JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 545, 273–284: Theory and Methods
https://doi.org/10.1080/01621459.2022.2110878

Off-Policy Confidence Interval Estimation with Confounded Markov Decision Process

Chengchun Shi[a], Jin Zhu[b], Ye Shen[c], Shikai Luo[d], Hongtu Zhu[e], and Rui Song[c]

[a]London School of Economics and Political Science, London, UK; [b]Sun Yat-sen University, Guangzhou, China; [c]North Carolina State University, Raleigh, NC; [d]Didi Chuxing, Peking, China; [e]University of North Carolina at Chapel Hill, Chapel Hill, NC

ABSTRACT

This article is concerned with constructing a confidence interval for a target policy's value offline based on a pre-collected observational data in infinite horizon settings. Most of the existing works assume no unmeasured variables exist that confound the observed actions. This assumption, however, is likely to be violated in real applications such as healthcare and technological industries. In this article, we show that with some auxiliary variables that mediate the effect of actions on the system dynamics, the target policy's value is identifiable in a confounded Markov decision process. Based on this result, we develop an efficient off-policy value estimator that is robust to potential model misspecification and provide rigorous uncertainty quantification. Our method is justified by theoretical results, simulated and real datasets obtained from ridesharing companies. A Python implementation of the proposed procedure is available at *https://github.com/Mamba413/cope*. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2021
Accepted August 2022

KEYWORDS
Infinite horizons; Off-policy evaluation; Reinforcement learning; Ridesourcing platforms; Statistical inference; Unmeasured confounders

1. Introduction

We consider reinforcement learning (RL) where the goal is to learn an optimal policy that maximizes the (discounted) cumulative rewards the decision maker receives (Sutton and Barto 2018). A (stationary) policy is a time-homogeneous decision rule that determines an action based on a set of observed state variables. Off-policy evaluation (OPE) aims to evaluate the impact of a given policy (called target policy) using observational data generated by a potentially different policy (called behavior policy). OPE is an important problem in settings where it is expensive or unethical to directly run an experiment that implements the target policy. This includes applications in precision medicine (Murphy 2003; Zhang et al. 2012, 2013; Chakraborty and Murphy 2014; Matsouaka, Li, and Cai 2014; Luedtke and Van Der Laan 2016; Wang et al. 2018; Gottesman et al. 2019; Wu and Wang 2020), autonomous driving (Li, Chan, and Chen 2020), robotics (Kober, Bagnell, and Peters 2013), natural language processing (Li et al. 2016), education (Mandel et al. 2014), among many others.

This article is concerned with OPE under infinite horizon settings where the number of decision points is not necessarily fixed and is allowed to diverge to infinity. We remark that most works in the statistics literature focused on learning and evaluating treatment decision rules for precision medicine with only a few treatment stages (see Tsiatis et al. 2019; Kosorok and Laber 2019, for an overview). These methods are not directly applicable to many other sequential decision making problems in reinforcement learning with infinite horizons (see e.g., Sutton and Barto 2018), such as autonomous driving, robotics, and mobile health (mHealth). Recently, there is a growing interest on policy learning and evaluation in mHealth applications (Ertefaie 2014; Hu et al. 2020; Luckett et al. 2020; Qi and Liao 2020; Xu et al. 2020; Liao, Qi, and Murphy 2020; Liao, Klasnja, and Murphy 2021; Shi et al. 2021, 2022). In the computer science literature, existing works for OPE in infinite horizons can be roughly divided into three categories. The first type of method directly derives the value estimates by learning the system transition matrix or the Q-function under the target policy (Le, Voloshin, and Yue 2019; Feng et al. 2020; Hao et al. 2021). The second type of method is built upon importance sampling (IS) that re-weights the observed rewards with the density ratio of the target and behavior policies (Thomas, Theocharous, and Ghavamzadeh 2015; Liu et al. 2018; Nachum et al. 2019; Dai et al. 2020). The last type of method combines the first two for more robust and efficient value evaluation. References include Jiang and Li (2016), Uehara, Huang, and Jiang (2020), and Kallus and Uehara (2019). In particular, Kallus and Uehara (2019) develops a double reinforcement learning (DRL) estimator that achieves the semiparametric efficiency limits for OPE. Informally speaking, a semiparametric efficiency bound can be viewed as the nonparametric extension of the Cramer–Rao lower bound in parametric models Bickel et al. (1993). It lower bounds the asymptotic variance among all regular estimators Van der Vaart (2000). However, all the above cited works rely on the sequential ignorability or the sequential randomization assumption (see e.g., Robins 2004, for a detailed definition). It essentially precludes the existence of unmeasured variables that confound the action-reward or action-next-state associations. However, this assumption is likely to be violated in applications such as healthcare and technological industries. We consider the following example to elaborate.

CONTACT Chengchun Shi *C.Shi7@lse.ac.uk* London School of Economics and Political Science, London, UK.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

Our work is motivated by the example of applying customer recommendation program in a ride-hailing platform. We consider evaluating the effects of applying certain customer recommendation program in large-scale ride-hailing platforms such as Uber, Lyft and Didi. These companies form a typical two-sided market which enables efficient interactions between passengers and drivers (Rysman 2009) and substantially transforms the transportation landscape of human beings (Jin et al. 2018).

Suppose a customer launches a ride-hailing application on their smart phone. When they enter their destination, the platform will decide whether to recommend them to join a program. This corresponds to the action. Different programs will apply different coupons to the customer to discount this ride. The purpose of such recommendation is to (i) increase the chance that the customer orders this particular ride, and reduce the local drivers' vacancy periods; (ii) increase the chance that the customer uses the app more frequently in the future. We remark that (i) and (ii) correspond to the short-term and long-term benefits for the company, respectively.

We would like to evaluate the cumulative effect of a given customer recommendation program given an observational dataset collected from the ride-hailing company. In addition to a point estimate on a target policy's value, many applications would benefit from having a confidence interval (CI) that quantifies the uncertainty of the value estimates. For instance, it allows us to infer whether the difference between two policies' values is statistically significant. This motivates us to study the off-policy confidence interval estimation problem.

Confounding is a serious issue in data generated from these applications. This is because the behavior policy involves not only an estimated automated policy to maximize the company's long term rewards but human interventions as well. For example, when there is severe weather like thunderstorms or large events like sports games and concerts in a certain area, there will be much more passengers than drivers in the local area. In that case, human interventions are needed to discourage passengers to request call orders. However, live events and extreme weather are not recorded, leading to a confounded dataset.

More recently, in the causal inference literature, a few methods have been proposed to deal with unmeasured confounders for treatment effects evaluation. Tchetgen Tchetgen et al. (2020) proposed a proximal g-computation algorithm in single-stage and two-stage studies. Shi et al. (2020) proposed to learn the average treatment effect (ATE) with double-negative control adjustment. See also Kallus, Mao, and Uehara (2021). These methods are not directly applicable to the infinite horizon setting, which is the focus of our paper. In the RL literature, a few works considered reinforcement learning with confounded datasets. Among those available, Wang, Yang, and Wang (2020) considered learning an optimal policy in an episodic confounded MDP setting. Namkoong et al. (2020) and Kallus and Zhou (2020) proposed partial identification bounds on the target policy's value under a single-decision confounding assumption and a memoryless unobserved confounding assumption, respectively. Bennett et al. (2021) introduced an optimal balancing algorithm for OPE in a confounded MDP, without requiring the mediators to exist. Tennenholtz, Shalit, and Mannor (2020) adopted the POMDP model to formulate the confounded OPE problem and develop value estimators in tabular settings using the idea of proxy variables. More recently, there are a few works that extend their method to more general settings (Bennett and Kallus 2021; Nair and Jiang 2021; Shi et al. 2022). However, none of the aforementioned methods considered constructing confidence intervals for the target policy's value in infinite horizons.

In this article, we model the observational data by a confounded Markov decision process (CMDP, Zhang and Bareinboim 2016). See Section 2.1 for a detailed description of the model. To handle unmeasured confounders, we make use of some intermediate variables (mediators) that mediate the effect of actions on the system dynamics. These mediators are required to be conditionally independent of the unmeasured confounders given the actions. We remark that these auxiliary variables exist in several applications.

For instance, in the ride-hailing example, the mediator corresponds to the final discount applied to each ride. It is worth mentioning that the final discount might be different from the discount included in the program, as it depends on other promotion strategies the platform applies to the ride, but is conditionally independent of other unmeasured variables that confound the action. In addition, the action will affect the immediate reward and future state variables only through the mediator (see our real data section for a detailed definition of the immediate reward). Consequently, the mediator satisfies the desired condition. Predictive policing is another example. Consider the Crime Incidents dataset (Elzayn et al. 2019). The action is whether a district is labeled as dangerous or not and the outcome is the total number of discovered crime incidents. Given the action, the police allocation (mediator) is determined by the current available policing resources and is thus conditionally independent of the confounder. In addition, in medicine, the treatment (action) and the patient's outcome might be confounded by that patient's attitude toward different treatments. For example, some patients might prefer conservative treatments, and others will strictly stick to the doctor's advice. However, given the treatment, the dosage that patients receive is determined by their age, weight and clinical conditions, and is thus conditionally independent of the confounder.

