^{\pi_e},\hat{\eta}^{G_e},\hat{r},\hat{p}_m,\hat{\omega}^{\pi_e},\hat{\pi}_b)-\hat{\phi}_j(O_{i,t};Q^{\pi_e},Q^{G_e},\eta^{\pi_e},\eta^{G_e},r,p_m,\hat{\omega}^{\pi_e},\hat{\pi}_b)\right\}=o_p(1),$$

for $j=2,3,4$. These two results further yields that

$$\frac{1}{NT}\sum_{i,t}\hat{\phi}_j(O_{i,t};\hat{Q}^{\pi_e},\hat{Q}^{G_e},\hat{\eta}^{\pi_e},\hat{\eta}^{G_e},\hat{r},\hat{p}_m,\hat{\omega}^{\pi_e},\hat{\pi}_b)=o_p(1)$$

for $j=2,3,4$. Then, it remains to show that $\hat{\phi}_1(\hat{\eta}^{\pi_e},\hat{\eta}^{G_e})$ is consistent to $\text{IDE}(\pi_e,\pi_0)$. Applying the arguments used to show $\Gamma_1=o_p(1)$ again, under the assumption that we have that $\hat{\eta}^{\pi_e}$ and $\hat{\eta}^{G_e}$ are consistent,

$$\hat{\phi}_1(\hat{\eta}^{\pi_e},\hat{\eta}^{G_e})\xrightarrow{P}\hat{\phi}_1(\eta^{\pi_e},\eta^{G_e})=\text{IDE}(\pi_e,\pi_0),$$

where the equation holds by definition. The proof of part IV is thus completed.

E. Proof of Theorem 6.2

First, we clarify the assumption of convergence. We required that each of $\hat{Q}^{(\cdot)},\hat{\omega}^{(\cdot)},\hat{p}_m,\hat{r},\hat{\pi}_b$, and $\hat{\eta}^{(\cdot)}$ converges to its corresponding oracle value in L_2-norm at a rate of N^{-k^*}, for some $k^*>1/4$. Specifically, taking $\hat{\omega}^{\pi_e}$ as an example, we assume that

$$\sqrt{\mathbb{E}_{s\sim p_\infty}|\hat{\omega}^{\pi_e}(s)-\omega^{\pi_e}(s)|}=O_p(N^{-k^*}).$$

The proof of the efficiency of the proposed estimator is similar for $\text{IDE}(\pi_e,\pi_0)$, $\text{IME}(\pi_e,\pi_0)$, $\text{DDE}(\pi_e,\pi_0)$, and $\text{DME}(\pi_e,\pi_0)$. Here, we take the MR estimator of IDE as an example. Adopting the notation used in the Appendix D, we have the proposed multiply robust estimator of IDE as

$$\text{MR-IDE}(\pi_e,\pi_0)=\frac{1}{NT}\sum_{i,t}[\hat{\phi}_1(O_{i,t})+\hat{\phi}_2(O_{i,t})-\hat{\phi}_3(O_{i,t})-\hat{\phi}_4(O_{i,t})].$$

Taking the oracle values of the estimators (i.e., $Q^{\pi_e}, Q^{G_e}, \eta^{\pi_e}, \eta^{G_e}, r, p_m, \omega^{\pi_e}, \pi_b$), we define the oracle estimator as MR-IDE$^*(\pi_e, \pi_0) = \frac{1}{NT}\sum_{i,t}[\hat{\phi}_1^*(O_{i,t}) + \hat{\phi}_2^*(O_{i,t}) - \hat{\phi}_3^*(O_{i,t}) - \hat{\phi}_4^*(O_{i,t})]$.

We decompose the proof into two parts. In part I, we show that the proposed estimator is asymptotically equivalent to the oracle estimator, such that MR-IDE(π_e, π_0) − MR-IDE$^*(\pi_e, \pi_0) = o_p(\frac{1}{\sqrt{NT}})$. In part II, we show that the oracle estimator is asymptotically normal such that $\sqrt{N}$[MR-IDE$^*(\pi_e, \pi_0)$ − IDE(π_e, π_0)] $\xrightarrow{d} N(0, \sigma_T^2)$, where σ_T^2 is the semiparametric efficiency bound. Noticing that $\psi_2(O_{i,t})$, $\psi_3(O_{i,t})$, and $\psi_4(O_{i,t})$ are the martingale difference sequence with respect to $\{O_{i,t}\}_{0\le t\le T-1}$, under the assumption of stationarity, we have that

$$\sigma_T^2 = \frac{1}{T} Var\left[\phi_2(O_t) - \phi_3(O_t) - \phi_4(O_t)\right].$$

Therefore, we have that

$$\sqrt{NT}[\text{MR-IDE}^*(\pi_e, \pi_0) - \text{IDE}(\pi_e, \pi_0)] \xrightarrow{d} N(0, \sigma^2),$$

where $\sigma^2 = Var[\phi_2(O_t) - \phi_3(O_t) - \phi_4(O_t)]$. Finally, by Slutsky's theorem, the proposed estimator is asymptotically normally distributed with mean 0 and a variance achieving the semiparametric efficiency bound. Specifically,

$$\sqrt{NT}\Big[\text{MR-IDE}(\pi_e, \pi_0) - \text{IDE}(\pi_e, \pi_0)\Big] \xrightarrow{d} N(0, \sigma^2).$$

In the following, we detail the proof of each part.

Part I. Let $\hat{\psi} = \{\hat{Q}^{\pi_e}, \hat{Q}^{G_e}, \hat{\eta}^{\pi_e}, \hat{\eta}^{G_e}, \hat{p}_m\}$. We first decompose the MR-IDE(π_e, π_0) − MR-IDE$^*(\pi_e, \pi_0)$ in to three parts, such that MR-IDE(π_e, π_0) − MR-IDE$^*(\pi_e, \pi_0)$ = MR-IDE$^{(1)}(\hat{\psi})$ + MR-IDE$^{(2)}(\hat{\psi})$ + MR-IDE$^{(3)}(\hat{\psi}, \hat{r})$, where

$$\text{MR-IDE}^{(1)}(\hat{\psi}) = \frac{1}{NT}\sum_{i,t}\Big\{\sum_{j=1}^{2}[\hat{\phi}_1(\hat{\psi}, \omega^{\pi_e}, \pi_b, r) - \hat{\phi}_1^*(O_{i,t})] - \sum_{j=3}^{4}[\hat{\phi}_j(O_{i,t}; \hat{\psi}, \omega^{\pi_e}, \pi_b, r) - \hat{\phi}_j^*(O_{i,t})]\Big\},$$

$$\begin{aligned}\text{MR-IDE}^{(2)}(\hat{\psi}) = \frac{1}{NT}\sum_{i,t}\Big\{&\sum_{j=1}^{2}[\hat{\phi}_1(\hat{\psi}, \omega^{\pi_e}, \pi_b, \hat{r}) - \hat{\phi}_1(\hat{\psi}, \omega^{\pi_e}, \pi_b, r)] \\ &- \sum_{j=3}^{4}[\hat{\phi}_j(O_{i,t}; \hat{\psi}, \omega^{\pi_e}, \pi_b, \hat{r}) - \hat{\phi}_j(O_{i,t}; \hat{\psi}, \omega^{\pi_e}, \pi_b, r)]\Big\},\end{aligned}$$

and

$$\begin{aligned}\text{MR-IDE}^{(3)}(\hat{\psi}, \hat{r}) = \frac{1}{NT}\sum_{i,t}\Big\{&\sum_{j=1}^{2}[\hat{\phi}_1(\hat{\psi}, \hat{\omega}^{\pi_e}, \hat{\pi}_b, \hat{r}) - \hat{\phi}_1(\hat{\psi}, \omega^{\pi_e}, \pi_b, \hat{r})] \\ &- \sum_{j=3}^{4}[\hat{\phi}_j(O_{i,t}; \hat{\psi}, \hat{\omega}^{\pi_e}, \hat{\pi}_b, \hat{r}) - \hat{\phi}_j(O_{i,t}; \hat{\psi}, \omega^{\pi_e}, \pi_b, \hat{r})]\Big\}.\end{aligned}$$

Following the arguments in part I and part II of the proof of Robustness in Appendix D, the expectation of MR-IDE$^{(2)}(\hat{\psi})$ is zero. Then, applying the same arguments used in showing that $\Gamma_3 = o_p(\frac{1}{\sqrt{NT}})$ in part I of the proof of Robustness, we can show that MR-IDE$^{(1)}(\hat{\psi}) = o_p(\frac{1}{\sqrt{NT}})$ under the assumption that each component in $\hat{\phi}$ converges to its oracle value in L_2 norm at a rate of N^{-k^*} for $k^* > \frac{1}{4}$.

Then, we focus on showing that MR-IDE$^{(2)}(\hat{\psi}) = o_p(\frac{1}{\sqrt{NT}})$. Noticing that MR-IDE$^{(2)}(\hat{\psi})$ can be further decomposed as

$$\text{MR-IDE}^{(2)}(\hat{\psi}) - \text{MR-IDE}^{(2)}(\psi) + \text{MR-IDE}^{(2)}(\psi),$$

it suffices to show that MR-IDE$^{(2)}(\hat{\psi})$ − MR-IDE$^{(2)}(\psi) = o_p(\frac{1}{\sqrt{NT}})$ and MR-IDE$^{(2)}(\psi) = o_p(\frac{1}{\sqrt{NT}})$. First, similar to the part III of the proof of Theorem 6.1, the expectation of MR-IDE$^{(2)}(\psi)$ is 0, for any $\hat{r} \in \mathcal{H}_r$. Then, applying the arguments

used in showing that $\Gamma_3 = o_p(\frac{1}{\sqrt{NT}})$, we can show that MR-IDE$^{(2)}(\psi, r) = o_p(\frac{1}{\sqrt{NT}})$ under the assumption that $\hat{\omega}^{\pi_e}$ and $\hat{\pi}_b$ converge to their oracle values. Then, it remains to show that MR-IDE$^{(2)}(\hat{\psi})$ − MR-IDE$^{(2)}(\psi) = o_p(\frac{1}{\sqrt{NT}})$. It suffices to show that

$$\frac{1}{NT}\sum_{i,t}[\hat{\phi}_j(O_{i,t};\hat{\psi},\omega^{\pi_e},\pi_b,\hat{r}) - \hat{\phi}_j(O_{i,t};\hat{\psi},\omega^{\pi_e},\pi_b,r)] - [\hat{\phi}_j(O_{i,t};\psi,\omega^{\pi_e},\pi_b,\hat{r}) - \hat{\phi}_j(O_{i,t};\psi,\omega^{\pi_e},\pi_b,r)] = o_p(\frac{1}{\sqrt{NT}}), \tag{27}$$

for $j = 1, 2, 3, 4$. Here, we prove that the above equation holds for $j = 3$ as an example. For $j = 1, 2, 4$, the proof can be completed using similar arguments.

We first observe that the LHS of (27) is upper bounded by

$$\begin{aligned}
&\frac{1}{NT}\sum_{i,t}|[\hat{\phi}_j(O_{i,t};\hat{\psi},\omega^{\pi_e},\pi_b,\hat{r}) - \hat{\phi}_j(O_{i,t};\hat{\psi},\omega^{\pi_e},\pi_b,r)] - [\hat{\phi}_j(O_{i,t};\psi,\omega^{\pi_e},\pi_b,\hat{r}) - \hat{\phi}_j(O_{i,t};\psi,\omega^{\pi_e},\pi_b,r)]| \\
=&\frac{1}{NT}\sum_{i,t}|\delta^{\pi_e}(S_{i,t},A_{i,t})||\hat{\rho}(S_{i,t},A_{i,t},M_{i,t}) - \rho(S_{i,t},A_{i,t},M_{i,t})|\frac{\pi_0(A_{i,t}|S_{i,t})}{\pi_e(A_{i,t}|S_{i,t})}|r(S_{i,t},A_{i,t},M_{i,t}) - \hat{r}(S_{i,t},A_{i,t},M_{i,t})| \\
\leq&\frac{C}{NT}\sum_{i,t}|\hat{\rho}(S_{i,t},A_{i,t},M_{i,t}) - \rho(S_{i,t},A_{i,t},M_{i,t})||r(S_{i,t},A_{i,t},M_{i,t}) - \hat{r}(S_{i,t},A_{i,t},M_{i,t})| \\
\leq&\frac{C}{2NT}\sum_{i,t}|\hat{\rho}(S_{i,t},A_{i,t},M_{i,t}) - \rho(S_{i,t},A_{i,t},M_{i,t})|^2 + \frac{C}{2NT}\sum_{i,t}|r(S_{i,t},A_{i,t},M_{i,t}) - \hat{r}(S_{i,t},A_{i,t},M_{i,t})|^2 \\
=&o_p(\frac{1}{\sqrt{NT}}),
\end{aligned}$$

where C is some positive constant. The first inequality holds under the assumption that Ω^{π_e} and Π_b are bounded function classes of ω^{π_e} and π_b, respectively. The second inequality holds by applying the Cauchy-Schwartz inequality such that $ab \leq \frac{a^2+b^2}{2}$. Using the similar arguments used to bound (18) in part I of the proof of Theorem 6.1, under the assumption that $\hat{p}_m$ and $\hat{r}$ converge to their oracle values respectively in L_2 norm at a rate of $O_p(N^{-k^*})$ for some $k^* > 1/4$, we can show that the final equality holds. Similarly, we can show that MR-IDE$^{(3)}(\hat{\psi}, \hat{r}) = o_p(\frac{1}{\sqrt{NT}})$ as well. The proof of part I is thus completed.

Part II. By Central Limit Theorem, when $N \to \infty$, we can show that

$$\sqrt{N}[\text{MR-IDE}^*(\pi_e, \pi_0) - \text{IDE}(\pi_e, \pi_0)] \xrightarrow{d} N(0, \sigma_T^2),$$

for some variance σ_T^2. Then it remains to show that σ_T^2 achieves the asymptotic semiparametric efficiency bound, which is the supreme of the Cramer-Rao lower bounds for all parametric submodels (Newey, 1990).

