Two-way Deconfounder for Off-policy Evaluation in Causal Reinforcement Learning

Shuguang Yu[*]
School of Statistics and Management
Shanghai University of Finance and Economics
Shanghai, China

Shuxing Fang[*]
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong, China

Ruixin Peng
School of Statistics and Management
Shanghai University of Finance and Economics
Shanghai, China

Zhengling Qi
Department of Decision Sciences
George Washington University
Washington D.C., USA

Fan Zhou[†]
School of Statistics and Management
Shanghai University of Finance and Economics
Shanghai, China

Chengchun Shi[†]
Department of Statistics
London School of Economics and Political Science
London, UK

Abstract

This paper studies off-policy evaluation (OPE) in the presence of unmeasured confounders. Inspired by the two-way fixed effects regression model widely used in the panel data literature, we propose a two-way unmeasured confounding assumption to model the system dynamics in causal reinforcement learning and develop a two-way deconfounder algorithm that devises a neural tensor network to simultaneously learn both the unmeasured confounders and the system dynamics, based on which a model-based estimator can be constructed for consistent policy value estimation. We illustrate the effectiveness of the proposed estimator through theoretical results and numerical experiments.

1 Introduction

Before deploying any newly developed policy, it is important to assess its impact. In many high-stakes domains, it is risky or unethical to implement such policies directly for online evaluation. This challenge highlights the essential role of off-policy evaluation (OPE).

There is a vast body of literature on OPE. Most studies assume there are no unmeasured confounders (NUC), also known as unconfoundedness (see e.g., Thomas et al. 2015, Jiang and Li 2016, Thomas and Brunskill 2016, Farajtabar et al. 2018, Liu et al. 2018, Irpan et al. 2019, Schlegel et al. 2019, Tang et al. 2019, Xie et al. 2019, Dai et al. 2020, Chandak et al. 2021, Hao et al. 2021, Liao et al. 2021b,

[*]Equal contribution.

[†]Corresponding authors: zhoufan@mail.shufe.edu.cn, c.shi7@lse.ac.uk.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Shi et al. 2021, Chen and Qi 2022, Kallus and Uehara 2022, Liao et al. 2022, Shi et al. 2022b, Xie et al. 2023, Zhou et al. 2023a). However, the NUC assumption is restrictive and untestable from the data. It can be violated in various domains such as urgent care [Namkoong et al., 2020], autonomous driving [Nyholm and Smids, 2020], ride-sharing [Shi et al., 2022c], and bidding [Xu et al., 2023]. Applying standard OPE methods that rely on the NUC assumption in these settings would result in a biased policy value estimator [see e.g., Bennett and Kallus, 2023, Section 6.1].

Causal reinforcement learning studies offline policy optimization or OPE in the presence of unmeasured confounding. Many existing works can be divided into one of the following three groups:

1. ***Methods under memoryless unmeasured confounding***: The first type of methods relies on a "memoryless unmeasured confounding" assumption to guarantee that the observed data satisfies the Markov property in the presence of unmeasured confounders [Zhang and Bareinboim, 2016, Kallus and Zhou, 2020, Li et al., 2021, Liao et al., 2021a, Wang et al., 2021, Chen et al., 2022, Fu et al., 2022, Shi et al., 2022c, Yu et al., 2022, Bruns-Smith and Zhou, 2023, Xu et al., 2023]. Many of these works also require external proxy variables (e.g., mediators and instrumental variables) to handle latent confounders. In contrast, the method proposed in this article neither relies on the Markov assumption nor requires external proxies.
2. ***POMDP-type methods***: The second category employs a partially observable Markov decision process (POMDP) to model unmeasured confounders as latent states, drawing on ideas from the proximal causal inference literature [Tchetgen et al., 2020] to address unmeasured confounding [Tennenholtz et al., 2020, Nair and Jiang, 2021, Miao et al., 2022, Shi et al., 2022a, Wang et al., 2022, Bennett and Kallus, 2023, Hong et al., 2023, Lu et al., 2023]. However, these works require restrict mathematical assumptions that are hard to verify in practice [Lu et al., 2018].
3. ***Deconfounding-type methods***: The final category originates from deconfounding methods in causal inference, which leverage the inherent structure within observed data, such as multiple treatment dependencies, network structures, and exposure models, to address unmeasured confounding [Louizos et al., 2017, Tran and Blei, 2017, Wang et al., 2018, Veitch et al., 2019, Wang and Blei, 2019, Zhang et al., 2019, Bica et al., 2020, Veitch et al., 2020, Shah et al., 2022, McFowland III and Shalizi, 2023, Shuai et al., 2023]. Notably, Wang and Blei [2019] directly estimate latent confounders for causal inference. However, their algorithm requires the unmeasured confounders to be a deterministic function of the actions [Ogburn et al., 2020, Wang and Blei, 2019], which has been criticized as being unreasonable [D'Amour, 2019, Ogburn et al., 2019]; refer to Appendix A.1. There have also been several extensions of deconfounding methods to reinforcement learning (RL) [Lu et al., 2018, Hatt and Feuerriegel, 2021, Kausik et al., 2022, 2023]. However, these methods either impose a one-way unmeasured confounding assumption, which can be overly restrictive (see Section 2), or require the correct specification of the latent variable [Rissanen and Marttinen, 2021].