To the best of our knowledge, this is the first article that systematically studies off-policy confidence interval estimation under infinite horizon settings with unmeasured confounders. Most prior work either requires the unmeasured confounders assumption, or focuses on point estimation. More importantly, our proposal addresses an important practical question in ride-sharing companies, allowing them to evaluate different customer recommendation programs more accurately in the presence of unmeasured confounders. Our proposal involves two key components. We first show that in the presence of mediators, the target policy's value can be represented using the probability distribution that generates the observational data. This result generalizes the front-door adjustment formula (see e.g., Pearl 2009) to infer the average treatment effect in single-stage decision making. Based on this result, we next apply the semiparametric theory (see e.g., Tsiatis 2007) to derive the efficiency limits for OPE under CMDP with mediators, and outline a robust and efficient value estimate that achieves this efficiency bound and its associated CI.

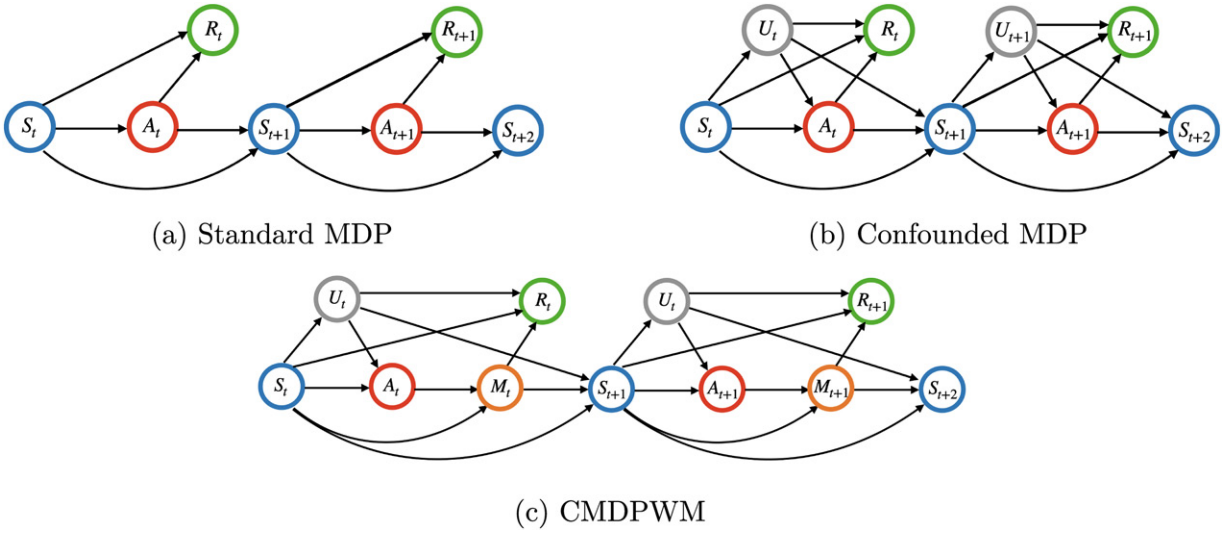

Figure 1. Causal diagrams.

The rest of the article is organized as follows. Section 2 lays out the basic model notation and data-generating process. Section 3 discusses the identifiability of the policy value and construct efficient and robust interval estimation. Section 4 presents the asymptotic properties of the proposed estimator with its inferential results. Section 5 presents two simulation studies to evaluate the performance of our proposed estimator and compare with the state-of-the-art methods by using synthetic data only. In Section 6, an application of the proposed estimator is used to analyze real data collected from a world-leading ride-hailing company. All proofs are given in the supplementary material.

2. Preliminaries

2.1. Data-Generating Process

We consider observational data generated from an confounded Markov decision process. Specifically, at a given time t, let (S_t, A_t, R_t) denote the observed state-action-reward triplet. A standard MDP without confounding is depicted in Figure 1(a). We assume both the state and action spaces are discrete, and the immediate rewards are uniformly bounded. The discrete state-space assumption is imposed only to simplify the theoretical analysis. Our proposal is equally applicable to continuous state space as well. Let U_t denote the set of unmeasured variables at time t that confounds either the A_t-R_t or A_t-S_{t+1} associations, as shown in Figure 1(b). Such a data-generating process excludes the existence of confounders that are influenced by past actions, leading to "memoryless unmeasured confounding" (Kallus and Zhou 2020). It yields the following Markov assumption:

Assumption 1. U_t and other observed variables at time t are conditionally independent of $\{U_j\}_{j<t}$ and past observed variables up to time $t-1$ given S_t.

To deal with unmeasured confounders, we assume there exist some observed immediate variables M_t that mediate the effect of A_t on R_t and S_{t+1} at time t, as shown in Figure 1(c). See Assumption 2. This assumption is similar to the front-door adjustment criterion (Pearl 2009) in single-stage decision making and is considered by Wang, Yang, and Wang (2020) as well for multi-stage decision making.

Assumption 2. (a) M_t intercepts every directed path from A_t to R_t or to S_{t+1};
(b) S_t blocks all backdoor paths from A_t to M_t;
(c) All back-door paths from M_t to R_t or S_{t+1} are blocked by (S_t, A_t).

For any two nodes X and Y, a backdoor path from X to Y is a path that would remain if we were to remove any arrows pointing out of X. We revisit Figure 1(c) to elaborate Assumption 2. Specifically, Assumption 2(a) requires the pathway that A_t has a direct effect on S_{t+1} absent M_t to be missing. Without Assumption 2(a), we can only identify the natural indirect treatment effect (Fulcher et al. 2020) and the policy value is not identifiable. Under Assumptions 2(b) and (c), U_t will not directly affect M_t. Assumption 2(b) essentially requires that there are no unmeasured variables that confound the A_t-M_t association. Similarly, Assumption 2(c) requires that there are no unmeasured variables that confound the M_t-S_{t+1} association.

We next detail the data generating process. At time t, we observe the state vector S_t and the environment randomly selects some unmeasured confounder $U_t \sim p_u(\bullet|S_t)$. Then the agent takes the action $A_t \sim p_a(\bullet|S_t, U_t)$ and the mediator M_t is generated using $p_m(\bullet|A_t, S_t)$ which is not confounded by U_t according to Assumption 2. Finally, the agent receives a reward $R_t \sim p_r(\bullet|M_t, A_t, S_t, U_t)$ and the environment transits into the next state $S_{t+1} \sim p_s(\bullet|M_t, A_t, S_t, U_t)$. We refer to such a stochastic process as the confounded MDP with mediators, or CMDPWM for short.

2.2. Problem Formulation

The data consist of N trajectories, summarized as $\{(S_{i,t}, A_{i,t}, M_{i,t}, R_{i,t}, S_{i,t+1})\}_{1\le i\le N, 0\le t<T_i}$ where T_i corresponds to the termination time of the ith trajectory. We assume these trajectories are iid copies of a CMDPWM model $\{(S_t, A_t, M_t, R_t, S_{t+1})\}_{t\ge 0}$.

Let π denote a given stationary policy that maps the state space to a probability mass function on the action space $\mathcal{A}$. Following π, at each time t, the decision maker will set $A_t = a$ with probability $\pi(a|S_t)$ for any $a \in \mathcal{A}$. Unlike the behavior policy p_a, the probability mass function π does not depend on the unmeasured confounders. For a given discounted factor $0 \leq \gamma < 1$, we define the corresponding (state) value function as

$$V^{\pi}(s) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}^{\pi}(R_t|S_0 = s), \tag{1}$$

where the expectation $\mathbb{E}^{\pi}$ is defined by assuming the system follows the policy π. Based on the observed data, our objective is to learn the aggregated value $\eta^{\pi} = \mathbb{E}\{V^{\pi}(S_0)\}$ where the expectation is taken with respect to the initial state distribution, and to construct its associated confidence interval.

We remark that we adopt a discounted reward formulation to investigate the policy evaluation problem. This formulation allows us to take customers' frequency of using the app into consideration in our application (see Section 6 for details). Meanwhile, our proposal can be easily extended to the average reward setting (see Appendix A.4, supplementary materials).

3. Off-Policy Confidence Interval Estimation

We first discuss the challenge of OPE in the presence of unmeasured confounders. We next show that η^{π} can be represented as a function of the observed dataset. This result implies that η^{π} is identifiable and forms the basis of our proposal. We then outline two potential estimators for η^{π}. Each estimator suffers from some limitations and requires some parts of the model to be correctly specified. This motivates our procedure that combines both estimators for more robust and efficient off-policy evaluation, based upon which a Wald-type CI is derived. Finally, we detail our method.