We first introduce some additional notations. Let $\pi_{b,\theta}$, $p_{m,\theta}$ and $p_{s',r,\theta}$, and ν_θ be some parametric models parameterized by θ for π_b, p_m and $p_{s',r}$, and ν, and $\mathcal{M}$ denotes the set of all such parametric models. Then, by Theorem 1, IDE(π_e, π_0) can be represented as a function of θ. We denote the IDE(π_e, π_0) parameterized by θ as IDE$_\theta(\pi_e, \pi_0)$. By definition, the Cramer-Rao lower bound for an unbiased estimator is

$$CR(\pi_{b,\theta}, p_{m,\theta}, p_{s',r,\theta}, \nu_\theta) = \frac{\partial \text{IDE}_\theta(\pi_e,\pi_0)}{\partial\theta}\left(\mathbb{E}\left\{\frac{\partial l(\{O_t\}_{0\leq t\leq T-1};\theta)}{\partial\theta}\frac{\partial l^T(\{O_t\}_{0\leq t\leq T-1};\theta)}{\partial\theta}\right\}\right)^{-1}\frac{\partial \text{IDE}_\theta(\pi_e,\pi_0)}{\partial\theta}^T,$$

where $l(\{O_t\}_{0\leq t\leq T-1};\theta)$ is the log-likelihood function.

Suppose that there exists some parameter θ_0 such that π_{b,θ_0}, p_{m,θ_0} and p_{s',r,θ_0}, and ν_{θ_0} are the corresponding true models. Then the semiparametric efficiency bound is

$$\sup_{\mathcal{M}} CR = \sup_{\pi_b, p_m, p_{s',r}, \nu\in\mathcal{M}} CR(\pi_b, p_m, p_{s',r}, \nu) = CR(\pi_{b,\theta_0}, p_{m,\theta_0}, p_{s',r,\theta_0}, \nu_{\theta_0}). \tag{28}$$

It suffices to show that $\sigma_T^2 = \sup_{\mathcal{M}} CR$.

On the one hand, from Appendix F, we have that

$$\frac{\partial \text{IDE}_{\theta_0}(\pi_e, \pi_0)}{\partial \theta} = \mathbb{E}[(\eta^{\pi_e} - \eta^{G_e}) S(\bar{O}_{T-1})] + D_1(\theta_0) - D_2(\theta_0), \tag{29}$$

where $\bar{O}_{T-1}$ is the sequence of observations such that $\bar{O}_{T-1} = \{O_1, O_2, \cdots, O_{T-1}\}$, $S(\cdot)$ is the gradient of the log-likelihood function evaluated at $\theta = \theta_0$ (i.e., $\frac{\partial l(\{O_t\}_{0 \le t \le T-1}; \theta_0)}{\partial \theta}$),

$$D_1(\theta_0) = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \omega^{\pi_e}(S_t) \frac{\pi_e(A_t|S_t)}{\pi_{b,\theta_0}(A_t|S_t)} \{R_t + \mathbb{E}^{\pi_e}_{a^*, m^*; \theta_0} Q^{\pi_e}(S_{t+1}, a, m) - \mathbb{E}_{m;\theta_0} Q^{\pi_e}(S_t, A_t, m) - \eta^{\pi_e}\} S(\bar{O}_{T-1})\right],$$

and

$$\begin{aligned} D_2(\theta_0) = \mathbb{E}\Big[\frac{1}{T}\sum_{t=0}^{T-1} \omega^{\pi_e}(S_t) \Big\{ & \frac{\sum_a p_{\theta_0}(M_t|S_t, a)\pi_e(a|S_t)}{p_{\theta_0}(M_t|S_t, A_t)} \frac{\pi_0(A_t|S_t)}{\pi_{b,\theta_0}(A_t|S_t)} [R_t - r_{\theta_0}(S_t, A_t, M_t)] + \frac{\pi_e(A_t|S_t)}{\pi_{b,\theta_0}(A_t|S_t)} \\ & \times \{\sum_{a'} r_{\theta_0}(S_t, a', M_t)\pi_0(a'|S_t) - \eta^{G_e} + \mathbb{E}^{\pi_e}_{a,m;\theta_0} Q^{G_e}(S_{t+1}, a, m) - \mathbb{E}_{m;\theta_0} Q^{G_e}(S_t, A_t, m)\}\Big\} S(\bar{O}_{T-1})\Big]. \end{aligned}$$

Adopting the notation used in Appendix E, (29) can be rewritten as

$$\mathbb{E}\left[\left\{\frac{1}{T}\sum_t [\phi_1(O_t) + \phi_2(O_t) - \phi_3(O_t) - \phi_4(O_t)]\right\} S(\bar{O}_{T-1})\right].$$

Furthermore, since the expectation of a score function is 0, we can show that $\mathbb{E}[\text{IDE}_{\theta_0}(\pi_e, \pi_0) \times S(\bar{O}_{T-1})] = \text{IDE}_{\theta_0}(\pi_e, \pi_0) \times \mathbb{E}[S(\bar{O}_{T-1})] = 0$. Therefore, $\frac{\partial \text{IDE}_{\theta_0}(\pi_e, \pi_0)}{\partial \theta}$ can be further represented as

$$\mathbb{E}\left[\left\{\frac{1}{T}\sum_t [\phi_1(O_t) + \phi_2(O_t) - \phi_3(O_t) - \phi_4(O_t)] - \text{IDE}_{\theta_0}(\pi_e, \pi_0)\right\} S(\bar{O}_{T-1})\right].$$

By Cauchy-Schwartz inequality (Tripathi, 1999), we have that

$$\begin{aligned} \sup_{\mathcal{M}} CR \le & \mathbb{E}\left[\left\{\frac{1}{T}\sum_t [\phi_1(O_t) + \phi_2(O_t) - \phi_3(O_t) - \phi_4(O_t)] - \text{IDE}_{\theta_0}(\pi_e, \pi_0)\right\}^2\right] \\ = & Var\left\{\frac{1}{T}\sum_t [\phi_1(O_t) + \phi_2(O_t) - \phi_3(O_t) - \phi_4(O_t)] - \text{IDE}_{\theta_0}(\pi_e, \pi_0)\right\} \\ = & \sigma_T^2. \end{aligned}$$

On the other hand, by Lemma 20 in Kallus & Uehara (2022), there exists model $\mathcal{M}_{\theta'} \in \mathcal{M}$ with sufficiently large number of parameters, having $CR(\pi_{b,\theta'}, p_{m,\theta'}, p_{s',r,\theta'}, \nu_{\theta'}) = \sigma_T^2$. Therefore, we have that $\sigma_T^2 = \sup_{\mathcal{M}} CR$. The proof is thus completed.

F. Derivation of Efficient Influence Functions (EIF)

In this section, we focus on deriving the efficient influence function for each component of the average treatment effect. Without loss of generality, we assume that the state, action, mediator and reward are all discrete. While adopting the notations used in the Appendix E, we omit the subscript in p_m and $p_{s',r}$ when there is no confusion. Let τ_t denote the data trajectory $\{(s_j, a_j, m_j, r_j, s_{j+1})\}_{0 \le j \le t}$.

F.1. EIF for Immediate Direct Effect

Let us first focus on the immediate direct effect (IDE). $\text{IDE}_{\theta_0}(\pi_e, \pi_0)$ can be represented as

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t}\big\{r_t p_{\theta_0}(s_{t+1},r_t|s_t,a_t,m_t)-\sum_{s^*,r^*,a'}r^* p_{\theta_0}(s^*,r^*|s_t,a',m_t)\pi_0(a'|s_t)\big\}p_{\theta_0}(m_t|s_t,a_t)\pi_e(a_t|s_t)$$
$$\times\prod_{j=0}^{t-1}p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\nu_{\theta_0}(s_0), \quad (30)$$

where ν denotes the initial state distribution, and

$$p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)=p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)p_{\theta_0}(m_j|s_j,a_j)\pi_e(a_j|s_j).$$

Taking the derivative of (30), we have

$$\frac{\partial \text{IDE}_{\theta_0}(\pi_e,\pi_0)}{\partial\theta}=C_1+D_1-D_2,$$

where

$$C_1=(30)\times\nabla_\theta\log(\nu_{\theta_0}(s_0)),$$

$$D_1=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t}r_t\prod_{j=0}^{t}p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\sum_{j=0}^{t}\left[\nabla_\theta\log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\right]\times\nu_{\theta_0}(s_0),$$

and

$$D_2=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{a',\tau_t}r_t p_{\theta_0}(s_{t+1},r_t|s_t,a_t,m_t)\pi_0(a_t|s_t)p_{\theta_0}(m_t|s_t,a')\pi_e(a'|s_t)$$
$$\times\prod_{j=0}^{t-1}p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\Big\{\nabla_\theta\log p_{\theta_0}^{\pi_e}(s_{t+1},r_t|m_t,s_t,a_t)+\nabla_\theta\log p_{\theta_0}(m_t|s_t,a')$$
$$+\sum_{j=0}^{t-1}\left[\nabla_\theta\log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j|s_j,a_j)\right]\Big\}\times\nu_{\theta_0}(s_0). \quad (31)$$

In the following sections, we will derive C_1, D_1, and D_2, respectively.

F.1.1. C_1

We first focus on C_1. Since the expectation of a score function is zero, we have that

$$C_1=\mathbb{E}[\text{IDE}_{\theta_0}(\pi_e,\pi_0)\times\nabla_\theta\log(\nu_{\theta_0}(s_0))]=\mathbb{E}[\text{IDE}_{\theta_0}(\pi_e,\pi_0)\times S(\bar{O}_{T-1})]=\mathbb{E}[(\eta^{\pi_e}-\eta^{G_e})\times S(\bar{O}_{T-1})].$$

F.1.2. D_1

We then focus on the derivation of D_1. Notice that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t}\eta^{\pi_e}\prod_{j=0}^{t}p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\sum_{j=0}^{t}\left[\nabla_\theta\log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\right]\nu_{\theta_0}(s_0),$$
$$=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\eta^{\pi_e}\mathbb{E}[\sum_{j=0}^{t}\nabla_\theta\log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)]\times\nu_{\theta_0}(s_0),$$
$$=0,$$

where the last equation holds using the fact that the expectation of a score function is 0. Therefore,

$$D_1 = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t}[r - \eta^{\pi_e}]\prod_{j=0}^{t}p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\sum_{j=0}^{t}\left[\nabla_\theta \log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\right]\nu_{\theta_0}(s_0).$$

Together with the trick of the equality $\star$ (See Appendix F.5 for a complete proof of it), we have that

$$D_1 \overset{\star}{=} \lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\sum_{\tau_j}[r-\eta^{\pi_e}+\mathbb{E}_{a^*,m^*}^{\pi_e}Q^{\pi_e}(s_{j+1},a^*,m^*)]\prod_{k=0}^{j}p_{\theta_0}^{\pi_e}(s_{k+1},r_k,m_k,a_k|s_k)$$
$$\times\nabla_\theta\log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\times\nu_{\theta_0}(s_0). \tag{32}$$

Then, we note that

$$\sum_{s_0}\prod_{k=0}^{j}p_{\theta_0}^{\pi_e}(s_{k+1},r_k,m_k,a_k|s_k)\nu_{\theta_0}(s_0)\overset{\star\star}{=}p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)p^{\pi_e}(s_j),$$

which is the probability of $\{S_{j+1}=s_{j+1},R_j=r_j,M_j=m_j,A_j=a_j\}$ under the target polity π_e. Further, we notice that

$$\nabla_\theta\log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)=\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j)$$

Using the fact that the expectation of a score function is 0, we have

$$\sum_{s_{j+1},r_j,m_j}[p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j)\nabla_\theta\log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j|a_j,s_j)]=0$$

for any j, which follows that

$$\begin{aligned}
&\lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\sum_{\tau_j}\mathbb{E}_{m^*}Q^{\pi_e}(s_j,a_j,m^*)p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)p^{\pi_e}(s_j)\times\nabla_\theta\log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j|a_j,s_j)\\
=&\lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\sum_{\tau_{j-1},a_j,s_j}\mathbb{E}_{m^*}Q^{\pi_e}(s_j,a_j,m^*)\pi_e(a_j|s_j)p^{\pi_e}(s_j)\\
&\qquad\times\sum_{s_{j+1},r_j,m_j}[p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j)\nabla_\theta\log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j|a_j,s_j)]\\
=&0
\end{aligned}$$

Thus, combined with the D_1 in equation (32), we have that

$$\begin{aligned}
D_1 &= \lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\sum_{\tau_j}[r_j-\eta^{\pi_e}+\mathbb{E}_{a^*,m^*}^{\pi_e}Q^{\pi_e}(s_{j+1},a^*,m^*)-\mathbb{E}_{m^*}Q^{\pi_e}(s_j,a_j,m^*)]\\
&\qquad\times p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)p^{\pi_e}(s_j)\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j),\\
&= \lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\sum_{\tau_j}[r_j-\eta^{\pi_e}+\mathbb{E}_{a^*,m^*}^{\pi_e}Q^{\pi_e}(s_{j+1},a^*,m^*)-\mathbb{E}_{m^*}Q^{\pi_e}(s_j,a_j,m^*)]\\
&\times\frac{\pi_e(a_j|s_j)p^{\pi_e}(s_j)}{\pi_{b,\theta_0}(a_j|s_j)p^{\pi_b}(s_j)}p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j)\pi_{b,\theta_0}(a_j|s_j)p^{\pi_b}(s_j)\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j),\\
&= \lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\sum_{\tau_j}[r_j-\eta^{\pi_e}+\mathbb{E}_{a^*,m^*}^{\pi_e}Q^{\pi_e}(s_{j+1},a^*,m^*)-\mathbb{E}_{m^*}Q^{\pi_e}(s_j,a_j,m^*)]\\
&\times\frac{\pi_e(a_j|s_j)p^{\pi_e}(s_j)}{\pi_{b,\theta_0}(a_j|s_j)p^{\pi_b}(s_j)}p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j)\pi_{b,\theta_0}(a_j|s_j)p^{\pi_b}(s_j)\nabla_\theta\log p_{\theta_0}^{\pi_b}(s_{j+1},r_j,m_j,a_j,s_j).
\end{aligned}$$