This paper aims to develop advanced deconfounding-type OPE methodologies, allowing for more flexible assumptions regarding latent variable modeling. Our proposal is inspired by the two-way fixed effects (2FE) model which is widely employed in applied economics Mundlak [1961], Baltagi and Baltagi [2008], Griliches [1979], Anderson and Hsiao [1982], Freyberger [2018], Callaway and Karami [2023] and causal inference with panel data Arkhangelsky and Imbens [2022], De Chaisemartin and d'Haultfoeuille [2020], Imai and Kim [2021], Sant'Anna and Zhao [2020], Athey and Imbens [2022]. More recently, Dwivedi et al. [2022] applied the 2FE model to counterfactual prediction in a contextual bandit setting and Bian et al. [2023] extended the 2FE model to RL. However, their investigations primarily consider an unconfounded setting. Additionally, the model proposed by Bian et al. [2023] imposes a restrictive additive assumption – requiring the latent factors to influence both the reward and transition functions in a purely additive manner. Furthermore, they rely on linear function approximation to estimate the policy value, which may fail to capture the inherently complex nonlinear dynamics. In contrast, we employ flexible neural networks to model the environment.

In this article, we propose a novel two-way unmeasured confounding assumption to effectively model latent confounders. This approach categorizes all unmeasured confounders into two distinct groups: those that are time-specific and those that are trajectory-specific. This assumption enhances the model's flexibility beyond one-way unmeasured confounding while ensuring that the total number of confounders remains much smaller than the sample size, making them estimable from the observed data. We further develop an original two-way deconfounder algorithm that constructs a neural tensor network to jointly learn the unmeasured confounders and the system dynamics. Based on the learned

model, we construct a model-based estimator to accurately estimate the policy value. Our proposed model for unmeasured confounders shares similarities with latent factor models used to capture two-way interactions, such as item-customer interactions in recommendation systems [Hu et al., 2008, He et al., 2017], and relationships between different entities in multi-relational learning [Socher et al., 2013, Nickel et al., 2015, Nguyen et al., 2017, Wang et al., 2017, Ji et al., 2021].

To summarize, our contributions include: (1) the introduction of a novel two-way unmeasured confounding assumption; (2) the development of a new two-way deconfounder algorithm for model-based OPE under unmeasured confounding; (3) the demonstration of the effectiveness of our model and algorithm.

2 Two-way Unmeasured Confounding

In this section, we begin by presenting the data generating process under unmeasured confounding and outlining our objectives. We then introduce the proposed two-way unmeasured confounding assumption and compare it against alternative assumptions.

We consider an offline setting with pre-collected observational data $\mathcal{D}$, containing N trajectories, each consisting of T time points. We use i to index the i-th trajectory and t to index the t-th time point. For each pair of indices (i, t), its associated data is given by the observation-action-reward triplet $(O_{i,t}, A_{i,t}, R_{i,t})$. In healthcare applications, each trajectory represents an individual patient where $O_{i,t}$ denotes the covariates of the i-th patient at time t, $A_{i,t}$ denotes the treatment assigned to the patient, and $R_{i,t}$ measures their clinical outcome at time t.