3.1. The Challenge with Unmeasured Confounders

In this section, we discuss the challenge of OPE with unmeasured confounders. To simplify the presentation, we assume π is a deterministic policy such that $\pi(\bullet|s)$ is a degenerate distribution for any s throughout this section and Section 3.2. For any such policy, we use $\pi(s)$ to denote the action that the agent selects after observing the state vector s. To begin with, we introduce the do-operator do to represent a (hard) intervention (see e.g., Pearl 2009). It amounts to lift A_t from the influence of the old functional mechanism $A_t \sim p_a(\bullet|S_t, U_t)$ and place it under the influence of a new mechanism that sets the value A_t while keeping all other mechanisms unperturbed. For instance, the notation $do(A_t = \pi(S_t))$ means that the action A_t is set to the value $\pi(S_t)$ irrespective of the value of U_t. In other words, whatever relationship exists between U_t and A_t, that relationship is no longer in effect when we perform the intervention. Adopting the do-operator, the expectation $\mathbb{E}^{\pi}$ in (1) can be represented as

$$\mathbb{E}\{R_t|do(A_j = \pi(S_j)), \forall 0 \leq j \leq t, S_0 = s\}. \tag{2}$$

In the presence of unmeasured confounders, the major challenge lies in that η^{π} is defined based on the intervention distribution under the do-operator and cannot be easily approximated via the distribution of the observed data. To elaborate this, we remark that the expectation in (2) is generally not equal to $\mathbb{E}\{R_t|A_j = \pi(S_j), \forall 0 \leq j \leq t, S_0 = s\}$. This is because the distribution under $do(A_t = \pi(S_t))$ is different from that given the observation $A_t = \pi(S_t)$. The latter corresponds to the conditional distribution generated by the causal diagram in Figure 1 given $A_t = \pi(S_t)$, whereas the former is the distribution generated by a slightly different graph, with the pathway $U_t \to A_t$ removed.

As an illustration, we apply DRL and the proposed method to a toy example detailed in Section 5.1. The data are generated according to a CMDPWM model. As we have commented, DRL is proposed by assuming no unmeasured confounders exist. As such, it can be seen from the left panel of Figure 2 that the DRL estimator has a non-diminishing bias under this example, due to the presence of unmeasured confounders. As shown in the right panel of Figure 2, the mean squared error (MSE) of DRL does not decay to zero as the number of trajectories increases to infinity.

In the next section, we address the above mentioned challenge by making use of the auxiliary variables M_t in the observed data. It can be seen from Figure 2 that the proposed estimator is consistent. Both its bias and MSE decay to zero as the number of trajectories diverges to infinity. Finally, we remark that in addition to the use of do-operator, one can adopt the potential outcome framework to formulate the policy evaluation problem (see e.g., Fulcher et al. 2020). We omit the details to save space.

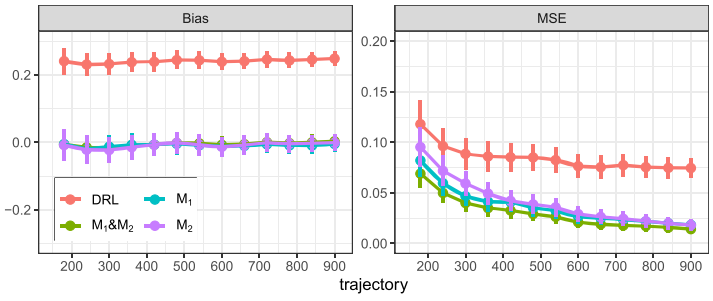

Figure 2. Bias and mean squared error (MSE) of DRL and the proposed estimator under different settings. $T = 100$ and the results are aggregated over 200 simulations. The error bar corresponds to 95% confidence interval for the bias and MSE, from left to right. The proposed estimator requires specification of two sets of models, $\mathcal{M}_1$ and $\mathcal{M}_2$ (see Section 5.1 for details). The green line depicts the estimator where both sets of models are correctly specified. The blue line depicts the estimator where models in $\mathcal{M}_1$ are correctly specified and $\mathcal{M}_2$ misspecified. The purple line depicts the estimator where $\mathcal{M}_2$ is correctly specified and $\mathcal{M}_1$ is misspecified.

3.2. Identification of η^π

In this section, we first show that η^π is identifiable based on the observed data. The main idea is to iteratively apply the Markov property and the front-door adjustment formula to represent the intervention distribution under the do-operator via the observed data distribution.

Recall that p_a is the conditional distribution of $A_t|(S_t, U_t)$. We use p_a^* to denote the corresponding conditional distribution $A_t|S_t$, marginalized over U_t. Similarly, we use $p_{s,r}^*$ to denote the conditional distribution $(S_{t+1}, R_t)|(M_t, A_t, S_t)$. We summarize the results in the following theorem.

Theorem 1. Let τ_t denote the data history $\{(s_j, a_j, m_j, r_j)\}_{0\le j\le t}$ up to time t and ν denote the initial state distribution. Under Assumptions 1 and 2, η^π is equal to

$$\left[\sum_{t=0}^{+\infty}\gamma^t\sum_{\tau_t,s_{t+1}} r_t\left\{\prod_{j=0}^{t} p_{s,r}^*(s_{j+1}, r_j|m_j, a_j, s_j)\right.\right.$$
$$\left.\left. p_m(m_j|\pi(s_j), s_j)p_a^*(a_j|s_j)\right\}\nu(s_0)\right].$$

Note that none of the distributions $p_{s,r}^*, p_m$ and p_a^* involves the unmeasured confounders. As such, these distribution functions can be consistently estimated based on the observational data. Consequently, Theorem 1 implies that η^π can be rewritten using the observed data distribution. Assumption 1 ensures the process satisfies the Markov property. Together with Assumption 2, it allows us to iteratively apply the front-door adjustment formula to replace the intervention distribution with the observed data distribution. See the proof of Theorem 1 in Appendix C.2, supplementary materials for details. We next outline two potential estimators for η^π.

3.3. Direct Estimator

The first estimator is Direct Estimator, where we estimate the Q-function based on the observed data and directly use it to derive the value estimator. In our setting, we define the Q-function $Q^\pi(m, a, s) = \mathbb{E}\{R_t + \gamma V^\pi(S_{t+1})|M_t = m, A_t = a, S_t = s\}$. We make a few remarks. First, our definition of the Q-function is slightly different from that in the existing RL literature, defined by $\mathbb{E}\{R_t + \gamma V^\pi(S_{t+1})|A_t = a, S_t = s\}$, as it involves mediators. Second, similar to Theorem 1, we can show V^π is identifiable from the observed data. It follows that Q^π is identifiable as well.

To motivate the first estimator, we notice that η^π can be rewritten as $\mathbb{E}^\pi\{Q^\pi(M_0, A_0, S_0)\}$, or equivalently, $\sum_a \mathbb{E}[\pi(a|S_0)\mathbb{E}\{Q^\pi(M_0, a, S_0)|do(A_0 = a), S_0\}]$. Applying the front-door adjustment formula, we obtain that

$$\eta^\pi = \sum_{m,a,a',s} p_m(m|a', s)\pi(a'|s)p_a^*(a|s)Q^\pi(m, a, s)\nu(s). \quad (3)$$

This motivates us to learn p_m, p_a^*, Q^π, and ν from the observed data and construct the value estimate by plugging-in these estimators. We refer to this estimator as the direct estimator, since the procedure shares similar spirits as the direct method in the RL literature.

3.4. Importance Sampling Estimator

The second estimator is Importance Sampling (IS) Estimator. This is motivated by the work of Liu et al. (2018) that develops a marginal IS estimator that breaks the curse of horizon, assuming no unmeasured confounders exist. Compared to the standard IS estimator (Zhang et al. 2013) whose variance will grow exponentially fast with respect to the number of decision points, the marginal IS estimator takes the stationary property of the state transitions into consideration and effectively breaks the curse of high variance in sequential decision making. Specifically, let $\omega^\pi(\bullet)$ be the marginal density ratio,

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

where $p_t^\pi(s)$ denotes the probability of $S_t = s$ by assuming the system follows π, and p_∞ denotes the limiting distribution of the stochastic process $\{S_t\}_{t\ge 0}$. Similar to Theorem 1, we can show for any $t > 1$, p_t^π is identifiable. So is ω^π.

A key observation is that, when the stochastic process $\{S_t\}_{t\ge 0}$ is stationary, it follows from the change of measure theorem that $\eta^\pi = (1-\gamma)^{-1}\sum_a \mathbb{E}\{\pi(a|S_t)R_t\omega^\pi(S_t)|do(A_t = a)\}$. When no unmeasured confounders exist, we have $\eta^\pi = (1-\gamma)^{-1}\mathbb{E}\{\pi(A_t|S_t)R_t\omega^\pi(S_t)/p_a^*(A_t, S_t)\}$, yielding the marginal IS estimator. To replace the intervention distribution with the observed data distribution, we apply the importance sampling method again and re-weight each reward by another probability ratio

$$\rho(M_t, A_t, S_t) = \frac{\sum_a \pi(a|S_t)p_m(M_t|a, S_t)}{p_m(M_t|A_t, S_t)}. \quad (4)$$

Such an importance sampling trick has been used by Fulcher et al. (2020) to handle unmeasured confounders in single-stage decision making. This yields the following estimate,

$$\frac{1}{(1-\gamma)\sum_i T_i}\sum_{i,t} R_{i,t}\widehat{\omega}(S_{i,t})\frac{\sum_a \pi(a|S_t)\widehat{p}_m(M_t|a, S_t)}{\widehat{p}_m(M_t|A_t, S_t)},$$

where $\widehat{\omega}$ and $\widehat{p}_m$ denote some estimators for ω^π and p_m.

To conclude this section, we discuss the limitations of the two estimators. First, each estimator requires some parts of the model to be correctly specified. Specifically, the direct estimator requires consistent estimates for Q^π, p_m, and p_a^*, and IS requires correct specification of ω^π and p_m. Second, generally speaking, the direct estimator suffers from a large bias due to potential model misspecification whereas the IS estimator suffers from a large variance due to inverse probability weighting. To address both limitations simultaneously, we develop a robust and efficient OPE procedure by carefully combining the two estimating strategies used in Sections 3.3 and 3.4. Meanwhile, the resulting estimator requires weaker assumptions to achieve consistency. We present the main idea in the next section.