The second equation holds by substituting $p^{\pi_e}(s_j)$ with $\frac{p^{\pi_e}(s_j)}{p^{\pi_b}(s_j)}p^{\pi_b}(s_j) = \omega^{\pi_e}(s_j)p^{\pi_b}(s_j)$ and $\pi_e(a_j|s_j)$ with $\frac{\pi_e(a_j|s_j)}{\pi_{b,\theta_0}(a_j|s_j)}\pi_{b,\theta_0}(a_j|s_j)$. The last equation holds, using the definition of $Q^{\pi_e}(s,a,m)$,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\sum_{\tau_{j-1},a_j,s_j}\frac{\pi_e(a_j|s_j)p^{\pi_e}(s_j)}{\pi_{b,\theta_0}(a_j|s_j)p^{\pi_b}(s_j)}\pi_{b,\theta_0}(a_j|s_j)p^{\pi_b}(s_j)\times\{\nabla_\theta\log\pi_{b,\theta_0}(a_j|s_j)+\log p^{\pi_b}(s_j)\}$$
$$\times\sum_{s_{j+1},r_j,m_j}[r_j-\eta^{\pi_e}+\mathbb{E}^{\pi_e}_{a^*,m^*}Q^{\pi_e}(s_{j+1},a^*,m^*)-\mathbb{E}_{m^*}Q^{\pi_e}(s_j,a_j,m^*)]p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j)=0$$

Therefore, implementing the fact that the expectation of a score function is zero and utilizing the Markov property, we obtain that,

$$D_1=\mathbb{E}\left[\omega^{\pi_e}(S)\frac{\pi_e(A|S)}{\pi_{b,\theta_0}(A|S)}\{R+\mathbb{E}^{\pi_e}_{a,m}Q^{\pi_e}(S',a,m)-\mathbb{E}_mQ^{\pi_e}(S,A,m)-\eta^{\pi_e}\}S(\bar{O}_{T-1})\right].$$

Since (S,A,M,R,S') is any arbitrary transaction tuple follows the corresponding distribution, we have that

$$D_1=\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\omega^{\pi_e}(S_t)\frac{\pi_e(A_t|S_t)}{\pi_{b,\theta_0}(A_t|S_t)}\{R_t+\mathbb{E}^{\pi_e}_{a,m}Q^{\pi_e}(S_{t+1},a,m)-\mathbb{E}_mQ^{\pi_e}(S_t,A_t,m)-\eta^{\pi_e}\}S(\bar{O}_{T-1})\right].$$

F.1.3. D_2

Finally, we focus on the derivation of D_2. Note that in equation (31), D_2 can be divided into two parts, where

$$D_2^{(1)}=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t}r_tp_{\theta_0}(s_{t+1},r_t|s_t,a_t,m_t)\pi_0(a_t|s_t)\sum_{a'}p_{\theta_0}(m_t|s_t,a')\pi_e(a'|s_t)$$
$$\times\prod_{j=0}^{t-1}p^{\pi_e}_{\theta_0}(s_{j+1},r_j,m_j,a_j|s_j)\nabla_\theta\log p^{\pi_e}_{\theta_0}(s_{t+1},r_t|m_t,s_t,a_t)\times\nu_{\theta_0}(s_0),$$

and

$$D_2^{(2)}=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{a_t,m_t,\tau_{t-1}}p_{\theta_0}(m_t|s_t,a_t)\pi_e(a_t|s_t)\sum_{s_{t+1},r_t,a'}r_tp_{\theta_0}(s_{t+1},r_t|s_t,a',m_t)\pi_0(a'|s_t)$$
$$\times\prod_{j=0}^{t-1}p^{\pi_e}_{\theta_0}(s_{j+1},r_j,m_j,a_j|s_j)\Big\{\nabla_\theta\log p_{\theta_0}(m_t|s_t,a_t)+\sum_{j=0}^{t-1}\left[\nabla_\theta\log p^{\pi_e}_{\theta_0}(s_{j+1},r_j,m_j|s_j,a_j)\right]\Big\}\nu_{\theta_0}(s_0),$$

note that here we switch the summation of a and a' and change the subscript of the summation accordingly.

Part I ($D_2^{(1)}$). Using the fact that the expectation of a score function is zero, we first obtain that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t}r_{\theta_0}(s_t,a_t,m_t)p_{\theta_0}(s_{t+1},r_t|s_t,a_t,m_t)\pi_0(a_t|s_t)\sum_{a'}p_{\theta_0}(m_t|s_t,a')\pi_e(a'|s_t)$$
$$\times\prod_{j=0}^{t-1}p^{\pi_e}_{\theta_0}(s_{j+1},r_j,m_j,a_j|s_j)\nabla_\theta\log p^{\pi_e}_{\theta_0}(s_{t+1},r_t|m_t,s_t,a_t)\times\nu_{\theta_0}(s_0)$$
$$=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{a_t,m_t,\tau_{t-1}}r_{\theta_0}(s_t,a_t,m_t)\pi_0(a_t|s_t)\sum_{a'}p_{\theta_0}(m_t|s_t,a')\pi_e(a'|s_t)\prod_{j=0}^{t-1}p^{\pi_e}_{\theta_0}(s_{j+1},r_j,m_j,a_j|s_j)$$
$$\times\sum_{s_{t+1},r_t}p_{\theta_0}(s_{t+1},r_t|s_t,a_t,m_t)\nabla_\theta\log p^{\pi_e}_{\theta_0}(s_{t+1},r_t|m_t,s_t,a_t)\times\nu_{\theta_0}(s_0)$$
$$=0$$

Therefore, it follows that

$$D_2^{(1)} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\tau_t} [r_t - r_{\theta_0}(s_t, a_t, m_t)] p_{\theta_0}(s_{t+1}, r_t | s_t, a_t, m_t) \pi_0(a_t|s_t) \sum_{a'} p_{\theta_0}(m_t|s_t, a') \pi_e(a'|s_t) \times \prod_{j=0}^{t-1} p_{\theta_0}^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j) \nabla_\theta \log p_{\theta_0}^{\pi_e}(s_{t+1}, r_t|m_t, s_t, a_t) \times \nu_{\theta_0}(s_0).$$

Furthermore, since

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{a_t, m_t, \tau_{t-1}} [\sum_{s_{t+1}, r_t} r_t p_{\theta_0}(s_{t+1}, r_t|s_t, a_t, m_t) - r_{\theta_0}(s_t, a_t, m_t)] \pi_0(a_t|s_t) \sum_{a'} p_{\theta_0}(m_t|s_t, a') \pi_e(a'|s_t) \times \prod_{j=0}^{t-1} p_{\theta_0}^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j) \sum_{j=0}^{t-1} \nabla_\theta \log p_{\theta_0}^{\pi_e}(s_{j+1}, r_j|m_j, s_j, a_j) \times \nu_{\theta_0}(s_0) = 0,$$

$D_2^{(1)}$ can be further written as

$$D_2^{(1)} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\tau_t} [r_t - r_{\theta_0}(s_t, a_t, m_t)] p_{\theta_0}(s_{t+1}, r_t|s_t, a_t, m_t) \pi_0(a_t|s_t) \sum_{a'} p_{\theta_0}(m_t|s_t, a') \pi_e(a'|s_t) \times \prod_{j=0}^{t-1} p_{\theta_0}^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j) \sum_{j=0}^{t} \nabla_\theta \log p_{\theta_0}^{\pi_e}(s_{j+1}, r_j|m_j, s_j, a_j) \times \nu_{\theta_0}(s_0).$$

Then, following the same steps in the proof of the equality D_1, we obtain that

$$D_2^{(1)} \overset{\star}{=} \lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} \sum_{\tau_j} [r_j - r_{\theta_0}(s_j, a_j, m_j)] p_{\theta_0}(s_{j+1}, r_j|s_j, a_j, m_j) \pi_0(a_j|s_j) \sum_{a'} p_{\theta_0}(m_j|s_j, a') \pi_e(a'|s_j) \times \prod_{k=0}^{j-1} p_{\theta_0}^{\pi_e}(s_{k+1}, r_k, m_k, a_k|s_k)\ \nu_{\theta_0}(s_0) \nabla_\theta \log p_{\theta_0}(s_{j+1}, r_j|s_j, a_j, m_j).$$

Similarly, using the equality $\star\star$,

$$D_2^{(1)} \overset{\star\star}{=} \lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} \sum_{s_j, a_j, m_j, r_j, s_{j+1}} [r_j - r_{\theta_0}(s_j, a_j, m_j)] p_{\theta_0}(s_{j+1}, r_j|s_j, a_j, m_j) \pi_0(a_j|s_j) \times \sum_{a'} p_{\theta_0}(m_j|s_j, a') \pi_e(a'|s_j) p^{\pi_e}(s_j) \nabla_\theta \log p_{\theta_0}(s_{j+1}, r_j|s_j, a_j, m_j).$$

Replacing $\pi_0(a_j|s_j)$ with $\frac{\pi_0(a_j|s_j)}{\pi_{b,\theta_0}(a_j|s_j)} \pi_{b,\theta_0}(a_j|s_j)$, $p^{\pi_e}(s_j)$ with $\frac{p^{\pi_e}(s_j)}{p^{\pi_b}(s_j)} p^{\pi_b}(s_j) = \omega^{\pi_e}(s_j) p^{\pi_b}(s_j)$, and $\sum_{a'} p_{\theta_0}(m_j|s_j, a') \pi_e(a'|s_j)$ with $\frac{\sum_{a'} p_{\theta_0}(m_j|s_j, a') \pi_e(a'|s_j)}{p_{\theta_0}(m_j|s_j, a_j)} p_{\theta_0}(m_j|s_j, a_j)$, we obtain that

$$D_2^{(1)} = \lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} \sum_{s_j, a_j, m_j, r_j, s_{j+1}} \omega^{\pi_e}(s_j) \frac{\sum_{a'} p_{\theta_0}(m_j|s_j, a') \pi_e(a'|s_j)}{p_{\theta_0}(m_j|s_j, a_j)} \frac{\pi_0(a_j|s_j)}{\pi_{b,\theta_0}(a_j|s_j)} [r_j - r_{\theta_0}(s_j, a_j, m_j)] \times p_{\theta_0}(s_{j+1}, r_j|s_j, a_j, m_j) p_{\theta_0}(m_j|s_j, a_j) \pi_{b,\theta_0}(a_j|s_j) p^{\pi_b}(s_j) \nabla_\theta \log p_{\theta_0}(s_{j+1}, r_j|s_j, a_j, m_j).$$

Further, since

$$\lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} \sum_{s_j, a_j, m_j} \omega^{\pi_e}(s_j) \frac{\sum_{a'} p_{\theta_0}(m_j|s_j, a') \pi_e(a'|s_j)}{p_{\theta_0}(m_j|s_j, a_j)} \frac{\pi_0(a_j|s_j)}{\pi_{b,\theta_0}(a_j|s_j)} p_{\theta_0}(m_j|s_j, a_j) \pi_{b,\theta_0}(a_j|s_j) p^{\pi_b}(s_j) \times \nabla_\theta \log p_{\theta_0}^{\pi_b}(m_j, a_j, s_j) \sum_{r_j, s_{j+1}} [r_j - r_{\theta_0}(s_j, a_j, m_j)] p_{\theta_0}(s_{j+1}, r_j|s_j, a_j, m_j) = 0,$$

we have that

$$D_2^{(1)} = \lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} \sum_{s_j,a_j,m_j,r_j,s_{j+1}} \omega^{\pi_e}(s_j) \frac{\sum_{a'} p_{\theta_0}(m_j|s_j,a')\pi_e(a'|s_j)}{p_{\theta_0}(m_j|s_j,a_j)} \frac{\pi_0(a_j|s_j)}{\pi_{b,\theta_0}(a_j|s_j)} [r_j - r_{\theta_0}(s_j,a_j,m_j)]$$
$$\times p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j) p_{\theta_0}(m_j|s_j,a_j)\pi_{b,\theta_0}(a_j|s_j) p^{\pi_b}(s_j) \nabla_\theta \log p_{\theta_0}^{\pi_b}(s_{j+1},r_j,m_j,a_j,s_j).$$

Then, combining the fact that the expectation of a score function is zero and the Markov property, we have that

$$D_2^{(1)} = \mathbb{E}\Big[\omega^{\pi_e}(S) \frac{\sum_a p_{\theta_0}(M|S,a)\pi_e(a|S)}{p_{\theta_0}(M|S,A)} \frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)} [R - r(S,A,M)] S(\bar{O}_{T-1})\Big]. \tag{33}$$

Part II ($D_2^{(2)}$). Note that, taking the sum over r_t and s_{t+1}, $D_2^{(2)}$ can be equally represented as

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{a',a_t,m_t,\tau_{t-1}} r_{\theta_0}(s_t,a',m_t)\pi_0(a'|s_t) p_{\theta_0}(m_t|s_t,a_t)\pi_e(a_t|s_t) \prod_{j=0}^{t-1} p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)$$
$$\times \Big\{ \nabla_\theta \log p_{\theta_0}(m_t|s_t,a_t) + \sum_{j=0}^{t-1} \big[\nabla_\theta \log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j|s_j,a_j)\big] \Big\} \times \nu_{\theta_0}(s_0).$$

Taking the average over r_t and s_{t+1}, and noticing that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{a',a_t,m_t,\tau_{t-1}} r_{\theta_0}(s_t,a',m_t)\pi_0(a'|s_t) p_{\theta_0}(m_t|s_t,a_t)\pi_e(a_t|s_t) \prod_{j=0}^{t-1} p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)$$
$$\times \sum_{s_{t+1},r_t} p_{\theta_0}(s_{t+1},r_t|s_t,a_t,m_t) \nabla_\theta \log p_{\theta_0}(s_{t+1},r_t|s_t,a_t,m_t)\nu_{\theta_0}(s_0) = 0,$$

we can rewrite the $D_2^{(2)}$ as

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{a',\tau_t} r_{\theta_0}(s_t,a',m_t)\pi_0(a'|s_t) \prod_{j=0}^{t} p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\nu_{\theta_0}(s_0) \sum_{j=0}^{t} \nabla_\theta \log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j|s_j,a_j).$$

Then, following the steps as we did in deriving D_1, we first show that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\tau_t} \eta^{G_e} \prod_{j=0}^{t} p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\nu_{\theta_0}(s_0) \sum_{j=0}^{t} \nabla_\theta \log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j|s_j,a_j) = 0,$$

which follows that

$$D_2^{(2)} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{a',\tau_t} [r_{\theta_0}(s_t,a',m_t)\pi_0(a'|s_t) - \eta^{G_e}] \prod_{j=0}^{t} p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\nu_{\theta_0}(s_0)$$
$$\sum_{j=0}^{t} \nabla_\theta \log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j|s_j,a_j).$$