We investigate a confounded setting characterized by the presence of unmeasured confounders (denoted by $Z_{i,t}$) that influence both $A_{i,t}$ and $(R_{i,t}, O_{i,t+1})$ for each pair (i, t). The offline data generating process (DGP) can be described as follows: (1) At each time t, the observation $O_{i,t}$ is recorded for the i-th trajectory. (2) Subsequently, an action $A_{i,t}$ is assigned for the i-th subject according to a behavior policy π_b, such that $A_{i,t} \sim \pi_b(\bullet|O_{i,t}, Z_{i,t})$. (3) Next, we obtain the immediate reward $R_{i,t}$ and the next observation $O_{i,t+1}$ such that $(R_{i,t}, O_{i,t+1}) \sim \mathcal{P}(\bullet|A_{i,t}, O_{i,t}, Z_{i,t})$ for some transition function $\mathcal{P}$. (4) Steps 2 and 3 are repeated until we reach the termination time T. See Figure 1(a) for an illustration.

In contrast, following a given target policy π we wish to evaluate, the data is generated as follows: (1) At each time t, the action $A_{i,t}$ is determined by the target policy $\pi(\bullet|O_{i,t})$, independent of the unmeasured confounder $Z_{i,t}$. (2) The immediate reward $R_{i,t}$ and next observation $O_{i,t+1}$ are generated according to the transition function $\mathcal{P}(\bullet|A_{i,t}, O_{i,t}, Z_{i,t})$. In this setup, the unmeasured confounders affect only the reward and next observation distributions, but not the action. This is a primary difference from the offline data generating process. Specifically, whatever relationship exists between the unmeasured confounders and the actions in the offline data, that relationship is no longer in effect when we perform the target policy π. Our objective lies in evaluating the expected cumulative reward under π, given by

$$\eta^{\pi} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}^{\pi}(R_{i,t}),$$

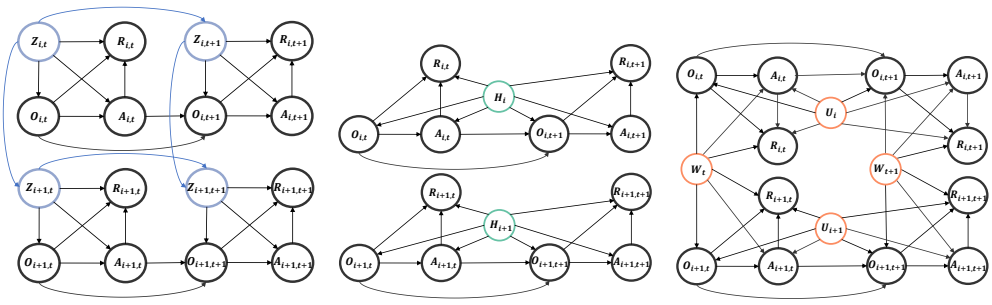
where $\mathbb{E}^{\pi}(R_{i,t})$ represents the expectation under the target policy π, irrespective of the unmeasured confounders.

To better understand the proposed two-way unmeasured confounding assumption, we first introduce two alternative assumptions concerning the unmeasured confounders as follows:

(a) ***Unconstrained unmeasured confounding*** (UUC): Each pair of indices (i, t) corresponds to an unmeasured confounder $Z_{i,t}$, and there are no restrictions on these values.

(b) ***One-way unmeasured confounding*** (OWUC): For any i, the unmeasured confounders remain the same across time, i.e., $H_i = Z_{i,1} = Z_{i,2} = \cdots = Z_{i,T}$.

These two assumptions represent two extremes. The UUC assumption offers maximal flexibility without imposing any specific conditions. In contrast, the OWUC assumption is restrictive, essentially excluding 'time-varying confounders'.



(a) UUC (b) OWUC (c) TWUC

Figure 1: The directed acyclic graphs of data generating processes under different assumptions. $(a) : \{z_{i,t}\}_{i,t}$ (colored in blue) are unconstrained unmeasured confounders. $(b) : \{h_i\}_i$ (colored in green) are one-way unmeasured confounders. $(c) : \{u_i\}_i$ and $\{w_t\}_t$ (colored in orange) are two-way unmeasured confounders.

The validity of the deconfounder algorithm [Wang and Blei, 2019] relies crucially on the consistent[3] estimation of the latent confounders. This is because it employs a plug-in method for constructing the average treatment effect estimator, which plugs in the estimated latent confounders into the model.