3.5. Our Proposal

We begin with some notations. Let O be a shorthand for a data tuple (S, A, M, R, S'). The key to our estimator is the estimating function, $\psi(O) = \psi_0 + \sum_{j=1}^3 \psi_j(O)$, where ψ_0 is the direct estimator outlined in (3), and $\psi_1(O), \psi_2(O), \psi_3(O)$ are some augmentation terms detailed below. Recall that ψ_0 depends on Q^π, p_m, and p_a^*. The purpose of adding the three augmentation terms is to offer additional protection against potential model

Algorithm 1 Proposed procedure for confounded off-policy confidence interval estimation.

Require: The data $\{(S_{i,t}, A_{i,t}, M_{i,t}, R_{i,t})\}_{i,t}$, and the significance level $0 < \alpha < 1$.

1: Compute the estimators for p_a^* and p_m via supervised learning algorithms. Estimate ν via the empirical initial state distribution.
2: Compute the Q-function and marginal density ratio estimator according to Section 3.6.
3: Plug-in the aforementioned estimated nuisance functions into (5) to construct the value estimator $\widehat{\eta}$.
4: Construct the Wald-type CI.

misspecification of these nuisance functions. As such, the proposed estimator achieves the desired robustness property. See Figure 2 for an illustration. A pseudocode summarizing the proposed algorithm is given in Algorithm 1.

We next present the explicit forms of the three augmentation terms. Specifically, $\psi_1(O)$ equals

$$\frac{1}{1-\gamma}\omega^\pi(S)\rho(M,A,S)\Big\{R+\gamma\sum_{m,a,a^*}Q^\pi(m,a,S')p_m(m,a^*,S')p_a^*(a|S')\pi(a^*|S')-Q^\pi(M,A,S)\Big\},$$

where ρ is the probability ratio defined in (4). The last term in the curly bracket corresponds to the temporal difference error under the CMDPWM model whose conditional mean given (M, A, S) equals zero. As such, $\psi_1(O)$ has zero mean when Q^π, p_m and p_a^* are correctly specified.

$\psi_2(O)$ equals

$$(1-\gamma)^{-1}\omega^\pi(S)\frac{\pi(A|S)}{p_a^*(A|S)}\sum_a p_a^*(a|S)\Big\{Q^\pi(M,a,S)-\sum_m p_m(m|A,S)Q^\pi(m,a,S)\Big\}.$$

When p_m is correctly specified, the last term in the curly bracket can be represented as the residual $Q^\pi(M,a,S) - \mathbb{E}\{Q^\pi(M,a,S)|A,S\}$. As such, $\psi_2(O)$ has zero mean when p_m is correctly specified.

$\psi_3(O)$ equals

$$(1-\gamma)^{-1}\sum_{m,a'}\omega^\pi(S)p_m(m|a',S)\pi(a'|S)\Big\{Q^\pi(m,A,S)-\sum_a Q^\pi(m,a,S)p_a^*(a|S)\Big\}.$$

Similarly, the last term in the curly bracket can be represented as the residual $Q^\pi(m,A,S) - \mathbb{E}\{Q^\pi(m,A,S)|S\}$. When p_a^* is correctly specified, we have $\mathbb{E}\psi_3(O) = 0$.

Based on the estimating function, the proposed estimator takes the following formula,

$$\widehat{\eta}=\frac{1}{\sum_i T_i}\sum_{i=1}^N\sum_{t=0}^{T_i-1}\psi(O_{i,t}), \tag{5}$$

where $O_{i,t} = (S_{i,t}, A_{i,t}, M_{i,t}, R_{i,t}, S_{i,t+1})$. Compared to the standard DRL estimator, the proposed estimator involves additional computations due to the inclusion of the mediator distribution function in the latter two augmentation terms. When there are no unmeasured confounders, the proposed estimator shares similar spirits with DRL.

To construct such an estimator, we need to learn Q^π, ω^π, p_a^*, p_m and the initial state distribution ν. Note that estimating p_a^* or p_m is essentially a regression problem. These functions can be conveniently estimated via existing supervised learning algorithms. We estimate ν via the empirical distribution of $\{S_{i,0}\}_{1\le i\le N}$. As for Q^π and ω^π, we discuss the corresponding estimating procedure later in Section 3.6.

Next, we discuss the relationship between the proposed estimator in (5) and the two estimators outlined in Sections 3.3 and 3.4. Suppose p_m is correctly specified. Let $\mathcal{M}_1$ denote the set of models $\{Q^\pi, p_a^*\}$, and $\mathcal{M}_2$ denote the model ω^π. First, when the models in $\mathcal{M}_1$ are correctly specified, the three augmentations terms have zero mean, as we have discussed earlier. By the weak law of large numbers, (5) is asymptotically equivalent to the direct estimator and is thus consistent. Second, when the models in $\mathcal{M}_2$ are correctly specified, we have $\mathbb{E}\psi_2(O) = 0$. In addition, using similar arguments in Part 3 of the proof of Theorem 2 in the appendix, supplementary materials, we can show that

$$\psi_0+\mathbb{E}\psi_3(O)=\frac{1}{1-\gamma}\mathbb{E}\left[\omega^\pi(S)\rho(M,A,S)\{Q^\pi(M,A,S)-\gamma V^\pi(S')\}\right].$$

By the definition of ψ_1, this in turn implies that (5) is unbiased to the IS estimator. It is thus consistent. The above discussion informally justifies the robustness property of (5). We will rigorously prove the claim in Theorem 2.

Finally, we observe that the proposed estimator can be written as $N^{-1}\sum_{i=1}^N \eta_i$ where η_i denotes the estimating function based on the ith trajectory only. Since the trajectories are independent, the proposed estimator is asymptotically normal, as shown in Theorem 3. A Wald-type CI $[\widehat{\eta} - z_{\alpha/2}N^{-1/2}\widehat{\sigma}_\eta, \widehat{\eta} + z_{\alpha/2}N^{-1/2}\widehat{\sigma}_\eta]$ is valid for off-policy interval estimation, where z_α denotes the upper αth quantile of a standard normal distribution and $\widehat{\sigma}_\eta^2$ denotes the sampling variance estimator of $\{\eta_i\}_i$.

3.6. Learning Q^π and ω^π

The estimating procedure for Q^π is motivated by the following Bellman equation,

$$\mathbb{E}\Big\{R+\gamma\sum_{m,a,a^*}Q^\pi(m,a,S')p_m(m,a^*,S')p_a^*(a|S')\pi(a^*|S')\,\Big|\,M,A,S\Big\}=Q^\pi(M,A,S).$$

Similar to the standard Bellman equation under settings without unmeasured confounders, it decomposes the Q-function into two parts, the immediate reward plus the discounted future state-action values.

To prove this identity, notice that similar to (3), we can show that

$$V^\pi(s)=\sum_{m,a,a^*}Q^\pi(m,a,s)p_m(m,a^*,s)p_a^*(a|s)\pi(a^*|s),$$

based on the front-door adjustment formula. This together with our definition of the Q-function yields the above Bellman equation.

Let $\widehat{p}_m$ and $\widehat{p}_a^*$ denote consistent estimators for p_m and p_a, based on the observed data. Given the Bellman equation, multiple methods can be applied to estimate Q^π. We employ the fitted Q-evaluation method Le, Voloshin, and Yue (2019) in our setup and propose to iteratively compute $\widehat{Q}^{\ell+1}$ by solving

$$\arg\min_{Q\in\mathcal{Q}} \sum_{i,t} \left\{R_{i,t} - Q(M_{i,t}, A_{i,t}, S_{i,t}) + \gamma \widehat{V}^{\ell}(S_{i,t+1})\right\}^2,$$

where $\widehat{V}^{\ell}(S_{i,t+1}) = \sum_{m,a,a^*} \widehat{Q}^{\ell}(m, a, S_{i,t+1})\widehat{p}_m(m, a^*, S_{i,t+1})\widehat{p}_a^*(a|S_{i,t+1})\pi(a^*|S_{i,t+1})$, for some function class $\mathcal{Q}$ and $\ell = 0, 1, \ldots$, until convergence. Similar to Fan et al. (2020), we can show that the resulting Q-estimator is consistent when $\mathcal{Q}$ is a class of universal function approximators such as neural networks.

We next consider ω^π. Similar to the work of Liu et al. (2018), we can show that when the process $\{S_t\}_{t\geq 0}$ is stationary, ω^π satisfies the equation $L(\omega^\pi, f) = 0$ for any discriminator function f in our setup, where $L(\omega^\pi, f)$ is given by

$$\mathbb{E}\omega^\pi(S_{i,t})\left\{f(S_{i,t}) - \gamma \frac{\sum_{a\sim\pi(\bullet|S_{i,t})} p_m(M_{i,t}|a, S_{i,t})}{p_m(M_{i,t}|A_{i,t}, S_{i,t})} f(S_{i,t+1})\right\} - (1-\gamma)\sum_s f(s)\nu(s). \quad (6)$$

As such, ω^π can be learned by solving the following mini-max problem,

$$arg\,min_{\omega\in\Omega} \sup_{f\in\mathcal{F}} L^2(\omega, f), \quad (7)$$

for some function classes Ω and $\mathcal{F}$. The expectation in (6) can be approximated by the sample average. p_m and ν in (6) can be substituted with their estimators. As pointed out by one of the referees, the minimax optimization is often not stable. To address this issue, we restrict attention to linear or kernel function classes to simply the calculation. See Appendix B, supplementary materials for details.