Next, similarly, given the definition of Q^{G_e}, following the steps in deriving D_1 and combining with the trick of score function together with the Markov property, we obtain that

$$D_2^{(2)} = \mathbb{E}[\omega^{\pi_e}(S) \frac{\pi_e(A|S)}{\pi_{b,\theta_0}(A|S)} \{\sum_{a'} r_{\theta_0}(S,a',M)\pi_0(a'|S) - \eta^{G_e} + \mathbb{E}_{a,m}^{\pi_e} Q^{G_e}(S',a,m) - \mathbb{E}_m Q^{G_e}(S,A,m)\} S(\bar{O}_{T-1})]. \tag{34}$$

Combining equation(33) and equation(34), we have that

$$D_2 = \mathbb{E}\Big[\omega^{\pi_e}(S)\Big\{\frac{\sum_a p_{\theta_0}(M|S,a)\pi_e(a|S)}{p_{\theta_0}(M|S,A)}\frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)}[R - r_{\theta_0}(S,A,M)] + \frac{\pi_e(A|S)}{\pi_{b,\theta_0}(A|S)} \times \{\sum_{a'} r_{\theta_0}(S,a',M)\pi_0(a'|S) - \eta^{G_e} + \mathbb{E}^{\pi_e}_{a,m}Q^{G_e}(S',a,m) - \mathbb{E}_m Q^{G_e}(S,A,m)\}\Big\}S(\bar{O}_{T-1})\Big].$$

Since (S, A, M, R, S') is any arbitrary transaction tuple follows the corresponding distribution, we have that

$$D_2 = \mathbb{E}\Big[\frac{1}{T}\sum_{t=0}^{T-1}\omega^{\pi_e}(S_t)\Big\{\frac{\sum_a p_{\theta_0}(M_t|S_t,a)\pi_e(a|S_t)}{p_{\theta_0}(M_t|S_t,A_t)}\frac{\pi_0(A_t|S_t)}{\pi_{b,\theta_0}(A_t|S_t)}[R_t - r_{\theta_0}(S_t,A_t,M_t)] + \frac{\pi_e(A_t|S_t)}{\pi_{b,\theta_0}(A_t|S_t)} \times \{\sum_{a'} r_{\theta_0}(S_t,a',M_t)\pi_0(a'|S_t) - \eta^{G_e} + \mathbb{E}^{\pi_e}_{a,m}Q^{G_e}(S_{t+1},a,m) - \mathbb{E}_m Q^{G_e}(S_t,A_t,m)\}\Big\}S(\bar{O}_{T-1})\Big].$$

F.1.4. Derivative of $\text{IDE}_{\theta_0}(\pi_e, \pi_0)$

Given C_1, D_1, and D_2, the derivative of $\text{IDE}_{\theta_0}(\pi_e, \pi_0)$ is $\eta^{\pi_e} - \eta^{G_e} + I_1 - I_2$, where

$$I_1 = \mathbb{E}[\omega^{\pi_e}(S)\frac{\pi_e(A|S)}{\pi_{b,\theta_0}(A|S)}\{R - \eta^{\pi_e} + \mathbb{E}^{\pi_e}_{a,m}Q^{\pi_e}(S',a,m) - \mathbb{E}_m Q^{\pi_e}(S,A,m)\}],$$

and

$$I_2 = \mathbb{E}\Big[\omega^{\pi_e}(S)\Big\{\frac{\sum_a p_{\theta_0}(M|S,a)\pi_e(a|S)}{p_{\theta_0}(M|S,A)}\frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)}[R - r_{\theta_0}(S,A,M)] + \frac{\pi_e(A|S)}{\pi_{b,\theta_0}(A|S)} \times \{\sum_{a'} r_{\theta_0}(S,a',M)\pi_0(a'|S) + \mathbb{E}^{\pi_e}_{a,m}Q^{G_e}(S',a,m) - \mathbb{E}_m Q^{G_e}(S,A,m) - \eta^{G_e}\}\Big\}\Big].$$

F.2. EIF for Immediate Mediator Effect

Immediate Mediator Effect (IME) can be represented as

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t} r_t p_{\theta_0}(s_{t+1},r_t|s_t,a_t,m_t)\pi_0(a_t|s_t)[\sum_{a'} p_{\theta_0}(m_t|s_t,a')\pi_e(a'|s_t) - p_{\theta_0}(m_t|a_t,s_t)] \times \prod_{j=0}^{t-1}\left[p^{\pi_e}_{\theta_0}(s_{j+1},r_j,m_j,a_j|s_j)\right]\nu_{\theta_0}(s_0). \quad (35)$$

Taking the derivative of $\text{IME}_{\theta_0}(\pi_e, \pi_0)$, we get that

$$\frac{\partial \text{IME}_{\theta_0}(\pi_e,\pi_0)}{\partial\theta_0} = C_2 + D_2 - D_3,$$

where

$$C_2 = (35) \times \nabla_\theta \log(\nu_{\theta_0}(s_0)) = \mathbb{E}[\text{IME}_{\theta_0}(\pi_e,\pi_0) \times S(\bar{O}_{T-1})] = \mathbb{E}[(\eta^{G_e} - \eta^{\pi_{e,0}}) \times S(\bar{O}_{T-1})],$$

D_2 is derived in Appendix F.1.3, and

$$D_3 = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t} r_t p_{\theta_0}(s_{t+1},r_t|s_t,a_t,m_t)p_{\theta_0}(m_t|s_t,a_t)\pi_0(a_t|s_t)\prod_{j=0}^{t-1}\left[p^{\pi_e}_{\theta_0}(s_{j+1},r_j,m_j,a_j|s_j)\right] \times \Big[\sum_{j=0}^{t-1}[\nabla_\theta \log p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j)] + \nabla_\theta \log p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)\Big]\nu_{\theta_0}(s_0),$$

which can be represented as the sum of two parts. Specifically, $D_3 = D_3^{(1)} + D_3^{(2)}$, where

$$D_3^{(1)} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\tau_t} r_t p_{\theta_0}(s_{t+1}, r_t|s_t, a_t, m_t) p_{\theta_0}(m_t|s_t, a_t)\pi_0(a_t|s_t) \prod_{j=0}^{t-1} \left[p_{\theta_0}^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j)\right] \times \nabla_\theta \log p_{\theta_0}(s_{t+1}, r_t, m_t|s_t, a_t)\nu_{\theta_0}(s_0),$$

and

$$D_3^{(2)} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\tau_t} r_t p_{\theta_0}(s_{t+1}, r_t|s_t, a_t, m_t) p_{\theta_0}(m_t|s_t, a_t)\pi_0(a_t|s_t) \prod_{j=0}^{t-1} \left[p_{\theta_0}^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j)\right] \times \sum_{j=0}^{t-1} \nabla_\theta \log p_{\theta_0}(s_{j+1}, r_j, m_j|a_j, s_j)\nu_{\theta_0}(s_0).$$

F.2.1. D_3

Part I ($D_3^{(1)}$). First, using the fact that the expectation of a score function is 0, we notice that,

$$\begin{aligned}
&\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\tau_t} \mathbb{E}_{m^*} r_{\theta_0}(s_t, a_t, m^*) p_{\theta_0}(s_{t+1}, r_t|s_t, a_t, m_t) p_{\theta_0}(m_t|s_t, a_t)\pi_0(a_t|s_t) \\
&\qquad \times \prod_{j=0}^{t-1} \left[p_{\theta_0}^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j)\right] \nabla_\theta \log p_{\theta_0}(s_{t+1}, r_t, m_t|s_t, a_t)\nu_{\theta_0}(s_0), \\
=& \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{a_t, \tau_{t-1}} \mathbb{E}_{m^*} r_{\theta_0}(s_t, a_t, m^*)\pi_0(a_t|s_t) \prod_{j=0}^{t-1} \left[p_{\theta_0}^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j)\right] \nu_{\theta_0}(s_0) \\
&\qquad \times \sum_{s_{t+1}, r_t, m_t} p_{\theta_0}(s_{t+1}, r_t, m_t|s_t, a_t) \nabla_\theta \log p_{\theta_0}(s_{t+1}, r_t, m_t|s_t, a_t), \\
=& 0,
\end{aligned}$$

which follows that

$$\begin{aligned}
D_3^{(1)} =& \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\tau_t} [r_t - \mathbb{E}_{m^*} r_{\theta_0}(s_t, a_t, m^*)] p_{\theta_0}(s_{t+1}, r_t|s_t, a_t, m_t) p_{\theta_0}(m_t|s_t, a_t)\pi_0(a_t|s_t) \\
&\qquad \times \prod_{j=0}^{t-1} \left[p_{\theta_0}^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j)\right] \nu_{\theta_0}(s_0) \nabla_\theta \log p_{\theta_0}(s_{t+1}, r_t, m_t|s_t, a_t), \\
=& \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{s_{t+1}, r_t, a_t, m_t, s_t} [r_t - \mathbb{E}_{m^*} r_{\theta_0}(s_t, a_t, m^*)] p_{\theta_0}(s_{t+1}, r_t, m_t|s_t, a_t)\pi_0(a_t|s_t) \\
&\qquad \times p^{\pi_e}(s_t) \nabla_\theta \log p_{\theta_0}(s_{t+1}, r_t, m_t|s_t, a_t).
\end{aligned}$$

Replacing the $\pi_0(a_t|s_t)$ with $\frac{\pi_0(a_t|s_t)}{\pi_{b,\theta_0}(a_t|s_t)}\pi_{b,\theta_0}(a_t|s_t)$, and $p^{\pi_e}(s_t)$ with $\frac{p^{\pi_e}(s_t)}{p^{\pi_b}(s_t)} p^{\pi_b}(s_t) = \omega^{\pi_e}(s_t) p^{\pi_b}(s_t)$, we obtain that

$$D_3^{(1)} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{s_{t+1}, r_t, a_t, m_t, s_t} \omega^{\pi_e}(s_t) \frac{\pi_0(a_t|s_t)}{\pi_{b,\theta_0}(a_t|s_t)} [r_t - \mathbb{E}_{m^*} r_{\theta_0}(s_t, a_t, m^*)] \times p_{\theta_0}(s_{t+1}, r_t, m_t|s_t, a_t)\pi_{b,\theta_0}(a_t|s_t) p^{\pi_b}(s_t) \nabla_\theta \log p_{\theta_0}(s_{t+1}, r_t, m_t|s_t, a_t).$$

Further, since

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{a_t,s_t}\omega^{\pi_e}(s_t)\frac{\pi_0(a_t|s_t)}{\pi_{b,\theta_0}(a_t|s_t)}\pi_{b,\theta_0}(a_t|s_t)p^{\pi_b}(s_t)\nabla_\theta\log p_{\theta_0}^{\pi_b}(a_t,s_t)$$
$$\times\sum_{s_{t+1},r_t,m_t}[r_t-\mathbb{E}_{m^*}r_{\theta_0}(s_t,a_t,m^*)]p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)=0,$$

we have that

$$D_3^{(1)}=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{s_{t+1},r_t,a_t,m_t,s_t}\omega^{\pi_e}(s_t)\frac{\pi_0(a_t|s_t)}{\pi_{b,\theta_0}(a_t|s_t)}[r_t-\mathbb{E}_{m^*}r_{\theta_0}(s_t,a_t,m^*)]$$
$$\times p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)\pi_{b,\theta_0}(a_t|s_t)p^{\pi_b}(s_t)\nabla_\theta\log p_{\theta_0}^{\pi_b}(s_{t+1},r_t,m_t,a_t,s_t).$$

Lastly, combining the fact that the expectation of a score function is 0 and the Markov property, we finalize the derivation of $D_3^{(1)}$ with

$$D_3^{(1)}=\mathbb{E}\Big[\omega^{\pi_e}(S)\frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)}[R-\mathbb{E}_m r_{\theta_0}(S,A,m)]S(\bar{O}_{T-1})\Big]. \tag{36}$$

Part II ($D_3^{(2)}$). We first rewrite the $D_3^{(2)}$ as

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{a_t,m_t,\tau_{t-1}}r_{\theta_0}(s_t,a_t,m_t)p_{\theta_0}(m_t|s_t,a_t)\pi_0(a_t|s_t)\prod_{j=0}^{t-1}\left[p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\right]$$
$$\times\sum_{j=0}^{t-1}\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j)\nu_{\theta_0}(s_0).$$

Taking the additional average over s^*, r^*, m^*, and a^*, and noticing that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{a_t,m_t,\tau_{t-1}}r_{\theta_0}(s_t,a_t,m_t)p_{\theta_0}(m_t|s_t,a_t)\pi_0(a_t|s_t)\prod_{j=0}^{t-1}\left[p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\right]$$
$$\times\sum_{s^*,r^*,m^*,a^*}p_{\theta_0}(s^*,r^*,m^*|s_t,a^*)\pi_e(a^*|s_t)\nabla_\theta\log p_{\theta_0}(s^*,r^*,m^*|s_t,a^*)\nu_{\theta_0}(s_0)=0,$$

we further represent $D_3^{(2)}$ as

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t}\sum_{m^*,a^*}p_{\theta_0}(m^*|s_t,a^*)r_{\theta_0}(s_t,a^*,m^*)\pi_0(a^*|s_t)\prod_{j=0}^{t}\left[p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\right]$$
$$\times\sum_{j=0}^{t}\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j)\nu_{\theta_0}(s_0).$$

Note that we change the subscript of the summations accordingly. Then, following the steps we processed to derive the D_1, we first show that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t}\eta^{\pi_{e,0}}\prod_{j=0}^{t}\left[p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\right]\sum_{j=0}^{t}\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j,m_j|a_j,s_j)\nu_{\theta_0}(s_0)=0.$$