Unconstrained unmeasured confounding imposes no restrictions on the latent variables, but requires to estimate a total of NT latent variables, which is equal to the sample size. This make consistent estimation infeasible without resorting to the 'deterministic unmeasured confounding' assumption discussed in Section 1. On the other hand, one-way unmeasured confounding only requires estimating N latent variables, but in reality, this assumption is often difficult to meet.

In this article, we propose the following two-way unmeasured confounding, which offers a middle ground between the UUC assumption and the OWUC assumption:

(c) ***Two-way unmeasured confounding*** (TWUC): There exist time-invariant confounders $\{U_i\}_i$ and trajectory-invariant confounders $\{W_t\}_t$ such that $Z_{i,t} = (U_i^\top, W_t^\top)^\top$.

As discussed in the introduction, TWUC requires that all unmeasured confounders belong to one of two groups: trajectory-specific time-invariant confounders and time-specific trajectory-invariant confounders. Notably, it excludes confounders that are both trajectory- and time-specific. The U_is can be interpreted as individual baseline information (e.g., salary or educational background) that remains consistent over time, while the W_ts represent external factors (e.g., weather or holidays) exerting a common influence across all trajectories. This assumption effectively relaxes one-way unmeasured confounding by accommodating time-varying confounders. Meanwhile, the number of latent confounders is confined to $N + T$, much smaller than the sample size $N \times T$ when both N and T grow to infinity. This ensures the feasibility of consistent estimation. See Figure 1 for a graphical visualization of the three assumptions.

To further elaborate the three modeling assumptions (a) – (c), we consider a linear model setup where the conditional means of the next observation and the immediate reward are linear functions of the current observation-action pair as well as the unmeasured confounders, and summarize the implications of adopting the three assumptions in the following corollaries.

Proposition 1 (Inconsistency of the unconstrained model). The least square estimator (LSE) based on the UUC assumption cannot yield consistent predictions. Its mean square error (MSE) remains constant as both N and T increase.

Proposition 2 (Inconsistency of the one-way model under misspecification). When time-varying unmeasured confounders exist, the LSE based on the one-way model cannot yield consistent predictions. Its MSE remains constant as both N and T increase.

Proposition 3 (Consistency of the two-way model). The LSE based on the two-way model can yield consistent predictions. Its MSE decays to zero as both N and T increase.

Furthermore, we design a linear simulation setting to numerically compare the estimators based on these three assumptions and report their MSEs in Figure 2(b). The results reveal that the unconstrained

[3]Here, consistency means that the estimators converge to their ground truth as the sample size increases [see e.g., Casella and Berger, 2021].

model tends to overfit the data, as evidenced by its lowest prediction error on the training dataset and higher error in off-policy value estimation. Conversely, the one-way model underfits the data. It achieves the largest errors in both training data prediction and off-policy value estimation. In contrast, our proposed two-way model strikes a balance, resulting in the lowest error in off-policy value estimation.

To conclude this section, we summarize the DGP under the proposed two-way unmeasured confounding assumption:

- At the initial time, each trajectory independent generates an observed $O_{i,1}$ and an unobserved U_i. Additionally, a latent W_1 is independently generated.
- Next, at each time t, $A_{i,t}$ and $(R_{i,t}, O_{i,t+1})$ are generated, according to π_b (or a target policy π) and $\mathcal{P}$, respectively. Moreover, a latent W_{t+1} is generated, whose distribution depends only on the past time-specific confounders.

Under this DGP, the two-way unmeasured confounders are policy-agnostic, i.e., unaffected by the actions or policies. Consequently, we can employ a plug-in approach to construct the policy value estimator (see (1)), eliminating the need to estimate their distributions.

3 Two-way Deconfounder

We introduce the proposed deconfounder algorithm in this section. We first present the proposed neural network architecture under two-way unmeasured confounding. We next define the loss function used for model training. Finally, we introduce a model-based policy value estimator, built upon the estimated model.

Network Architecture. The proposed model contains the following components: (1) d-dimensional embedding vectors $\{u_i\}_i$ to model the trajectory-specific latent confounders; (2) d-dimensional embedding vectors $\{w_t\}_t$ to model the time-specific latent confounders; (3) A transition function $\widehat{\mathcal{P}}(\bullet|a, o, u_i, w_t)$ that takes an action-observation pair (a, o) and a pair of embedding vectors (u_i, w_t) as input to model the conditional distribution of the reward-next-observation pair given the current observation-action pair and the two-way unobserved confounders; (4) An actor network $\widehat{\pi}_b(\bullet|o, u_i, w_t)$ that takes o and (u_i, w_t) as input to model the behavior policy.