4. Statistical Guarantees

We prove the robustness and efficiency of our estimator as well as the validity of our CI in this section. Without loss of generality, we assume $T_i = T$ for any i. To derive the asymptotic theories, we require the number of trajectories N to diverge to infinity. The termination time T can either be bounded, or diverge with N. The assumption on N is imposed to ensure that the initial state distribution ν can be well-approximated by the empirical distribution of $\{S_{i,0}\}_{1\leq i\leq N}$. We first introduce some conditions. Let $\mathcal{H}_m$ and $\mathcal{H}_a$ be the function classes used to model p_m and p_a, respectively.

Assumption 2. The function classes $\mathcal{Q}$, Ω, $\mathcal{H}_m$ and $\mathcal{H}_a$ are bounded and belong to VC type classes (Definition 2.1, Chernozhukov, Chetverikov, and Kato 2014) with VC indices upper bounded by $\nu = O(N^\kappa)$ for some $0 \leq \kappa < 1/2$.

Assumption 2 is mild as the function classes are user-specified. VC type classes contains a wide variety of functional classes, including neural networks and regression trees. The VC index controls the model complexity. It generally increases with the number of parameters in the model. We allow the VC index to diverge with the sample size to reduce the bias of the estimator due to model misspecification.

Theorem 2 (Robustness). Suppose the process $\{S_t\}_{t\geq 0}$ is stationary, $p_m(M_{i,t}|A_{i,t}, S_{i,t})$, $p_a(A_{i,t}|S_{i,t})$, $\widehat{p}_m(M_{i,t}|A_{i,t}, S_{i,t})$, $\widehat{p}_a(A_{i,t}|S_{i,t})$ and $p_\infty(A_{i,t}|S_{i,t})$ are uniformly bounded away from zero, Assumptions 1, 2 hold, and $\widehat{p}_m$ is consistent. Then as $N \to \infty$, the proposed estimator is consistent when either $\widehat{Q}$, $\widehat{p}_a^*$ or $\widehat{\omega}$ converges in L_2-norm to their oracle values.

To save space, we present the detailed definition of L_2-norm convergence in Appendix C.1, supplementary materials. Theorem 2 formally establishes the robustness property. Notice that the proposed estimator equals the direct estimator outlined in Section 3.3 when $\omega^\pi = 0$ and equals the IS estimator in Section 3.4 when $\widehat{Q} = 0$. As a byproduct, we obtain the following corollary.

Corollary 1. (i) Suppose the conditions in Theorem 2 hold. Suppose $\widehat{Q}$ and $\widehat{p}_a^*$ converge in L_2-norm to their oracle values. Then the direct estimator is consistent as $N \to \infty$. (ii) Suppose $\widehat{\omega}$ converges in L_2-norm to their oracle value. Then the IS estimator is consistent as $N \to \infty$.

To achieve efficiency, we need the following assumption:

Assumption 3. Suppose $\widehat{Q}$, $\widehat{p}_a^*$, $\widehat{p}_m$, $\widehat{\omega}$ converge in L_2-norm to their oracle values at a rate of $N^{-\kappa^*}$ for some $\kappa^* > 1/4$.

Assumption 3 characterizes the theoretical requirements on the nuisance function estimators. Suppose some parametric models (e.g., linear) are imposed to learn these nuisance functions. When the models are correctly specified, then we have $\kappa^* = 1/2$ (Uehara et al. 2021). Here, we do not impose parametric assumptions and only require $\kappa^* > 1/4$. For instance, when using kernels or neural networks for function approximation, the corresponding convergence rates of $\widehat{Q}$ and $\widehat{\omega}$ are provided in Fan et al. (2020) and Liao, Qi, and Murphy (2020). $\widehat{p}_a^*$ and $\widehat{p}_m$ can be computed via standard supervised learning algorithms. Their rates of convergence are available for most often used machine learning approaches including random forests (Wager and Athey 2018) and deep learning (Schmidt-Hieber 2020).

Theorem 3 (Efficiency). Suppose the conditions in Theorem 2 hold and Assumption 3 holds. Then the proposed estimator achieves the semiparametric efficiency bound.

We make a few remarks. First, we show in the proof of Theorem 3 that the proposed estimator is asymptotically normal and satisfies $\sqrt{N}(\widehat{\eta} - \eta^\pi) \xrightarrow{d} N(0, \sigma_T^2)$ where the explicit form of σ_T^2 is detailed in Appendix C.4, supplementary materials. The asymptotic variance estimator for σ_T^2 can be constructed via the sampling-variance formula. Consequently, a two-sided Wald-type confidence interval (CI) can be derived for η^π. Second, the asymptotic variance σ_T^2 decays with T. Specifically, it can be decomposed into $\sigma_0^2 + T^{-1}\sigma_*^2$ for some σ_0, σ_*. See Appendix C.4, supplementary materials for the explicit forms of these quantities. The first term σ_0^2 accounts for the variation of the initial state distribution in the plug-in estimator. The second

term $T^{-1}\sigma_*^2$ is the variance of the augmentation terms and decays to zero as $T \to \infty$. Third, Kallus and Uehara (2019) derives the efficiency bound for OPE in infinite horizon settings where no unmeasured confounders exist and the initial state distribution under the target policy is known. Our proof for Theorem 3 differs from theirs in that we allow the initial state distribution to be unknown and allow unmeasured confounders to exist. Fourth, in Assumption 3, we require all the nuisance function estimators to converge to their oracle values and thus exclude the case with model misspecification. When the model is misspecified, the semiparametric efficiency bound cannot be achieved. Finally, in our proposal, we use the same dataset twice to estimate the nuisance functions and construct the final value estimator. We do not use cross-fitting. Because of that, we impose certain metric entropy conditions in Assumption 2 to establish the robustness and efficiency of the proposed value estimator. To remove Assumption 2, we can couple our procedure with sample-splitting and cross-fitting (see e.g., Chernozhukov et al. 2018; Kallus and Uehara 2019). However, in our setup, we find that the proposed estimator without cross-fitting has better finite sample properties.

Theorem 4 (*Validity*). Suppose the conditions in Theorem 3 hold. Then the coverage probability of the proposed CI approaches to the nominal level as N diverges to infinity.

We remark that Theorems 3 and 4 are concerned with the asymptotic distribution of the value estimator under a single target policy. In Appendix A.3 of the supplementary materials, we establish the joint asymptotic distribution of the proposed value estimators under multiple target policies, introduce the proposed CI for the value difference between two target policies and prove its validity.

5. Simulation Studies

In this section, we evaluate the finite sample performance of the proposed estimator using two simulation studies. The first toy example aims to illustrate the robustness properties of our estimator to unmeasured confounding and model misspecification. In the second simulation study, we demonstrate that our method is superior to state-of-the-art policy evaluation methods.

5.1. A Toy Example

We first describe the detailed setting for the toy example. We fix time $T = 100$ and the initial state is sampled from a Bernoulli distribution with support $\{0, 1\}$ and satisfies that $\mathbb{P}(S_0 = 1) = \mathbb{P}(S_0 = 0) = 0.5$. The unmeasured confounders $\{U_t\}_{t=1}^T$ are iid sampled from a Bernoulli distribution with support $\{-1, 1\}$ and satisfy that $\mathbb{P}(U_t = 1) = \mathbb{P}(U_t = -1) = 0.5$. The action is discrete-valued and the behavior policy p_a satisfies that $p_a(1|S_t, U_t) = p_a(-1|S_t, U_t) = 0.5\text{sigmoid}(0.1S_t + 0.9U_t)$, and $p_a(0|S_t, U_t) = 1 - \text{sigmoid}(0.1S_t + 0.9U_t)$. The mediator is drawn from a Bernoulli distribution with binary support. We set $p_m(1|A_t, S_t) = \text{sigmoid}(0.1S_t - 0.9(A_t - 0.5))$ which does not depend on U_t. Assumption 1 is thus satisfied. The reward R_t and the next-state S_{t+1} are Bernoulli random variables with support $\{0, 10\}$ and $\{0, 1\}$, respectively, and satisfy $\mathbb{P}(R_t = 10|S_t, U_t, M_t) = \mathbb{P}(S_{t+1} = 1|S_t, U_t, M_t) = \text{sigmoid}(0.5I(U_t = 1)(S_t + M_t) - 0.1S_t)$. We are interested in evaluating a random policy that outputs 0 with probability $1 - \text{sigmoid}(0.3S_t)$, and outputs -1 or 1 with probability $0.5\text{sigmoid}(0.3S_t)$ after observing S_t. Under this toy example, we are able to derive Q^π, p_m, p_a^*, and η^π theoretically, and we calculate the true value of ω^π via Monte Carlo method.