Therefore,

$$D_3^{(2)} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{\tau_t} \Big[\sum_{m^*,a^*} r_{\theta_0}(s_t, a^*, m^*) p_{\theta_0}(m^*|s_t, a^*) \pi_0(a^*|s_t) - \eta^{\pi_{e,0}} \Big] \prod_{j=0}^{t} \big[p_{\theta_0}^{\pi_e}(s_{j+1}, r_j, m_j, a_j|s_j) \big] \times \sum_{j=0}^{t} \nabla_\theta \log p_{\theta_0}(s_{j+1}, r_j, m_j|a_j, s_j) \nu_{\theta_0}(s_0).$$

Next, using the equality properties $\star$ and $\star\star$, together with the definition of $Q^{\pi_{e,0}}(s, a, m)$ and the trick of score functions, we can show that

$$\begin{aligned} D_3^{(2)} = \lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} \sum_{\tau_j} \Big[\sum_{m^*,a^*} & r_{\theta_0}(s_t, a^*, m^*) p_{\theta_0}(m^*|s_t, a^*) \pi_0(a^*|s_t) - \eta^{\pi_{e,0}} + \mathbb{E}_{a^*,m^*}^{\pi_e} Q^{\pi_{e,0}}(s_{j+1}, a^*, m^*) \\ & - \mathbb{E}_{m^*} Q^{\pi_{e,0}}(s_j, a_j, m^*) \Big] \frac{\pi_e(a_j|s_j) p^{\pi_e}(s_j)}{\pi_{b,\theta_0}(a_j|s_j) p^{\pi_b}(s_j)} p_{\theta_0}(s_{j+1}, r_j, m_j|a_j, s_j) \\ & \times \pi_{b,\theta_0}(a_j|s_j) p^{\pi_b}(s_j) \nabla_\theta \log p_{\theta_0}^{\pi_b}(s_{j+1}, r_j, m_j, a_j, s_j). \end{aligned}$$

Implementing the fact that the expectation of a score function is zero and utilizing the Markov property, we finally obtain that,

$$\begin{aligned} D_3^{(2)} = \mathbb{E}\Big[\omega^{\pi_e}(S) \frac{\pi_e(A|S)}{\pi_{b,\theta_0}(A|S)} \{ \sum_{a'} & \mathbb{E}_{m\sim p_{\theta_0}(\bullet|S,a')} r_{\theta_0}(S, a', m) \pi_0(a'|S) \\ & + \mathbb{E}_{a,m}^{\pi_e} Q^{\pi_{e,0}}(S', a, m) - \mathbb{E}_m Q^{\pi_{e,0}}(S, A, m) - \eta^{\pi_{e,0}} \} S(\bar{O}_{T-1}) \Big]. \end{aligned} \tag{37}$$

Combining equation (36) and equation (37), we have that

$$\begin{aligned} D_3 = \mathbb{E}\Big[\omega^{\pi_e}(S) \Big\{ \frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)} [R - \mathbb{E}_m r_{\theta_0}(S, A, m)] + \frac{\pi_e(A|S)}{\pi_{b,\theta_0}(A|S)} \{ \sum_{a'} & \mathbb{E}_{m\sim p_{\theta_0}(\bullet|S,a')} r_{\theta_0}(S, a', m) \pi_0(a'|S) \\ & + \mathbb{E}_{a,m}^{\pi_e} Q^{\pi_{e,0}}(S', a, m) - \mathbb{E}_m Q^{\pi_{e,0}}(S, A, m) - \eta^{\pi_{e,0}} \} \Big\} S(\bar{O}_{T-1}) \Big]. \end{aligned}$$

Since (S, A, M, R, S') is any arbitrary transaction tuple follows the corresponding distribution, we have that

$$\begin{aligned} D_3 = \mathbb{E}\Big[\frac{1}{T} \sum_{t=0}^{T-1} \omega^{\pi_e}(S_t) \Big\{ \frac{\pi_0(A_t|S_t)}{\pi_{b,\theta_0}(A_t|S_t)} [R_t - \mathbb{E}_m r_{\theta_0}(S_t, A_t, m)] + \frac{\pi_e(A_t|S_t)}{\pi_{b,\theta_0}(A_t|S_t)} \{ \sum_{a'} & \mathbb{E}_{m\sim p(\bullet|S_t,a')} r_{\theta_0}(S_t, a', m) \pi_0(a'|S_t) \\ & + \mathbb{E}_{a,m}^{\pi_e} Q^{\pi_{e,0}}(S_{t+1}, a, m) - \mathbb{E}_m Q^{\pi_{e,0}}(S_t, A_t, m) - \eta^{\pi_{e,0}} \} \Big\} S(\bar{O}_{T-1}) \Big]. \end{aligned}$$

F.2.2. Efficient Function

Given C_2, D_2, and D_3, the efficient influence function for $\text{IME}_{\theta_0}(\pi_e, \pi_0)$ is $\eta^{G_e} - \eta^{\pi_{e,0}} + I_2 - I_3$, where

$$\begin{aligned} I_2 = \mathbb{E}\Big[\omega^{\pi_e}(S) \Big\{ \frac{\sum_a p_{\theta_0}(M|S, a) \pi_e(a|S)}{p_{\theta_0}(M|S, A)} \frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)} [R - r_{\theta_0}(S, A, M)] + \frac{\pi_e(A|S)}{\pi_{b,\theta_0}(A|S)} & \\ \times \{ \sum_{a'} r_{\theta_0}(S, a', M) \pi_0(a'|S) + \mathbb{E}_{a,m}^{\pi_e} Q^{G_e}(S', a, m) - \mathbb{E}_m Q^{G_e}(S, A, m) - \eta^{G_e} \} \Big\} \Big], \end{aligned}$$

and

$$\begin{aligned} I_3 = \mathbb{E}\Big[\omega^{\pi_e}(S) \Big\{ \frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)} [R - \mathbb{E}_m r_{\theta_0}(S, A, m)] + \frac{\pi_e(A|S)}{\pi_{b,\theta_0}(A|S)} \{ \sum_{a'} & \mathbb{E}_{m\sim p_{\theta_0}(\bullet|S,a')} r_{\theta_0}(S, a', m) \pi_0(a'|S) \\ & + \mathbb{E}_{a,m}^{\pi_e} Q^{\pi_{e,0}}(S', a, m) - \mathbb{E}_m Q^{\pi_{e,0}}(S, A, m) - \eta^{\pi_{e,0}} \} \Big\} \Big]. \end{aligned}$$

F.3. EIF for Delayed Direct Effect

Delayed Direct Effect (DDE) can be represented as

$$\begin{aligned}\text{DDE}(\pi_e,\pi_0) = &\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t} r_t p(s_{t+1},r_t|s_t,a_t,m_t)p(m_t|s_t,a_t)\pi_0(a_t|s_t)\\ &\times\Big\{\prod_{j=0}^{t-1}p^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j) - \sum_{\bar{a}^*_{t-1}}\prod_{j=0}^{t-1}p(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p(m_j|s_j,a^*_j)\pi_e(a^*_j|s_j)\Big\}\nu(s_0). \quad (38)\end{aligned}$$

Taking the derivative of $\text{DDE}_{\theta_0}(\pi_e,\pi_0)$, we get that

$$\frac{\partial \text{DDE}_{\theta_0}(\pi_e,\pi_0)}{\partial\theta_0} = C_3 + D_3 - D_4,$$

where

$$C_3 = (38)\times\nabla_\theta\log(\nu_{\theta_0}(s_0)) = \mathbb{E}[\text{DDE}_{\theta_0}(\pi_e,\pi_0)\times S(\bar{O}_{T-1})] = \mathbb{E}[(\eta^{\pi_{e,0}} - \eta^{G_0})\times S(\bar{O}_{T-1})],$$

D_3 is derived in Appendix F.2.1, and

$$\begin{aligned}D_4 = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t,\bar{a}^*_{t-1}} &r_t p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)\pi_0(a_t|s_t)\\ &\prod_{j=0}^{t-1}p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p_{\theta_0}(m_j|s_j,a^*_j)\pi_e(a^*_j|s_j)\nu_{\theta_0}(s_0)\\ &\Big\{\underbrace{\nabla_\theta\log p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)}_{D_4^{(1)}} + \underbrace{\sum_{j=0}^{t-1}\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)}_{D_4^{(2)}} + \underbrace{\sum_{j=0}^{t-1}\nabla_\theta\log p^{\pi_e}_{\theta_0}(m_j|s_j,a^*_j)}_{D_4^{(3)}}\Big\}.\end{aligned}$$

F.3.1. D_4

Part I ($D_4^{(1)}$). First, using the fact that the expectation of a score function is 0, we notice that,

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t,\bar{a}^*_{t-1}}\mathbb{E}_{m'}r_{\theta_0}(s_t,a_t,m')p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)\pi_0(a_t|s_t)\nabla_\theta\log p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)\\ &\qquad\prod_{j=0}^{t-1}p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p_{\theta_0}(m_j|s_j,a^*_j)\pi_e(a^*_j|s_j)\nu_{\theta_0}(s_0),\\ =&\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{a_t,\tau_{t-1},\bar{a}^*_{t-1}}\mathbb{E}_{m'}r_{\theta_0}(s_t,a_t,m')\pi_0(a_t|s_t)\prod_{j=0}^{t-1}p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p_{\theta_0}(m_j|s_j,a^*_j)\pi_e(a^*_j|s_j)\\ &\qquad\sum_{s_{t+1},r_t,m_t}p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)\nabla_\theta\log p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)\nu_{\theta_0}(s_0),\\ =&0,\end{aligned}$$

which follows that

$$D_4^{(1)} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t,\bar{a}_{t-1}^*}[r_t-\mathbb{E}_{m'}r_{\theta_0}(s_t,a_t,m')]p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)\pi_0(a_t|s_t)\nabla_\theta\log p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)$$
$$\prod_{j=0}^{t-1}p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p_{\theta_0}(m_j|s_j,a_j^*)\pi_e(a_j^*|s_j)\nu_{\theta_0}(s_0),$$
$$= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{s_{t+1},r_t,m_t,a_t,s_t}[r_t-\mathbb{E}_{m'}r_{\theta_0}(s_t,a_t,m')]p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)\pi_0(a_t|s_t)p_{\theta_0}^G(s_t)$$
$$\nabla_\theta\log p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t).$$

The last equation holds, since

$$\sum_{s_0,\tau_{t-1},\bar{a}_{t-1}^*}\prod_{j=0}^{t-1}p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p_{\theta_0}(m_j|s_j,a_j^*)\pi_e(a_j^*|s_j)\nu_{\theta_0}(s_0)=p_{\theta_0}^G(s_t).$$

Replacing the $\pi_0(a_t|s_t)$ with $\frac{\pi_0(a_t|s_t)}{\pi_{b,\theta_0}(a_t|s_t)}\pi_{b,\theta_0}(a_t|s_t)$, and $p_{\theta_0}^G(s_t)$ with $\frac{p_{\theta_0}^G(s_t)}{p^{\pi_b}(s_t)}p^{\pi_b}(s_t)=\omega_{\theta_0}^G(s_t)p^{\pi_b}(s_t)$, we obtain that

$$D_4^{(1)} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{s_{t+1},r_t,m_t,a_t,s_t}\omega_{\theta_0}^G(s_t)\frac{\pi_0(a_t|s_t)}{\pi_{b,\theta_0}(a_t|s_t)}[r_t-\mathbb{E}_{m'}r_{\theta_0}(s_t,a_t,m')]$$
$$p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)\pi_{b,\theta_0}(a_t|s_t)p^{\pi_b}(s_t)\nabla_\theta\log p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t).$$

Further, since

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{a_t,s_t}\omega_{\theta_0}^G(s_t)\frac{\pi_0(a_t|s_t)}{\pi_{b,\theta_0}(a_t|s_t)}\pi_{b,\theta_0}(a_t|s_t)p^{\pi_b}(s_t)\nabla_\theta\log p_{\theta_0}^{\pi_b}(a_t,s_t)$$
$$\times\sum_{s_{t+1},r_t,m_t}[r_t-\mathbb{E}_{m'}r_{\theta_0}(s_t,a_t,m')]p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)=0,$$

we have that

$$D_4^{(1)} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{s_{t+1},r_t,a_t,m_t,s_t}\omega_{\theta_0}^G(s_t)\frac{\pi_0(a_t|s_t)}{\pi_{b,\theta_0}(a_t|s_t)}[r_t-\mathbb{E}_{m'}r_{\theta_0}(s_t,a_t,m')]$$
$$\times p_{\theta_0}(s_{t+1},r_t,m_t|s_t,a_t)\pi_{b,\theta_0}(a_t|s_t)p^{\pi_b}(s_t)\nabla_\theta\log p_{\theta_0}^{\pi_b}(s_{t+1},r_t,m_t,a_t,s_t).$$

Lastly, combining the fact that the expectation of a score function is 0 and the Markov property, we finalize the derivation of $D_4^{(1)}$ with

$$D_4^{(1)} = \mathbb{E}\Big[\omega_{\theta_0}^G(S)\frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)}[R-\mathbb{E}_m r_{\theta_0}(S,A,m)]S(\bar{O}_{T-1})\Big]. \tag{39}$$

Part II ($D_4^{(2)}$). Taking the additional average over s', r', a', m', and $\tilde{a}$, and noticing that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{a_t,m_t,\tau_{t-1},\bar{a}_{t-1}^*}r_{\theta_0}(s_t,a_t,m_t)p_{\theta_0}(m_t|s_t,a_t)\pi_0(a_t|s_t)$$
$$\times\prod_{j=0}^{t-1}p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p_{\theta_0}(m_j|s_j,a_j^*)\pi_e(a_j^*|s_j)\nu_{\theta_0}(s_0)$$
$$\times\sum_{s',r',a',m',\tilde{a}}p_{\theta_0}(s',r'|s_t,a',m')\pi_0(a'|s_t)p_{\theta_0}(m'|s_t,\tilde{a})\pi_e(\tilde{a}|s_t)\nabla_\theta\log p_{\theta_0}(s',r'|s_t,a',m')=0,$$

we further represent $D_4^{(2)}$ as

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t,\bar{a}_t^*}\sum_{m',a'}p_{\theta_0}(m'|s_t,a')r_{\theta_0}(s_t,a',m')\pi_0(a'|s_t)$$
$$\times\prod_{j=0}^{t}p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p_{\theta_0}(m_j|s_j,a_j^*)\pi_e(a_j^*|s_j)$$
$$\times\sum_{j=0}^{t}\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\nu_{\theta_0}(s_0).$$