Our objective is to simultaneously learn the embedding representations and the parameters in both the transition and actor networks from the observed data. Toward that end, we treat each pair of embedding vectors (u_i, w_t) as two entities. To accurately capture their joint effects on both the transition function and the behavior policy, we adopt the neural tensor network [NTN, Socher et al., 2013], which is known for its ability to capture the intricate interactions between pairs of entity vectors.

Specifically, we parameterize $\widehat{\mathcal{P}}(\bullet|a, o, u_i, w_t)$ via a conditional Gaussian model given by $\mathcal{N}\left(\widehat{\mu}\left(a, o, u_i, w_t\right), \text{diag}(\widehat{\sigma}\left(a, o, u_i, w_t\right))\right)$, where $\text{diag}(\widehat{\sigma}\left(a, o, u_i, w_t\right))$ is a diagonal matrix and $\widehat{\sigma}\left(a, o, u_i, w_t\right)$ denotes the vector consisting of all diagonal elements. Its conditional mean and variance functions are modeled jointly with the behavior policy, specified by

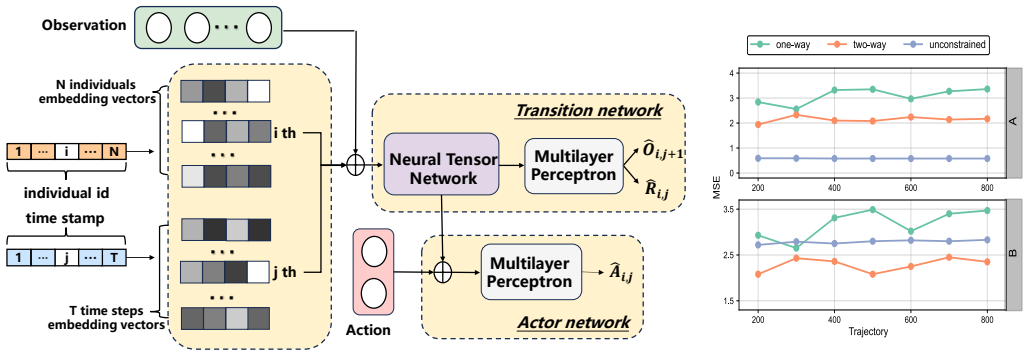
$$(\widehat{\mu}^\top, \widehat{\sigma}^\top)^\top = \text{MLP}_{\mathcal{P}}(a_{i,t}, \text{NTN}(o, u_i, w_t)),\ \widehat{\pi}_b(\bullet \mid o, u_i, w_t) = \text{MLP}_{\pi_b}(\text{NTN}(o, u_i, w_t)),$$

where $\text{NTN}(o, u_i, w_t)$ denotes the d-dimensional output vector from an NTN, and $\text{MLP}_{\mathcal{P}}$, MLP_{π_b} are the two multilayer perceptrons that take this output vector as input. As such, the NTN layer captures the joint information from both the observation o and latent confounders u_i, w_t and is shared among the actor and transition networks. We provide an overview of the proposed model architecture in Figure 2 (a).

Finally, we remark that while we model the transition function using a conditional Gaussian like other works in the RL literature [see e.g., Janner et al., 2019, Yu et al., 2020], more complex generative models such as transformers or diffusion models are equally applicable.

Loss Function. As mentioned earlier, the proposed model contains both the transition network and the actor network. Accordingly, our loss function is given by

$$L\left(\mathcal{D}; \{u_i\}_i, \{w_t\}_t\right) = (1-\alpha) \cdot L_T + \alpha \cdot L_A,$$



(a) network architecture (b) numerical experiments

Figure 2: (a) : An overview of the proposed network architecture. (b) : The upper panel reports MSEs under different unmeasured confounding assumptions for fitting the observed data whereas the bottom panel displays the MSEs for off-policy value prediction. The unconstrained unmeasured confounding model shows the best fit for the training data, due to overfitting. The OPE estimator under the proposed two-way unmeasured confounding achieves the smallest MSE. More details are referred to Appendix D.1.

where L_T is the negative log likelihood of conditional Gaussian model for the transition network, L_A is the cross-entropy loss between the actual action and actor network, and $\alpha \in (0, 1)$ is a hyperparameter that balances the two losses. This loss is optimized to compute both the latent embeddings vectors (u_i, w_t)s and the parameters in the three neural networks: NTN, $\mathrm{MLP}_{\mathcal{P}}$ and MLP_{π_b}. Further details are relegated to Appendix C.1 to save space.