Recall that $\mathcal{M}_1$ is a combination of Q^π, p_a^* and $\mathcal{M}_2 = \{\omega^\pi\}$. We evaluated the performance of the proposed estimator under the following scenarios: (i) all the models p_m, $\mathcal{M}_1$, $\mathcal{M}_2$ are correct; (ii) p_m and $\mathcal{M}_1$ are correct, $\mathcal{M}_2$ is misspecified; (iii) p_m and $\mathcal{M}_2$ are correct, $\mathcal{M}_1$ is misspecified. Specifically, to misspecify Q^π, we inject a Gaussian noise with unit variance to the true Q^π. To misspecify p_a^*, we multiply p_a^* by a variable sampled from a uniform distribution with lower boundary 0.75 and higher boundary 1. To misspecify ω^π, we increase the value of $\omega^\pi(0)$ by 0.5 and reduce the value of $\omega^\pi(1)$ by 0.5. As shown in Figure 2, our proposed estimator is robust to unmeasured confounding and model misspecification.

5.2. Comparison with State-of-the-Art Methods

We compare the proposed method with the state-of-the-art methods in the existing reinforcement learning literature. The simulated data are generated as follows. The initial state is sampled from a standard normal distribution with dimension $d_S = 1$ or 3. The distributions of unmeasured confounders are the same as those in the toy example. Let 1_t be a length t vector with values 1, and $C_t = 1_{d_S}^\top S_t$, the sum of the state. The action is binary-valued and is generated according to the behavior policy $p_a(1|S_t, U_t) = \text{sigmoid}(0.1C_t + 0.9U_t)$. The mediator is drawn from a Bernoulli distribution with binary support. We set $p_m(1|A_t, S_t) = \text{sigmoid}(0.1C_t + 0.9(A_t - 0.5))$ which does not depend on U_t. Assumption 1 is thus satisfied. The reward R_t is sampled from a normal distribution with conditional mean $0.5I(U_t = 1)(M_t + C_t) - 0.1C_t$ and standard deviation 0.1. The future state S_{t+1} is sampled from a multivariate normal distribution with mean $0.5I(U_t = 1)(M_t 1_{d_S} + S_t) - 0.1S_t$ and covariance matrix $0.25I_{d_S}$. The target policy selects action 1 with probability $\text{sigmoid}(0.3C_t)$.

We compare the proposed estimator with three types of baseline methods. All these methods are developed by assuming no unmeasured confounders. The first one is the direct estimator, computed based on an estimated Q-function, $\widehat{\eta}_{\text{REG}} = N^{-1}\sum_{i=1}^N \widehat{Q}(S_{i,0}, \pi(S_{i,0}))$ (denoted by REG). In our implementation, we compute $\widehat{Q}$ via the fitted Q-evaluation algorithm. The second one is the marginal importance sampling (MIS) estimator (Liu et al. 2018). To implement this method, the marginal sampling ratio is estimated by assuming no unmeasured confounders exist and is different from the proposed estimator for ω^π. The third one is the DRL estimator that combines the first two estimators for value evaluation. None of these methods uses the mediator. For fair comparison, we also include the mediator in the state to construct the value estimates. Denote the resulting three estimators by REG-M, MIS-M, and DRL-M, respectively. We further estimate their variances based on the sampling variance formula (see Appendix B, supplementary

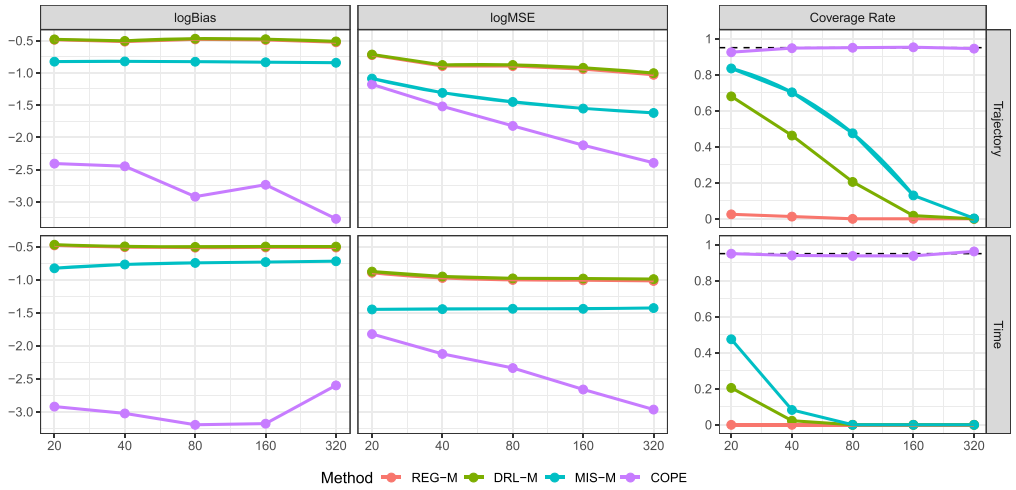

Figure 3. Logarithms of the bias and mean square error (MSE), and 95% CI's coverage rate of various OPE methods with different combinations of N and T when S_t is three-dimensional. The black dash line corresponds to the confidence level 95%. Top panels: T is fixed to 20 and $N \in \{20, 40, 80, 160, 320\}$, bottom panels: N is fixed to 20 and $T \in \{20, 40, 80, 160, 320\}$.

materials for details) and construct the associated confidence interval. The proposed estimator is denoted by COPE, short for confounded off-policy interval estimation.

The linear basis function models $\widehat{p}_a^*$, $\widehat{p}_m$, $\widehat{Q}^\pi$, $\widehat{\omega}^\pi$ employ randomly generated Fourier features based on the Python `RBFsampler` function. We find that the performance of the value estimator is not overly sensitive to the number of basis functions (see Appendix B, supplementary materials for details). Let η^π be the ground truth and $\widehat{\eta}^\pi$ be a given OPE estimator, we define logBias as $\log_{10}(|\mathbb{E}\widehat{\eta}^\pi - \eta^\pi|)$ and logMSE as $\log_{10}\{\mathbb{E}(\widehat{\eta}^\pi - \eta^\pi)^2\}$, where the expectation $\mathbb{E}(\cdot)$ is approximated by Monte Carlo simulations. We report these metrics, as well as the empirical coverage probabilities of all the confidence intervals for the target policy's value in Figures 3 and S3 (see Appendix B in the supplementary materials). We also calculate the standard deviation of these metrics in 400 replications and report them in Appendix B, supplementary materials.

It can be seen that COPE achieves the least bias and MSE among all methods. In addition, its MSE decays with N and T in general and the empirical coverage rate of our CI is close to the nominal level. We also notice that the squared bias of our estimator is much smaller than its MSE. This demonstrates the consistency of our method and is in line with our theoretical findings. In contrast, other baseline estimators are severely biased, since they cannot handle the unmeasured confounders.

6. Real Data Application

In this section, we apply our method to a real dataset from a world-leading ride-hailing company. We focus on a particular recommendation program applied to customers in regions where there are more taxi drivers than the call orders. As we have commented, in the short term, this helps balance the taxi supply and passenger demand across different areas of the city. In the long term, this increases the frequency that the customer uses the app to request the trip.

The dataset consists of all the call orders at a given a city from September 16th to September 22th. The features available to us consist of each order's time, origin, destination and a supply-demand equilibrium metric that characterizes the degree that supply meets the demand. For each of the call order, the customer might receive a coupon for 20% off. This yields a binary action. The mediator is the actual discount applied to the order. As we have commented, the mediator is calculated by the platform using the action and other promotion strategies and differs from the action, but is conditionally independent of those unmeasured variables that confound the action. Assumption 2(b) is thus satisfied. The reward is zero if the customer does not request the ride at the end, and one minus the actual discount times the price of the order otherwise. We present the empirical quantiles of the reward and state in Table S4.

By definition, the reward depends on the action only through its effect on the mediator. In addition, the customer observes the final discount applied to their ride on the application, but is not aware of which promotion strategy yields the discount. As such, it is reasonable to assume that these promotion strategies will affect their behaviors through the final discount only. Consequently, the reward and future state are conditionally independent of the action and other promotion strategies given the mediator. Assumptions 2(a) and (c) thus hold.

We first fit a MDP model to this dataset, and use the estimated MDP to generate synthetic data to mimic the real dataset. Specifically, the distribution of initial state is approximated by a multivariate normal distribution. The state transition $S_{t+1}|A_t, S_t, M_t$ is modeled by a multivariate normal distribution $N(\mu(S_t, A_t, M_t), \sigma^2)$ where the conditional mean function μ is estimated using regularized linear basis function models. Similarly, we estimate the reward function $\mathbb{E}(R_t|S_t, A_t, M_t)$ using a regularized linear basis function model as well. All the tuning parameters are selected by 5-fold cross-validation. Based on the fitted MDP, synthetic dataset can be generated to evaluate different OPE methods.