Note that we change the subscript of the summations accordingly. Then, following the steps we processed to derive the D_1, we first show that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t,\bar{a}_t^*}\eta^{G_0}\prod_{j=0}^{t}p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p_{\theta_0}(m_j|s_j,a_j^*)\pi_e(a_j^*|s_j)$$
$$\times\sum_{j=0}^{t}\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j|a_j,s_j,m_j)\nu_{\theta_0}(s_0)=0.$$

Therefore,

$$D_4^{(2)}=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t,\bar{a}_t^*}\Big[\sum_{m',a'}p_{\theta_0}(m'|s_t,a')r_{\theta_0}(s_t,a',m')\pi_0(a'|s_t)-\eta^{G_0}\Big]$$
$$\prod_{j=0}^{t}p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p_{\theta_0}(m_j|s_j,a_j^*)\pi_e(a_j^*|s_j)\sum_{j=0}^{t}\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j|a_j,s_j,m_j)\nu_{\theta_0}(s_0).$$

Similar to $\star\star$, we have that

$$\sum_{s_0,\tau_{t-1},\bar{a}_{t-1}^*}\prod_{j=0}^{t-1}p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p_{\theta_0}(m_j|s_j,a_j^*)\pi_e(a_j^*|s_j)\nu_{\theta_0}(s_0)\overset{\star\star\star}{=}p_{\theta_0}^G(s_t).$$

Next, using the equality properties $\star$ and $\star\star\star$, together with the definition of $Q^{G_0}(s,a,m)$ and the trick of score functions, we can show that

$$D_4^{(2)}=\lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\sum_{s_{j+1},r_j,a_j,s_j,m_j,a_j^*}\Big[\sum_{m',a'}p_{\theta_0}(m'|s_j,a')r_{\theta_0}(s_j,a',m')\pi_0(a'|s_j)-\eta^{G_0}+\mathbb{E}_{a,m}^G Q^{G_0}(s_{j+1},a,m)\Big]$$
$$\times p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j)p_{\theta_0}(m_j|s_j,a_j^*)\pi_e(a_j^*|s_j)p^G(s_j)\nabla_\theta\log p_{\theta_0}(s_{j+1},r_j|a_j,s_j,m_j).\quad(40)$$

Then, following the steps in deriving D_1, we have that

$$D_4^{(2)}=\lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\sum_{s_{j+1},r_j,a_j,s_j,m_j}\omega_{\theta_0}^G(s_j)\frac{\sum_{a'}p_{\theta_0}(m_j|s_j,a')\pi_e(a'|s_j)}{p_{\theta_0}(m_j|s_j,a_j)}\frac{\pi_0(a_j|s_j)}{\pi_{b,\theta_0}(a_j|s_j)}$$
$$\times\Big[\sum_{m',a'}p_{\theta_0}(m'|s_j,a')r_{\theta_0}(s_j,a',m')\pi_0(a'|s_j)-\eta^{G_0}+\mathbb{E}_{a,m}^G Q^{G_0}(s_{j+1},a,m)-Q^{G_0}(s_j,a_j,m_j)\Big]$$
$$\times p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)p_{\theta_0}(m_j|s_j,a_j)\pi_{b,\theta_0}(a_j|s_j)p^{\pi_b}(s_j)\nabla_\theta\log p_{\theta_0}^{\pi_b}(s_{j+1},r_j,m_j,a_j,s_j).$$

Finally, combining the fact that the expectation of a score function is zero and the Markov property, we have that

$$D_4^{(2)}=\mathbb{E}\Big[\Big\{\sum_{m',a'}p_{\theta_0}(m'|S,a')r_{\theta_0}(S,a',m')\pi_0(a'|S)-\eta^{G_0}+\mathbb{E}_{a,m}^G Q^{G_0}(S',a,m)-Q^{G_0}(S,A,M)\Big\}$$
$$\times\omega_{\theta_0}^G(S)\frac{\sum_{a'}p_{\theta_0}(M|S,a')\pi_e(a'|S)}{p_{\theta_0}(M|S,A)}\frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)}S(\bar{O}_{T-1})\Big].\quad(41)$$

Part III ($D_4^{(3)}$). Following the same steps used in deriving the equation (40), we can show that,

$$D_4^{(3)} = \lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} \sum_{a_j,s_j,m_j,a_j^*} \sum_{s_{j+1},r_j} \Big[\sum_{m',a'} p_{\theta_0}(m'|s_j,a') r_{\theta_0}(s_j,a',m') \pi_0(a'|s_j) - \eta^{G_0} + \mathbb{E}^G_{a,m} Q^{G_0}(s_{j+1},a,m) \Big]$$
$$\times p_{\theta_0}(s_{j+1},r_j|s_j,a_j,m_j)\pi_0(a_j|s_j) p_{\theta_0}(m_j|s_j,a_j^*)\pi_e(a_j^*|s_j) p^G(s_j) \nabla_\theta \log p_{\theta_0}(m_j|s_j,a_j^*). \quad (42)$$

Based on the definition of Q^{G_0} and the corresponding Bellman equation, we have that

$$\mathbb{E}\Big[\sum_{m',a'} p_{\theta_0}(m'|s_j,a') r_{\theta_0}(s_j,a',m') \pi_0(a'|s_j) - \eta^{G_0} + \mathbb{E}^G_{a,m} Q^{G_0}(s_{j+1},a,m)|s_j,a_j,m_j \Big] = Q^{G_0}(s_j,a_j,m_j).$$

Therefore, (42) can be rewritten as

$$\lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} \sum_{s_j,m_j,a_j} [\sum_{a'} Q^{G_0}(s_j,a',m_j)\pi_0(a'|s_j)] p_{\theta_0}(m_j|s_j,a_j)\pi_e(a_j|s_j) p^G(s_j) \nabla_\theta \log p_{\theta_0}(m_j|s_j,a_j).$$

Notice that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} \sum_{s_j,m_j,a_j} \Big\{ \sum_{a',m_j^*} Q^{G_0}(s_j,a',m_j^*) p_{\theta_0}(m_j^*|s_j,a_j)\pi_0(a'|s_j) \Big\} p_{\theta_0}(m_j|s_j,a_j)\pi_e(a_j|s_j) p^G(s_j)$$
$$\times \nabla_\theta \log p_{\theta_0}(m_j|s_j,a_j)$$
$$= \lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} \sum_{s_j,a_j} \Big\{ \sum_{a',m_j^*} Q^{G_0}(s_j,a',m_j^*) p_{\theta_0}(m_j^*|s_j,a_j)\pi_0(a'|s_j) \Big\} \pi_e(a_j|s_j) p^G(s_j) \sum_{m_j} p_{\theta_0}(m_j|s_j,a_j))$$
$$\times \nabla_\theta \log p_{\theta_0}(m_j|s_j,a_j$$
$$=0.$$

Therefore, we have that

$$D_4^{(3)} = \lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} \sum_{s_j,m_j,a_j} \Big\{ [\sum_{a'} Q^{G_0}(s_j,a',m_j)\pi_0(a'|s_j)] - \sum_{a',m_j^*} Q^{G_0}(s_j,a',m_j^*) p_{\theta_0}(m_j^*|s_j,a_j)\pi_0(a'|s_j) \Big\}$$
$$\times p_{\theta_0}(m_j|s_j,a_j)\pi_e(a_j|s_j) p^G(s_j) \nabla_\theta \log p_{\theta_0}(m_j|s_j,a_j).$$

Following the same steps we used in getting the final expression of $D_4^{(1)}$, we can show that

$$D_4^{(3)} = \mathbb{E}\Big[\omega^G_{\theta_0}(S) \frac{\pi_e(A|S)}{\pi_{b,\theta_0}(A|S)} \Big\{ [\sum_{a'} Q^{G_0}(S,a',M)\pi_0(a'|S)] - \sum_{a',m'} Q^{G_0}(S,a',m') p_{\theta_0}(m'|S,A)\pi_0(a'|S) \Big\} S(\bar{O}_{T-1}) \Big]. \quad (43)$$

Combining $D_4^{(1)}$, $D_4^{(2)}$, and $D_4^{(3)}$, we have that

$$D_4 = \mathbb{E}\Big[\omega^G(S) \frac{\pi_0(A|S)}{\pi_b(A|S)} \Big[\Big\{ R - \mathbb{E}_m r(S,A,m) \Big\} + \rho(S,A,M) \Big\{ \mathbb{E}^{\pi_0}_{a',m} r(S,a',m) + \mathbb{E}^G_{a,m} Q^{G_0}(S',a,m)$$
$$- Q^{G_0}(S,A,M) - \eta^{G_0} \Big\} \Big] + \omega^G(S) \frac{\pi_e(A|S)}{\pi_b(A|S)} \sum_a \pi_0(a|S) \Big\{ Q^{G_0}(S,a,M) - \sum_m p(m|S,A) Q^{G_0}(S,a,m) \Big\} \Big].$$

Since (S,A,M,R,S') is any arbitrary transaction tuple follows the corresponding distribution, we have that

$$D_4 = \mathbb{E}\Big[\frac{1}{T} \sum_{t=0}^{T-1} \omega^G(S_t) \frac{\pi_0(A_t|S_t)}{\pi_b(A_t|S_t)} \Big[\Big\{ R_t - \mathbb{E}_m r(S_t,A_t,m) \Big\} + \rho(S_t,A_t,M_t) \Big\{ \mathbb{E}^{\pi_0}_{a',m} r(S_t,a',m) + \mathbb{E}^G_{a,m} Q^{G_0}(S_{t+1},a,m)$$
$$- Q^{G_0}(S_t,A_t,M_t) - \eta^{G_0} \Big\} \Big] + \omega^G(S_t) \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} \sum_a \pi_0(a|S_t) \Big\{ Q^{G_0}(S_t,a,M_t) - \sum_m p(m|S_t,A_t) Q^{G_0}(S_t,a,m) \Big\} \Big].$$

F.3.2. EFFICIENT FUNCTION

Given C_3, D_3, and D_4, the efficient influence function for $\text{DDE}_{\theta_0}(\pi_e, \pi_0)$ is $\eta^{\pi_{e,0}} - \eta^{G_0} + I_3 - I_4$, where

$$I_3 = \mathbb{E}\Big[\omega^{\pi_e}(S)\Big\{\frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)}[R - \mathbb{E}_m r_{\theta_0}(S, A, m)] + \frac{\pi_e(A|S)}{\pi_{b,\theta_0}(A|S)}\{\sum_{a'} \mathbb{E}_{m\sim p_{\theta_0}(\bullet|S,a')} r_{\theta_0}(S, a', m)\pi_0(a'|S) + \mathbb{E}^{\pi_e}_{a,m} Q^{\pi_{e,0}}(S', a, m) - \mathbb{E}_m Q^{\pi_{e,0}}(S, A, m) - \eta^{\pi_{e,0}}\}\Big\}\Big],$$

and

$$I_4 = \mathbb{E}\Big[\omega^G(S)\frac{\pi_0(A|S)}{\pi_b(A|S)}\Big[\Big\{R - \mathbb{E}_m r(S, A, m)\Big\} + \rho(S, A, M)\Big\{\mathbb{E}^{\pi_0}_{a',m} r(S, a', m) + \mathbb{E}^G_{a,m} Q^{G_0}(S', a, m) - Q^{G_0}(S, A, M) - \eta^{G_0}\Big\}\Big] + \omega^G(S)\frac{\pi_e(A|S)}{\pi_b(A|S)}\sum_a \pi_0(a|S)\Big\{Q^{G_0}(S, a, M) - \sum_m p(m|S, A)Q^{G_0}(S, a, m)\Big\}\Big].$$

F.4. EIF for Delayed Mediator Effect

Delayed Mediator Effect (DME) can be represented as

$$\text{DME}(\pi_e, \pi_0) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t} r_t p(s_{t+1}, r_t|s_t, a_t, m_t)p(m_t|s_t, a_t)\pi_0(a_t|s_t) \times \Big\{\sum_{\bar{a}^*_{t-1}}\prod_{j=0}^{t-1} p(s_{j+1}, r_j|s_j, a_j, m_j)\pi_0(a_j|s_j)p(m_j|s_j, a^*_j)\pi_e(a^*_j|s_j) - \prod_{j=0}^{t-1} p^{\pi_0}(s_{j+1}, r_j, m_j, a_j|s_j)\Big\}\nu(s_0). \quad (44)$$

Taking the derivative of $\text{DME}_{\theta_0}(\pi_e, \pi_0)$, we get that

$$\frac{\partial \text{DME}_{\theta_0}(\pi_e, \pi_0)}{\partial \theta_0} = C_4 + D_4 - D_5,$$

where

$$C_4 = (44) \times \nabla_\theta \log(\nu_{\theta_0}(s_0)) = \mathbb{E}[\text{DME}_{\theta_0}(\pi_e, \pi_0) \times S(\bar{O}_{T-1})] = \mathbb{E}[(\eta^{G_0} - \eta^{\pi_0}) \times S(\bar{O}_{T-1})],$$

D_4 is derived in Appendix F.3.1, and

$$D_5 = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau_t} r_t \prod_{j=0}^{t} p^{\pi_0}_{\theta_0}(s_{j+1}, r_j, m_j, a_j|s_j)\sum_{j=0}^{t}\left[\nabla_\theta \log p^{\pi_0}_{\theta_0}(s_{j+1}, r_j, m_j, a_j|s_j)\right] \times \nu_{\theta_0}(s_0),$$

Notice that D_5 is similar as D_1, and can be derived similarly as D_1 by replacing the π_e in D_1 with π_0. Therefore, with the definition of $Q^{\pi_0}(s, a, m)$, we can show that

$$D_5 = \mathbb{E}\Big[\omega^{\pi_0}(S)\frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)}\{R + \sum_{a'}\mathbb{E}_m Q^{\pi_0}(S', a', m)\pi_0(a'|S') - \mathbb{E}_m Q^{\pi_0}(S, A, m) - \eta^{\pi_0}\}S(\bar{O}_{T-1})\Big].$$