Alternative to our loss function which involves both the actor and transition networks, one can consider minimizing other losses focused exclusively on either the transition or the actor network, but not necessarily both. However, in the presence of unmeasured confounding, it is essential to note that both the behavior policy and transition function are influenced by these latent confounders. Consequently, our joint learning approach is expected to more effectively identify the unmeasured confounders compared to these transition-only or actor-only approaches.

Model-based OPE. Based on the estimated model, we develop a model-based estimator to learn η^π via Monte Carlo policy evaluation [Sutton and Barto, 2018]. A key observation is that, since the two-way unmeasured confounders are policy-agnostic, the empirical sum shown below is an unbiased intermediate estimator for the evaluation target η^π:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}^{\pi}\left(R_{i,t} \mid O_{i,1}, U_i, \{W_{t'}\}_{t'=1}^{t}\right). \tag{1}$$

Note that in this formulation, the latent factors in the conditioning set can be substituted with our estimated embedding vectors. Additionally, based on our estimated transition network, the expectation $\mathbb{E}^{\pi}$ can be effectively approximated using the Monte Carlo method. This approach allows us to construct a model-based plug-in estimator for Equation (1).

To elaborate, the Monte Carlo simulation begins at $t = 1$ for each individual i. We sample an action $\widehat{A}_{i,t}$ according to the target policy $\pi(\bullet \mid \widehat{O}_{i,t})$, draw samples for $\widehat{R}_{i,t}$ and $\widehat{O}_{i,t+1}$ from the learned transition network $\widehat{\mathcal{P}}(\bullet \mid \widehat{A}_{i,t}, \widehat{O}_{i,t}, \widehat{U}_i, \widehat{W}_t)$ with the estimated latent factors $\{\widehat{U}_i\}_i$, $\{\widehat{W}_t\}_t$, and iterate this procedure until the terminal time T is reached. Next, we replicate this simulation multiple times to reduce the Monte Carlo error. The final step is to aggregate all the estimated results across these simulations to construct the final OPE estimator.

4 Theoretical Results

In this section, we provide a finite-sample error bound for the expected difference between the estimated policy value and the ground truth η^π. We first introduce some notations. We consider a tabular setting where the observation space $\mathcal{O}$, action space $\mathcal{A}$ and latent factor spaces $\mathcal{U}$, $\mathcal{W}$ are all discrete. Let $d_{\bar{\mathcal{D}}}$ denote the data distribution of quadruples (a, o, u, w), $d_{\bar{\mathcal{D}}}(a, o, u, w) = (NT)^{-1}\sum_{(A,O,U,W)\in\bar{\mathcal{D}}}\mathbb{P}(A = a, O = o, U = u, W = w)$, where the summation $\sum_{(A,O,U,W)\in\bar{\mathcal{D}}}$

is carried out over all quadruples in the augmented dataset $\bar{\mathcal{D}} = \mathcal{D} \cup \{U_i\}_i \cup \{W_t\}_t$ containing both the observed data and the latent confounders. Furthermore, let d^π represent the visitation distribution under π. Specifically, $d^\pi(a, o, u, w)$ denotes the average probability of a given quadruple (a, o, u, w) appearing at any time step under π (refer to Appendix B.1 for the detailed definition).

We next introduce some assumptions to derive the finite-sample error bound of the proposed policy value estimator.

Assumption 1 (Coverage). The data distribution covers the visitation distribution induced by the target policy π, i.e., $C = \sup_{a,o,u,w} \frac{d^\pi(a,o,u,w)}{d_{\bar{\mathcal{D}}}(a,o,u,w)} < \infty$.

Assumption 2 (Boundedness). The absolute values of the immediate reward are upper bounded by some constant $R_{\max} < \infty$.

Assumption 3 (Error bound of estimated transition function). $\mathbb{E}\|\widehat{\mathcal{P}} - \mathcal{P}\|^2_{d_{\bar{\mathcal{D}}}} \leq \varepsilon^2_{\mathcal{P}}$ for some $\varepsilon_{\mathcal{P}} > 0$ where $\|\widehat{\mathcal{P}} - \mathcal{P}\|_{d_{\bar{\mathcal{D}}}}$ denotes the total variation distance between $\widehat{\mathcal{P}}$ and $\mathcal{P}$ (see Appendix B.1 for the detailed definition).