We are interested in evaluating two recommendation policies. One of them is a random policy (denote by π_1), with which each customer would have an equal chance to get a 20% discount with probability 0.5. Another policy (denote by π_2) relies on the imbalance measure between supply and demand. Specifically,

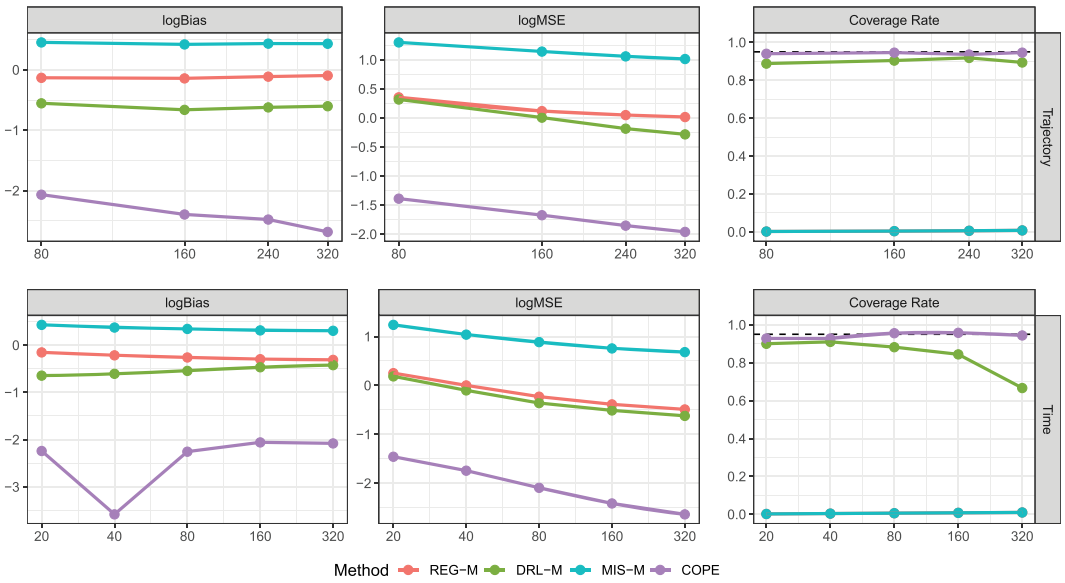

Figure 4. Logarithms of the bias and MSE, and 95% CI's coverage rate of various OPE methods with different combinations of N and T in simulated real data environment. The black dash line in the right most panel corresponds to the confidence level 95%. Top panels: T is fixed to 20 and $N \in \{80, 160, 240, 320\}$, bottom panels: N is fixed to 100 and $T \in \{20, 40, 80, 160, 320\}$.

for the region with extremely more vacant drivers than requests, we randomly choose 70% of the customers for getting the discount. For the rest of regions, the customers would have a 30% chance for obtaining the discount. We expect the second policy would yield larger values, as it has better immediate reward and encourages customers to request rides more often.

To take customers' frequency of using the app into consideration, we use a slightly different definition of the value function as in Xu et al. (2018), and adjust the proposed method and other baselines accordingly to reflect this change. In addition to $\{(S_{i,t}, A_{i,t}, M_{i,t}, R_{i,t})\}_{i,t}$, the observed data consist of another sequence of variables $\{T_{i,t}\}$, corresponding to the time that the ith customer launches the app and enters the destination. We initialize $T_{i,0}$ to zero, for all i. The target policy's value is defined as $\eta^{\pi} = \sum_{t=0}^{+\infty} \mathbb{E}^{\pi}(\gamma^{T_{i,t}} R_{i,t})$. To reflect this change, the proposed estimator takes the following form,

$$\widehat{\eta} = \psi(O) = \psi_0 + \frac{1}{NT} \sum_{j=1}^{3} \sum_{i,t} \psi_j'(O_{i,t}),$$

where ψ_0 is the same as the direct estimator with the Q-estimator $\widehat{Q}$ replaced by $\widehat{Q}'$ detailed below, and for any j, $\psi_j'(O_{i,t})$ is a version of $\psi_j(O_{i,t})$ with γ replaced by $\gamma^{T_{i,t+1}-T_{i,t}}$, $\widehat{Q}$ replaced by $\widehat{Q}'$ and $\widehat{\omega}$ replaced by $\widehat{\omega}'$. Specifically, $\widehat{Q}'$ is computed by solving a slightly different Bellman equation

$$\mathbb{E}\Big\{R_{i,t} + \gamma^{T_{i,t+1}-T_{i,t}} \sum_{m,a',a} p_m(m|a', S_{i,t+1}) p_a^*(a|S_{i,t+1}) \pi(a'|S_{i,t+1})$$
$$Q^{\pi}(m, a, S_{i,t+1}) - Q^{\pi}(M_{i,t}, A_{i,t}, S_{i,t}) \mid M_{i,t}, A_{i,t}, S_{i,t}\Big\} = 0,$$

and $\widehat{\omega}'$ is computed by solving (7) with γ replaced by $\gamma^{T_{i,t+1}-T_{i,t}}$ in (6). The DRL estimator can be similarly modified to adapt to this change.

We apply REG-M, MIS-M, DRL-M and the proposed method COPE to evaluate the value difference $\eta^{\pi_2} - \eta^{\pi_1}$. The ground truth OPE is approximated via Monte Carlo based on the fitted MDP model and equals 0.17. This is consistent with our expectation that the second policy yields a larger value. We evaluate the estimation accuracy by logBias and logMSE as in the simulation section, and the coverage probability of the confidence interval. See Appendix A.3 in the supplementary materials for the construction of the confidence interval of the value difference. The simulation results are aggregated over 500 replications. The discounted factor γ is set to 0.99, as we are interested in the long-term treatment effects.

Figure 4 depicts the performance of four methods. Results are summarized as follows. First, COPE has the best estimation accuracy among the four methods. Second, the coverage probability of the proposed CI is close to the nominal level. In contrast, the baseline methods fail to achieve the nominal coverage when N or T is large. These results are consistent with our simulation findings.

We next apply our method and DRL to the real dataset to evaluate the value difference $\eta^{\pi_2} - \eta^{\pi_1}$. The proposed method yields a value difference of 0.63. The 95% associated confidence interval is [0.03, 1.23]. As such, the second policy is significantly better than the first one. The result is consistent with our expectation. On the contrary, DRL yields a value difference of -0.96. The associated confidence interval is $[-2.07, 0.14]$. According to DRL, the random policy is much better. This is due to that DRL cannot handle unmeasured confounders, leading to a biased estimator. Combining this with our theoretical and simulations results, we have more confidence about the findings of our proposed CI.

7. Discussion

In this section, we discuss several extensions. First, our current proposal relies on the "memoryless unmeasured confounding" assumption to simplify the derivation. In Appendix A.1 of the supplementary materials, we discuss several possible relaxations of this assumption. Second, we assume the mediators

are discrete to simplify the presentation. In Appendix A.2, supplementary materials, we extend the proposed method to settings with continuous mediators. Third, we adopt a discounted reward formulation to investigate the policy evaluation problem. We extend our proposal to the average reward setting in Appendix A.4 of the supplementary materials. Fourth, we assume the mediator variable is completely observed. In the causal inference literature, Chernofsky, Bosch, and Lok (2021) considered settings with left censored mediators. They proposed three estimation methods, including (i) mediator model extrapolation; (ii) numerical integration and optimization of the observed data likelihood function; (iii) the Monte Carlo Expectation-Maximization algorithm. In cases with partially observed mediators, we can couple their ideas with our proposal for value evaluation.

Supplementary Materials

The supplementary article consists of some further discussions of the memoryless unmeasured confounding assumption, extensions of the proposal to settings with the average reward objective, continuous mediators and multiple target policies, as well as some implementation details, technical definitions and proofs.

Acknowledgments

The authors wish to thank the AE, and two anonymous reviewers for their constructive comments, which have led to a significant improvement of the proposed methodology.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

CS' research is partly supported by the EPSRC grant EP/W014971/1. RS' research is partly supported by the NSF grants DMS-1555244 and DMS-2113637.

ORCID

Rui Song https://orcid.org/0000-0003-1875-2115

References

Bennett, A., and Kallus, N. (2021), "Proximal Reinforcement Learning: Efficient Off-Policy Evaluation in Partially Observed Markov Decision Processes," arXiv:2110.15332. [274]

Bennett, A., Kallus, N., Li, L., and Mousavi, A. (2021), "Off-policy Evaluation in Infinite-Horizon Reinforcement Learning with Latent Confounders," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (Vol. 130), eds. A. Banerjee and K. Fukumizu, pp. 1999–2007, PMLR. [274]

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993), *Efficient and Adaptive Estimation for Semiparametric Models* (Vol. 4), Baltimore, MD: Johns Hopkins University Press. [273]

Chakraborty, B., and Murphy, S. A. (2014), "Dynamic Treatment Regimes," *Annual Review of Statistics and its Application*, 1, 447–464. [273]

Chernofsky, A., Bosch, R. J., and Lok, J. J. (2021), "Causal Mediation Analysis with Mediator Values Below an Assay Limit," arXiv:2107.14782. [283]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68. [280]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014), "Gaussian Approximation of Suprema of Empirical Processes," *The Annals of Statistics*, 42, 1564–1597. [279]

Dai, B., Nachum, O., Chow, Y., Li, L., Szepesvari, C., and Schuurmans, D. (2020), "Coindice: Off-policy Confidence Interval Estimation," in *Advances in Neural Information Processing Systems* (Vol. 33). [273]

Elzayn, H., Jabbari, S., Jung, C., Kearns, M., Neel, S., Roth, A., and Schutzman, Z. (2019), "Fair Algorithms for Learning in Allocation Problems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 170–179. [274]

Ertefaie, A. (2014), "Constructing Dynamic Treatment Regimes in Infinite-Horizon Settings," arXiv:1406.0764. [273]

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020), "A Theoretical Analysis of Deep q-earning," *Learning for Dynamics and Control*, PMLR, pp. 486–489. [279]

Feng, Y., Ren, T., Tang, Z., and Liu, Q. (2020), "Accountable Off-policy Evaluation with Kernel Bellman Statistics," arXiv:2008.06668. [273]