Since (S, A, M, R, S') is any arbitrary transaction tuple follows the corresponding distribution, we have that

$$D_5 = \mathbb{E}\Big[\frac{1}{T}\sum_{t=0}^{T-1}\omega^{\pi_0}(S_t)\frac{\pi_0(A_t|S_t)}{\pi_{b,\theta_0}(A_t|S_t)}\{R_t + \sum_{a'}\mathbb{E}_m Q^{\pi_0}(S_{t+1}, a', m)\pi_0(a'|S_{t+1}) - \mathbb{E}_m Q^{\pi_0}(S_t, A_t, m) - \eta^{\pi_0}\}S(\bar{O}_{T-1})\Big].$$

F.4.1. EFFICIENT FUNCTION

Given C_4, D_4, and D_5, the efficient influence function for $\text{DME}_{\theta_0}(\pi_e, \pi_0)$ is $\eta^{G_0} - \eta^{\pi_0} + I_4 - I_5$, where

$$I_4 = \mathbb{E}\Big[\omega^G(S)\frac{\pi_0(A|S)}{\pi_b(A|S)}\Big[\Big\{R - \mathbb{E}_m r(S, A, m)\Big\} + \rho(S, A, M)\Big\{\mathbb{E}^{\pi_0}_{a',m} r(S, a', m) + \mathbb{E}^G_{a,m} Q^{G_0}(S', a, m) - Q^{G_0}(S, A, M) - \eta^{G_0}\Big\}\Big] + \omega^G(S)\frac{\pi_e(A|S)}{\pi_b(A|S)}\sum_a \pi_0(a|S)\Big\{Q^{G_0}(S, a, M) - \sum_m p(m|S, A)Q^{G_0}(S, a, m)\Big\}\Big].$$

and

$$I_5 = \mathbb{E}\Big[\omega^{\pi_0}(S)\frac{\pi_0(A|S)}{\pi_{b,\theta_0}(A|S)}\{R + \sum_{a'}\mathbb{E}_m Q^{\pi_0}(S',a',m)\pi_0(a'|S') - \mathbb{E}_m Q^{\pi_0}(S,A,m) - \eta^{\pi_0}\}\Big].$$

F.5. Proof of the Equality $\star$

The equality can be proved with the following three steps:

Step 1. We first exchange the summation of t and j in the first line of the equation D_1, which yields that

$$D_1 = \lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\sum_{t=j}^{T-1}\sum_{\tau_t}[r_t - \eta^{\pi_e}]\prod_{k=0}^{t} p_{\theta_0}^{\pi_e}(s_{k+1},r_k,m_k,a_k|s_k) \times \left[\nabla_\theta \log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\right] \times \nu_{\theta_0}(s_0). \quad (45)$$

Step 2. Then we split the summation $\sum_{t=j}^{T-1}$ into $t=j$ and $\sum_{t=j+1}^{T-1}$, and split the product $\prod_{k=0}^{t}$ into $\prod_{k=0}^{j}$ and $\prod_{k=j+1}^{t}$, which leads to

$$D_1 = \lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\Big\{\sum_{\tau_j}[r_j - \eta^{\pi_e}] + \sum_{t=j+1}^{T-1}\sum_{\tau_t}[r_t - \eta^{\pi_e}]\prod_{k=j+1}^{t} p_{\theta_0}^{\pi_e}(s_{k+1},r_k,m_k,a_k|s_k)\Big\} \times \prod_{k=0}^{j} p_{\theta_0}^{\pi_e}(s_{k+1},r_k,m_k,a_k|s_k)\left[\nabla_\theta \log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\right] \times \nu_{\theta_0}(s_0). \quad (46)$$

Step 3. By the definition of Q^{π_e} (see equation (2)), we have that

$$\begin{aligned}\sum_{t=j+1}^{T-1}\sum_{\tau_t}[r_t - \eta^{\pi_e}]\prod_{k=j+1}^{t} p_{\theta_0}^{\pi_e}(s_{k+1},r_k,m_k,a_k|s_k) &= \mathbb{E}_{m_{j+1},a_{j+1}}^{\pi_e} Q^{\pi_e}(s_{j+1},a_{j+1},m_{j+1}) \\ &= \mathbb{E}_{a^*,m^*}^{\pi_e} Q^{\pi_e}(s_{j+1},a^*,m^*).\end{aligned}$$

Substituting this equation, we conclude the proof of $\star$ with that

$$D_1 = \lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T-1}\Big\{\sum_{\tau_j}[r_j - \eta^{\pi_e}] + \mathbb{E}_{a^*,m^*}^{\pi_e} Q^{\pi_e}(s_{j+1},a^*,m^*)\Big\} \times \prod_{k=0}^{j} p_{\theta_0}^{\pi_e}(s_{k+1},r_k,m_k,a_k|s_k)\left[\nabla_\theta \log p_{\theta_0}^{\pi_e}(s_{j+1},r_j,m_j,a_j|s_j)\right]\nu_{\theta_0}(s_0).$$

G. Settings for Numerical Examples

G.1. Toy Example 1 & Toy Example 2

Settings. We consider a scenario with discrete states, actions, mediators, and rewards. We set time $T = 50$, and S_0 for each trajectory is sampled from a Bernoulli distribution with a mean probability of 0.5. Denote the sigmoid function as $expit(\cdot)$. Following the behavior policy, the action $A_t \in \{0,1\}$ is sampled from a Bernoulli distribution, where $\Pr(A_t = 1|S_t) = expit(1.0 - 2.0S_t)$. Observing S_t and A_t, the mediator $M_t \in \{0,1\}$ is drawn from a Bernoulli distribution with $\Pr(M_t = 1|S_t, A_t) = expit(1.0 - 1.5S_t + 2.5A_t)$. The distributions of R_t and S_{t+1} are both Bernoulli and conditional on S_t, A_t, and M_t. Specifically, the reward distribution of $R_t \in \{0,10\}$ satisfies that $\Pr(R_t = 10|S_t, A_t, M_t) = expit(1.0 + 2.0S_t - 1.0A_t - 2.5M_t)$, while the distribution of next state $S_{t+1} \in \{0,1\}$ satisfies that $\Pr(S_{t+1} = 1|S_t, A_t, M_t) = expit(.5 + 3.0S_t - 2.5A_t - .5M_t)$. We are interested in estimating the treatment effect of the target policy π_e, which applies a treatment with $\Pr(A_t = 1|S_t) = expit(1.5 + 1.0S_t)$, compared to the control policy π_0, which always applies no treatment

(i.e., $\Pr(A_t = 1) = 0$). Monte Carlo (MC) simulations are used to calculate the oracle distributions of $\omega^{(\cdot)}$ and $Q^{(\cdot)}$, and the oracle values of $\eta^{(\cdot)}$, IDE(π_e, π_0), IME(π_e, π_0), DDE(π_e, π_0), and DME(π_e, π_0). Based on 40K simulated trajectories with 1K observations each, we obtained that IDE $= -1.277$, IME $= -1.222$, DDE $= -2.982$, and DME $= -.085$. Considering the true distributions of Q^{π_e}, $Q^{\pi_e,a0}$, $Q^{\pi_e,a0*}$, Q^G, and Q^{a_0}, we approximate each of them by assuming linear equation models (Shi et al., 2022a).

Misspecification. To misspecify the ω^{π_e}, we add .25 to the $\omega^{\pi_e}(S_t = 1)$ and subtract .25 from the $\omega^{\pi_e}(S_t = 0)$. Similarly, we subtract .3 from the $\omega^{a_0}(S_t = 1)$ ($\omega^G(S_t = 1)$) and add .3 to the $\omega^{a_0}(S_t = 0)$ ($\omega^G(S_t = 0)$). For $Q^{(\cdot)}$, and r functions, we inject Gaussian noises into each parameter involved in the true model. For the misspecification of p_m and π_b, we multiply the true value by a random variable drawn from a bounded uniform distribution and then clip the probabilities to ensure that they are within the range of .01 and .99.

G.2. Semi-Synthetic Data

The spaces for reward, state, and mediator are continuous, and the action space is binary. Specifically, the semi-synthetic data is generated as follows. The initial states are i.i.d. sampled from the standard normal distribution. π_b follows a Bernoulli distribution, satisfying that $\Pr(A_t = 1) = \Pr(A_t = 0) = .5$. We consider a 2-dimensional mediator, where $M_{t,1}$ and $M_{t,2}$ are independent and normally distributed with a standard deviation of 2. While the mean of $M_{t,1}$ is $\sqrt{|S_t|} + (A_t - .5)$, the mean of $M_{t,2}$ is $.5(A_t - .5) * \sqrt{|S_t|} - .5S_t$. We set $R_t = S_{t+1}$, and R_t is drwan from a normal distribution with a mean of $.75[S_t + \sqrt{|S_t|} + (1 + \sqrt{|M_{t1}| + |M_{t2}|})(A_t - .5)] + 1.5(M_{t1} + M_{t2})$ and a standard deviation of 2. The control policy always takes action $A_t = 0$, while the target policy follows a Bernoulli distribution with $\Pr(A_t = 1) = expit(.7 * S_t)$. Monte Carlo (MC) simulations are used to calculate the oracle the oracle values of $\eta^{(\cdot)}$, IDE(π_e, π_0), IME(π_e, π_0), DDE(π_e, π_0), and DME(π_e, π_0). Based on 20K simulated trajectories with 6400 observations each, we obtained that IDE $= 2.680$, IME $= 3.654$, DDE $= 1.244$, and DME $= .689$.

H. Baseline Estimators

H.1. Baseline_DM

Following the definitions of the direct and indirect effect in Robins & Greenland (1992), the first set of baseline estimators we considered are constructed by inputting estimated probability functions directly. Specifically, the estimator for IDE(π_e, π_0) is

$$\frac{1}{NT} \sum_{i,t,a,m} \Big\{ r(S_{i,t}, a, m) - \sum_{a'} r(S_{i,t}, a', m) \pi_0(a'|S_{i,t}) \Big\} p(m|S_{i,t}, a) \pi_e(a|S_{i,t}),$$

and the estimator for IME(π_e, π_0) is

$$\frac{1}{NT} \sum_{i,t,a,m} r(S_{i,t}, a, m) \Big\{ \sum_{a'} p(m|S_{i,t}, a') \pi_e(a'|S_{i,t}) - p(m|S_{i,t}, a) \Big\} \pi_0(a|S_{i,t}).$$

H.2. Baseline_IPW

The inverse probability weighting estimators, proposed in Lange et al. (2012) and Hong et al. (2010), are the second set of baseline estimators. Specifically, the estimator for IDE(π_e, π_0) is

$$\frac{1}{NT} \sum_{i,t} \frac{\pi_e(A_{i,t}|S_{i,t}) - \pi_0(A_{i,t}|S_{i,t}) \rho(S, A, M)}{\pi_b(A_{i,t}|S_{i,t})} R_{i,t}$$

and the estimator for IME(π_e, π_0) is

$$\frac{1}{NT} \sum_{i,t} \frac{\pi_0(A_{i,t}|S_{i,t})}{\pi_b(A_{i,t}|S_{i,t})} \{\rho(S, A, M) - 1\} R_{i,t}.$$

H.3. Baseline_MR

Finally, the multiply robust estimators, proposed in Tchetgen & Shpitser (2012), are the third set of baseline estimators. Specifically, the estimator for η^{π} is

$$\frac{1}{NT}\sum_{i,t}\frac{\pi(A_{i,t}|S_{i,t})}{\pi_b(A_{i,t}|S_{i,t})}[R_{i,t}-r(S_{i,t},A_{i,t})]+r(S_{i,t},\pi)$$

and the estimator for η^{G_e} is

$$\frac{1}{NT}\sum_{i,t}\Big\{\frac{\pi_0(A_{i,t}|S_{i,t})}{\pi_b(A_{i,t}|S_{i,t})}\rho(S,A,M)[R_{i,t}-r(S_{i,t},M_{i,t},A_{i,t})]+ \\ \frac{\pi_e(A_{i,t}|S_{i,t})}{\pi_b(A_{i,t}|S_{i,t})}[r(S_{i,t},\pi_0,M_{i,t})-\int_m p(m|S_{i,t},A_{i,t})r(S_{i,t},\pi_0,m)]+r(S_{i,t},\pi_0,\pi_e)\Big\},$$

where $r(S_{i,t},\pi_0,\pi_e)=\sum_{a,a',m}r(S_{i,t},a',m)\pi_0(a'|S_{i,t})p(m|S_{i,t},a)\pi_e(a|S_{i,t})$. Plugging in these estimators, we then get the corresponding estimators for IDE(π_e,π_0) and IME(π_e,π_0).

I. Estimating Optimal Policy

To estimate the optimal policy, we first estimate a Q function based on the observational data, following the same methods described in Section 5.4. Specifically, let

$$Q(s,a,m)=\sum_{t\geq 0}\mathbb{E}[R_t-\eta],$$

which leads to a Bellmen equation model, such that

$$Q(S_t,A_t,M_t)=\sum\mathbb{E}[R_t+\sum_a\int_m\mathbb{E}Q(S_{t+1},a,m)-\eta].$$

We then approximate the Q function using linear sieves. Finally, the estimated optimal policy is defined as

$$\hat{\pi}^{opt}(s)=\arg\max_{a\in\mathcal{A}}\sum_a\int_m\mathbb{E}\hat{Q}(s,a,m).$$

It is worth noting that we used cross-validation to estimate the ATE of $\hat{\pi}_{opt}$. To be more specific, we divide the observed trajectories into two folds. In each round k, we first estimate the $\hat{\pi}^k_{opt}$ based on the trajectories within fold k, and then estimate the ATE of $\hat{\pi}^k_{opt}$ on another fold of trajectories.

J. Additional Numerical Experiments

In this section, we conducted additional numerical experiments to evaluate the estimation performance of the proposed estimators under diverse settings.

J.1. Performance Under Tabular Setting

First, we investigate the estimation performance under the tabular setting as discussed in the toy examples (see Section 7.2). Results are summarized in Figure 7. Similar to what we concluded from the semi-synthetic simulation in Section 7.3, all three sets of the proposed estimators provide unbiased estimation for all four effect components, with the MSE decreasing continuously as the sample size increases. In contrast, all baseline estimators ignoring the fact of state transition continue to yield biased estimates no matter how large the sample size is.