Assumption 4 (Error bound of estimated latent confounders). $\sum_{1\leq i\leq N} \mathbb{E}\|\widehat{U}_i - U_i\|_2^2/N \leq \varepsilon^2_{U,W}$ and $\sum_{1\leq t\leq T} \mathbb{E}\|\widehat{W}_t - W_t\|_2^2/N \leq \varepsilon^2_{U,W}$ for some $\varepsilon_{U,W} > 0$.

Assumption 5 (Autocorrelation). For any $t_1, t_2 \geq 1$, let $\rho(t_1, t_2)$ denote the correlation coefficient between $\mathbb{E}^\pi(R_{1,t_1}|\{W_{t'}\}_{t'=1}^{t_1})$ and $\mathbb{E}^\pi(R_{1,t_2}|\{W_{t'}\}_{t'=1}^{t_2})$. There exists some $0 \leq \alpha \leq 1$ such that $\sum_{1\leq t_1\neq t_2\leq T} \rho(t_1, t_2) = O(T^{2\alpha})$.

We remark that conditions similar to Assumptions 1-2 are widely imposed in the RL literature to simplify the theoretical analysis [see e.g., Chen and Jiang, 2019, Fan et al., 2020, Liu et al., 2020, Uehara and Sun, 2021]. Assumption 3 is concerned with the estimation error of the transition function. This error is expected to be minimal, since we use neural networks for function approximation [see e.g., Schmidt-Hieber, 2020, Farrell et al., 2021]. Assumptions 4 is concerned with the estimation errors of the estimated two-way unmeasured confounders. According to Proposition 3, these errors are negligible under simple models. Finally, Assumption 5 is purely technical. It measures the autocorrelation of the time series $\mathbb{E}^\pi(R_{1,t}|\{W_{t'}\}_{t'=1}^{t})$. Here, $\alpha = 0$ indicates independence over time. This condition is automatically satisfied when $\alpha = 1$.

Theorem 1 (Finite-sample error bound). Suppose the two-way unmeasured confounding assumption holds, and Assumptions 1-5 are satisfied. Then

$$\mathbb{E}|\widehat{\eta}^\pi - \eta^\pi| \leq CTR_{\max}\varepsilon_{\mathcal{P}} + cTR_{\max}\varepsilon_{U,W} + cR_{\max}N^{-1/2} + cR_{\max}T^{\alpha-1},$$

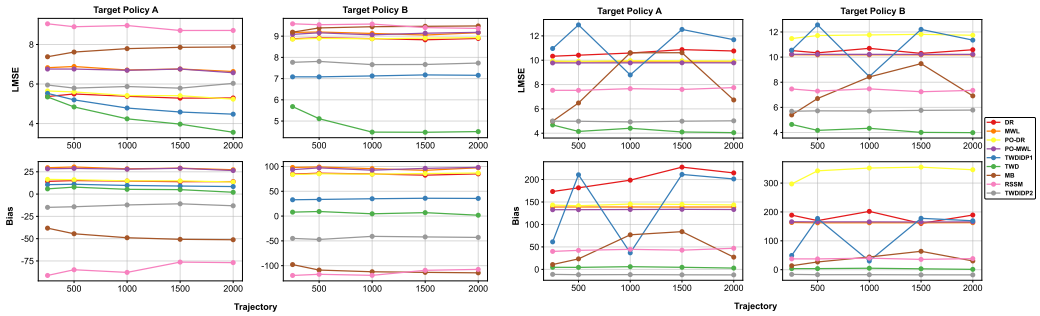
for some constant $c > 0$.

It can be seen from Theorem 1 that the mean absolute error of the proposed policy value estimator involves two components:

1. **Estimation errors:** The first two terms on the right side of the inequality correspond to the estimation errors of the transition function and latent confounders respectively. Notably, both terms are linear in the time horizon T, due to error accumulation (see Appendix B.5). However, such a linear dependence is the best one can hope in general [Jiang, 2024], although it is possible to eliminate the dependence upon T under additional ergodicity assumptions [Liao et al., 2022].
2. **Standard deviations:** The last two terms measure the standard deviations of the average values across trajectories and over time, respectively. These upper bounds decays to zero, as both N and T approach infinity, provided that the exponent α in Assumption 5 – which measures the autocorrelation of the time series $\mathbb{E}^\pi(R_{1,t}|\{W_{t'}\}_{t'=1}^{t})$ – is strictly smaller than 1.