Fulcher, I. R., Shpitser, I., Marealle, S., and Tchetgen Tchetgen, E. J. (2020), "Robust Inference on Population Indirect Causal Effects: The Generalized Front Door Criterion," *Journal of the Royal Statistical Society*, Series B, 82, 199–214. [275,276,277]

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. (2019), "Guidelines for Reinforcement Learning in Healthcare," *Nature Medicine*, 25, 16–18. [273]

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvári, C., and Wang, M. (2021), "Bootstrapping Statistical Inference for Off-policy Evaluation," arXiv:2102.03607. [273]

Hu, X., Qian, M., Cheng, B., and Cheung, Y. K. (2020), "Personalized Policy Learning Using Longitudinal Mobile Health Data," arXiv:2001.03258. [273]

Jiang, N., and Li, L. (2016), "Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning," *International Conference on Machine Learning*, PMLR, pp. 652–661. [273]

Jin, S. T., Kong, H., Wu, R., and Sui, D. Z. (2018), "Ridesourcing, the Sharing Economy, and the Future of Cities," *Cities*, 76, 96–104. [274]

Kallus, N., Mao, X., and Uehara, M. (2021), "Causal Inference under Unmeasured Confounding with Negative Controls: A Minimax Learning Approach," arXiv:2103.14029. [274]

Kallus, N., and Uehara, M. (2019), "Efficiently Breaking the Curse of Horizon in Off-policy Evaluation with Double Reinforcement Learning," arXiv arXiv–1909. [273,280]

Kallus, N., and Zhou, A. (2020), "Confounding-Robust Policy Evaluation in Infinite-Horizon Reinforcement Learning," in *Advances in Neural Information Processing Systems* (Vol. 33), eds. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, pp. 22293–22304, Curran Associates, Inc. [274,275]

Kober, J., Bagnell, J. A., and Peters, J. (2013), "Reinforcement Learning in Robotics: A Survey," *The International Journal of Robotics Research*, 32, 1238–1274. [273]

Kosorok, M. R., and Laber, E. B. (2019), "Precision Medicine," *Annual Review of Statistics and its Application*, 6, 263–286. [273]

Le, H., Voloshin, C., and Yue, Y. (2019), "Batch Policy Learning under Constraints," *International Conference on Machine Learning*, pp. 3703–3712. [273,279]

Li, C., Chan, S. H., and Chen, Y.-T. (2020), "Who make Drivers Stop? Towards Driver-Centric Risk Assessment: Risk Object Identification via Causal Inference," arXiv:2003.02425. [273]

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., and Jurafsky, D. (2016), "Deep Reinforcement Learning for Dialogue Generation," arXiv:1606.01541. [273]

Liao, P., Klasnja, P., and Murphy, S. (2021), "Off-Policy Estimation of Long-Term Average Outcomes with Applications to Mobile Health," *Journal of the American Statistical Association*, 116, 382–391. [273]

Liao, P., Qi, Z., and Murphy, S. (2020), "Batch Policy Learning in Average Reward Markov Decision Processes," arXiv:2007.11771. [273,279]

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018), "Breaking the Curse of Horizon: Infinite-Horizon Off-policy Estimation," in *Advances in Neural Information Processing Systems* (Vol. 31), pp. 5356–5366. [273,277,279,280]

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020),"Estimating Dynamic Treatment Regimes in Mobile Health Using v-Learning," *Journal of the American Statistical Association*, 115, 692–706. [273]

Luedtke, A. R., and Van Der Laan, M. J. (2016), "Statistical Inference for the Mean Outcome under a Possibly Non-unique Optimal Treatment Strategy," *Annals of Statistics*, 44, 713–742. [273]

Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., and Popovic, Z. (2014), "Offline Policy Evaluation across Representations with Applications to Educational Games," in *AAMAS*, pp. 1077–1084. [273]

Matsouaka, R. A., Li, J. and Cai, T. (2014), "Evaluating Marker-Guided Treatment Selection Strategies," *Biometrics*, 70, 489–499. [273]

Murphy, S. A. (2003), "Optimal Dynamic Treatment Regimes," *Journal of the Royal Statistical Society*, Series B, 65, 331–355. [273]

Nachum, O., Chow, Y., Dai, B., and Li, L. (2019), "Dualdice: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections," in *Advances in Neural Information Processing Systems* (Vol. 32), pp. 2318–2328. [273]

Nair, Y., and Jiang, N. (2021), "A Spectral Approach to Off-policy Evaluation for POMDPs," arXiv:2109.10502. [274]

Namkoong, H., Keramati, R., Yadlowsky, S., and Brunskill, E. (2020), "Off-policy Policy Evaluation for Sequential Decisions under Unobserved Confounding," in *Advances in Neural Information Processing Systems* (Vol. 33), 18819–18831. [274]

Pearl, J. (2009), "*Causality*, Cambridge: Cambridge University Press. [274,275,276]

Qi, Z., and Liao, P. (2020), "Robust Batch Policy Learning in Markov Decision Processes," arXiv:2011.04185. [273]

Robins, J. M. (2004), "Optimal Structural Nested Models for Optimal Sequential Decisions," in *Proceedings of the Second Seattle Symposium in Biostatistics*, pp. 189–326, Springer. [273]

Rysman, M. (2009), "The Economics of Two-Sided Markets," *Journal of Economic Perspective*, 23, 125–143. [274]

Schmidt-Hieber, J. (2020), "Nonparametric Regression using Deep Neural Networks with ReLU Activation Function," *Annals of Statistics*, 48, 1875–1897. [279]

Shi, C., Uehara, M., Huang, J., and Jiang, N. (2022), "A Minimax Learning Approach to Off-Policy Evaluation in Confounded Partially Observable Markov Decision Processes," *International Conference on Machine Learning*, pp. 20057–20094, PMLR. [274]

Shi, C., Wan, R., Chernozhukov, V., and Song, R. (2021), "Deeply-Debiased Off-policy Interval Estimation," in *Proceedings of the 38th International Conference on Machine Learning*, eds. M. Meila and T. Zhang (Vol. 139), pp. 9580–9591, PMLR. [273]

Shi, C., Zhang, S., Lu, W., and Song, R. (2022), "Statistical Inference of the Value Function for Reinforcement Learning in Infinite Horizon Settings," *Journal of the Royal Statistical Society*, Series B, 84, 765–793. [273]

Shi, X., Miao, W., Nelson, J. C., and Tchetgen Tchetgen, E. J. (2020), "Multiply Robust Causal Inference with Double-Negative Control Adjustment for Categorical Unmeasured Confounding," *Journal of the Royal Statistical Society*, Series B, 82, 521–540. [274]

Sutton, R. S., and Barto, A. G. (2018), *Reinforcement learning: An introduction*, Cambridge, MA: MIT Press. [273]

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. (2020), "An Introduction to Proximal Causal Learning," arXiv:2009.10982. [274]

Tennenholtz, G., Shalit, U., and Mannor, S. (2020), "Off-policy Evaluation in Partially Observable Environments," *AAAI*, pp. 10276–10283. [274]

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. (2015), "High-Confidence Off-policy Evaluation," *Twenty-Ninth AAAI Conference on Artificial Intelligence*. [273]

Tsiatis, A. (2007), *Semiparametric Theory and Missing Data*, New York: Springer. [274]

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. (2019), *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*, Boca Raton, FL: CRC Press. [273]

Uehara, M., Huang, J., and Jiang, N. (2020), "Minimax Weight and q-function Learning for Off-policy Evaluation," in *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119), eds. H. D. III and A. Singh, PMLR, pp. 9659–9668. [273]

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., and Xie, T. (2021), "Finite Sample Analysis of Minimax Offline Reinforcement Learning: Completeness, Fast Rates and First-Order Efficiency," arXiv:2102.02981. [279]

Van der Vaart, A. W. (2000), *Asymptotic Statistics* (Vol. 3), Cambridge: Cambridge University Press. [273]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [279]

Wang, L., Yang, Z., and Wang, Z. (2020), "Provably Efficient Causal Reinforcement Learning with Confounded Observational Data," arXiv:2006.12311. [274,275]

Wang, L., Zhou, Y., Song, R., and Sherwood, B. (2018), "Quantile-Optimal Treatment Regimes," *Journal of the American Statistical Association*, 113, 1243–1254. [273]

Wu, Y., and Wang, L. (2020), "Resampling-based Confidence Intervals for Model-Free Robust Inference on Optimal Treatment Regimes," *Biometrics*, 77, 465–476. [273]

Xu, Z., Laber, E., Staicu, A.-M., and Severus, E. (2020), "Latent-State Models for Precision Medicine," arXiv:2005.13001. [273]

Xu, Z., Li, Z., Guan, Q., Zhang, D., Li, Q., Nan, J., Liu, C., Bian, W., and Ye, J. (2018), "Large-Scale Order Dispatch in On-demand Ride-Hailing Platforms: A Learning and Planning Approach," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 905–913. [282]

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012), "A Robust Method for Estimating Optimal Treatment Regimes," *Biometrics*, 68, 1010–1018. [273]

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013), "Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions," *Biometrika*, 100, 681–694. [273,277]

Zhang, J., and Bareinboim, E. (2016)"Markov Decision Processes with Unobserved Confounders: A Causal Approach," *Technical Report*, Technical Report R-23, Purdue AI Lab. [274]