J.2. Impact of Variance on Performance

Second, under the semi-synthetic data setting that we used in Section 7.3, we further vary the variances of the random noises of states, mediators, and rewards. Specifically, we considered three settings where the standard deviation of the random

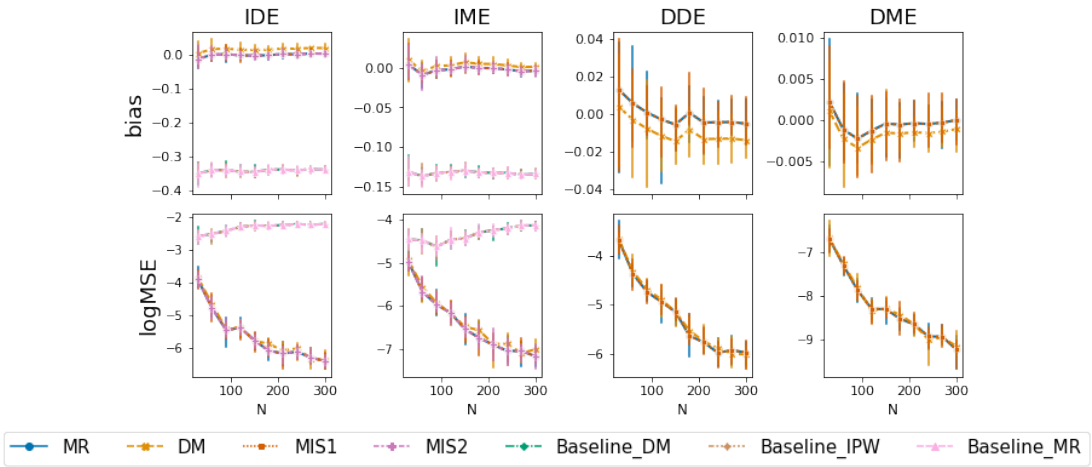

Figure 7. The bias and logMSE of various estimators under the setting with discrete spaces. The results are aggregated over 200 random seeds.

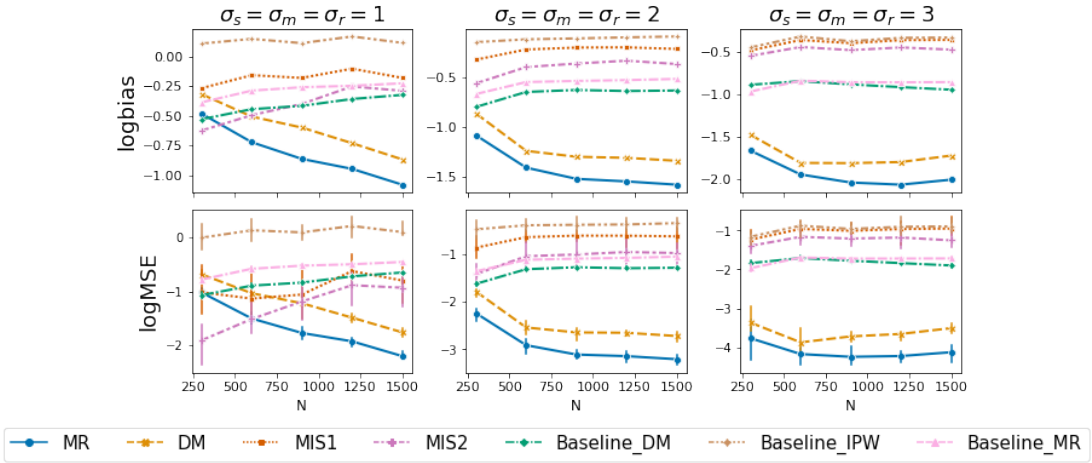

Figure 8. The logBias and the logMSE of estimators for IDE, under different data generation scenarios. Fix T = 50.

noise takes values of 1, 2, and 3, i.e., $\sigma_m = \sigma_s = \sigma_r = 1$, $\sigma_m = \sigma_s = \sigma_r = 2$, and $\sigma_m = \sigma_s = \sigma_r = 3$. Results are summarized in Figure 8, Figure 9, Figure 10, and Figure 11. For each choice of the variance, the proposed MR estimators always achieve the smallest bias and MSE. While all the baseline estimators and MIS estimators continue to provide biased estimates with non-decreasing MSE, we observe that the difference between DM and MR estimators becomes smaller as the variance increases.

J.3. Multidimensional State and Mediator

Finally, we investigate the estimation performance in a more complicated system where both the state and mediator are multi-dimensional. Specifically, we set the number of state variables to be 2 and the number of mediators to be 2. The data

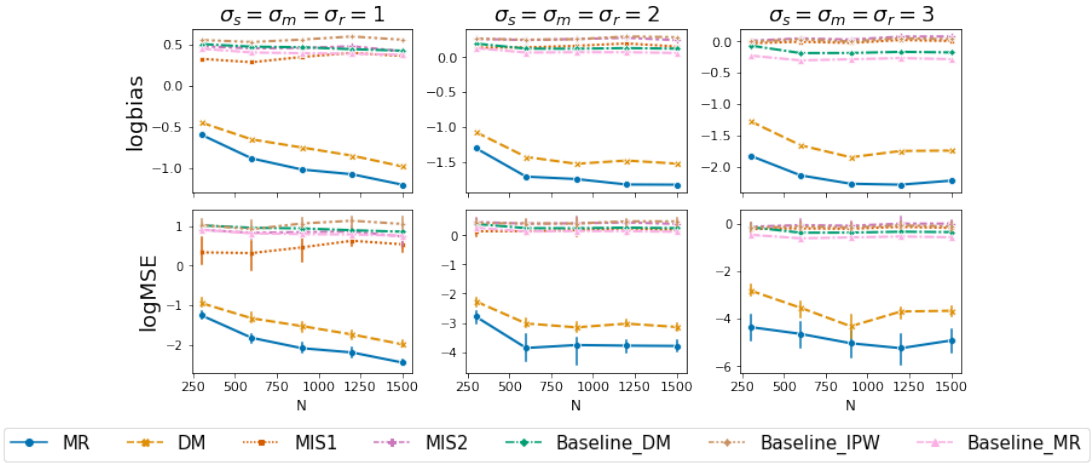

Figure 9. The logBias and the logMSE of estimators for IME, under different data generation scenarios. Fix T = 50.

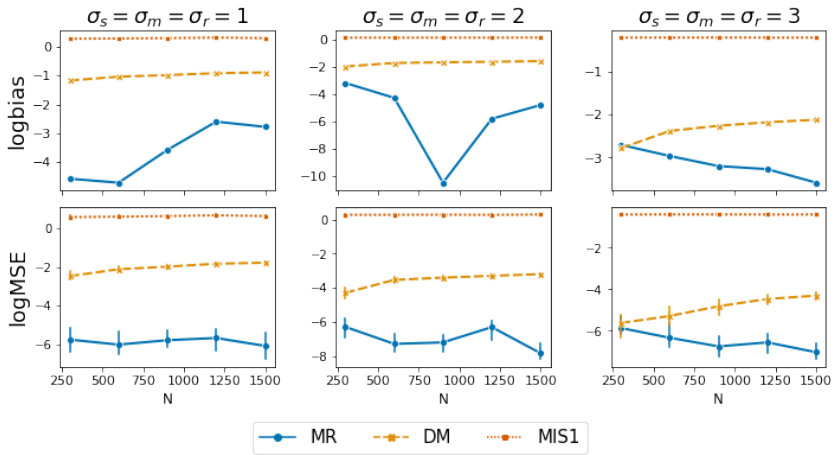

Figure 10. The logBias and the logMSE of estimators for DDE, under different data generation scenarios. Fix T = 50.

generating mechanism is as follows:

$$
\begin{aligned}
S_{0,i} &\sim \mathcal{N}(0,1), \text{ for } i \in \{1,2\};\\
A_t &\sim \pi_b = \text{Bernoulli}(.5);\\
M_{t1} &\sim \mathcal{N}\left(.5\left(\sqrt{|S_{t,1}|}+\sqrt{|S_{t,2}|}\right)+(A_t-.5),1\right);\\
M_{t2} &\sim \mathcal{N}\left(-.25\left(S_{t,1}+S_{t,2}\right)+.25\left(A_t-.5\right)\left(\sqrt{|S_{t,1}|}+\sqrt{|S_{t,2}|}\right),1\right);\\
S_{t+1,1} &\sim \mathcal{N}\left(.75\left\{S_{t,1}+\sqrt{|S_{t,1}|}+\left(1+\sqrt{|M_{t1}|+|M_{t2}|}\right)(A_t-.5)\right\}+1.5\left\{M_{t1}+M_{t2}\right\},1\right);\\
S_{t+1,2} &\sim \mathcal{N}\left(.75\left\{S_{t,2}+\sqrt{|S_{t,2}|}+\left(1+\sqrt{|M_{t1}|+|M_{t2}|}\right)(A_t-.5)\right\}+1.5\left\{M_{t1}+M_{t2}\right\},1\right);\\
R_t &\sim \mathcal{N}\left(.75\left\{.5\left(S_{t,1}+S_{t,2}+\sqrt{|S_{t,1}|}+\sqrt{|S_{t,2}|}\right)+\left(1+\sqrt{|M_{t1}|+|M_{t2}|}\right)(A_t-.5)\right\}+\frac{3}{2}\left\{M_{t1}+M_{t2}\right\},1\right);\\
\pi_e &\sim \text{Bernoulli}\left(\text{expit}\left(.3\left(S_{t,1}+S_{t,2}\right)\right)\right);\\
a_0 &= 0.
\end{aligned}
$$

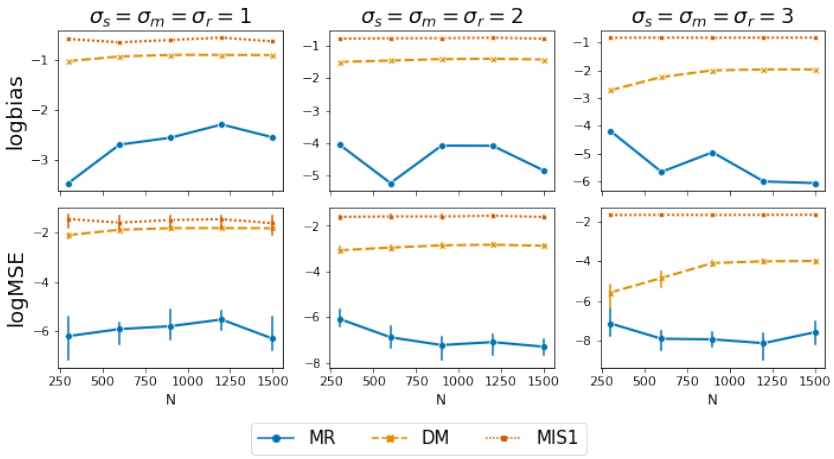

Figure 11. The logBias and the logMSE of estimators for DME, under different data generation scenarios. Fix T = 50.

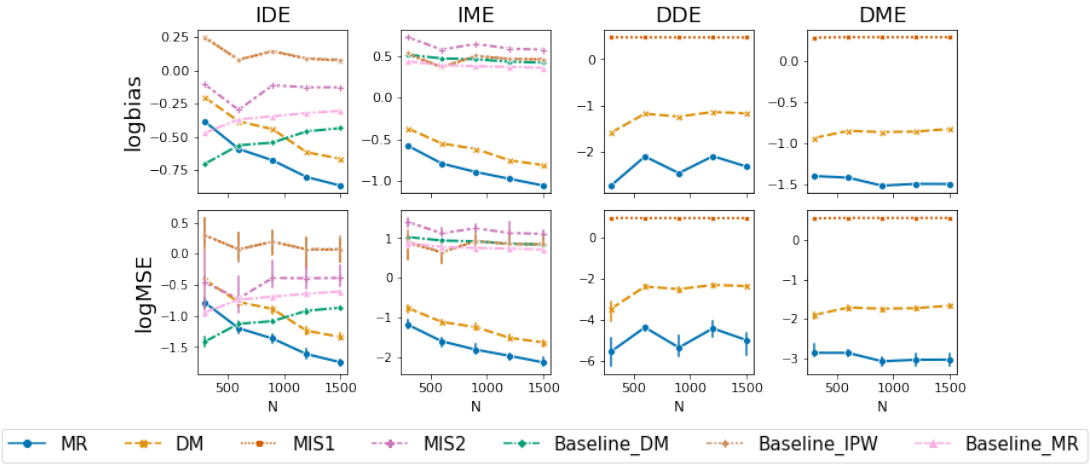

Figure 12. The logBias and the logMSE of estimators, under settings with multidimensional state. Fix T = 100.

The results are summarized in Figure 12 and Figure 13, which show the same trend as the semi-synthetic simulation in Section 7.3.

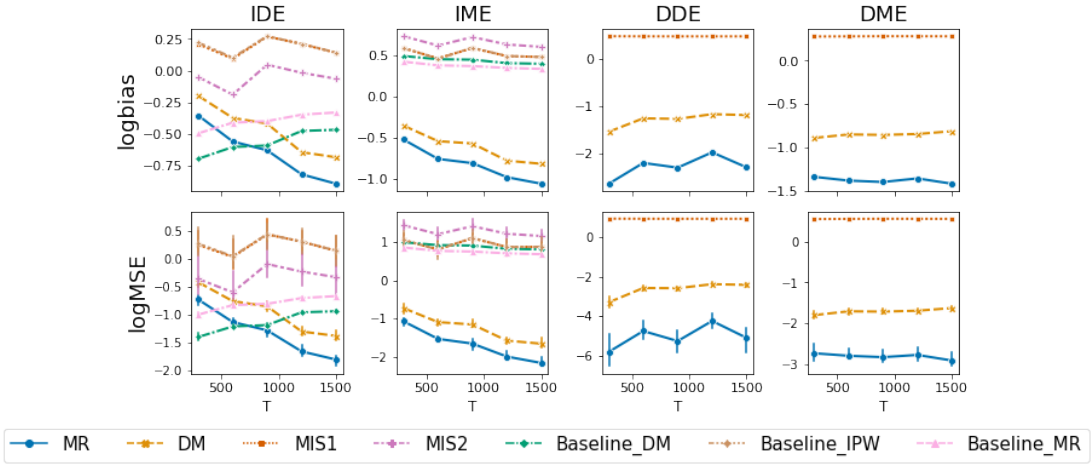

Figure 13. The logBias and the logMSE of estimators, under settings with multidimensional state. Fix N = 100.