5 Experiments

In this section, we perform numerical experiments using two simulated datasets and one real-world dataset to demonstrate the effectiveness of the proposed two-way deconfounder (denoted by TWD) in handling unmeasured confounders. We consider two simulated examples in Section 5.1: a simple dynamic process and a tumor growth example. For each simulated example, the true value of η^π is computed based on 10,000 Monte Carlo experiments. We also explore a real-world example using the



(a) The simulated dynamic process (b) Tumor growth example

Figure 3: Logarithmic MSE and Bias of various estimators for the simulated dynamic process and tumor growth example.

MIMIC-III dataset in Section 5.2, and conduct a sensitivity analysis and an ablation study in Sections 5.3 and 5.4, respectively. The source code is available on github: https://github.com/fsmiu/Two-way-Deconfounder

We use two metrics to evaluate different OPE estimators: the logarithmic mean squared error (LMSE) and bias. Both are estimated based on 20 simulations. Comparison is made between TWD and the following set of baseline methods, which covers a wide range of model-based and model-free approaches:

(1) ***Model-based method*** (MB) that learns a transition model from the offline data based on the proposal by Yu et al. [2020] and applies the Monte Carlo method to construct the OPE estimator;

(2) ***Minimax weight learning*** [MWL, Uehara et al., 2020] that learns a marginalized importance sampling (MIS) ratio from the offline data to constructs an MIS estimator for OPE;

(3) ***Double robust method*** (DR) that combines the MIS ratio and an estimated Q-function computed via minimax learning [Uehara et al., 2020] to enhance robustness of OPE;

(4) ***Partially observable MWL*** [PO-MWL, Shi et al., 2022a] – a POMDP-type method that extends MWL to handle unmeasured confounders;

(5) ***Partially observable DR*** [PO-DR, Shi et al., 2022a] – another POMDP-type method that extends DR to handle unmeasured confounders;

(6) ***Recurrent state-space method*** [RSSM, Hafner et al., 2019a,b] that models unmeasured confounders as latent states;

(7) ***Model-free two-way doubly inhomogeneous decision process*** (TWDIDP1) – a deconfounding-type method that uses the model-free algorithm developed by Bian et al. [2023];

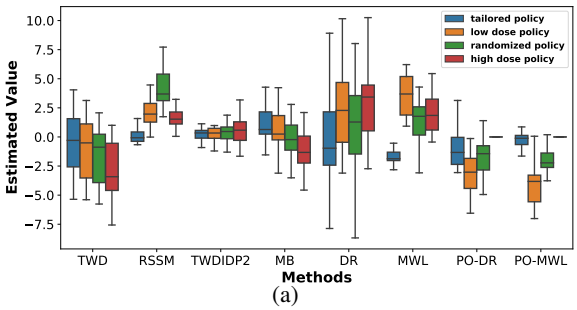
(8) ***Model-based two-way doubly inhomogeneous decision process*** (TWDIDP2) – another model-based deconfounding-type algorithm, also developed by Bian et al. [2023].

Notably, the first three methods require the NUC assumption. The next three methods are POMDP-type, whereas the last two are deconfounding-type algorithms. Given our focus on settings without external proxies, we do not compare against methods developed under memoryless unmeasured confounding, which typically rely on these proxies, as commented in Section 1.

5.1 Simulation studies

Simulated Dynamic Process. We first consider a dynamic process with four-dimensional observations and binary actions. The data is generated under the proposed TWUC assumption; see Appendix D.2 for its detailed DGP. We fix $T = 50$, and vary the number of trajectories from 250 to 2000. As shown in Figure 3(a), our proposed TWD estimator frequently achieves the smallest LMSE with bias closer to 0 in all cases. Additionally, the LMSE of TWD generally decreases as the number of trajectories increases, demonstrating its consistency. In contrast, most other methods under the assumption of NUC or POMDP setting are severely biased, highlighting the risks of ignoring or improperly handling unmeasured confounding.

Tumor Growth Example. We consider a tumor growth example and utilize the pharmacokinetic-pharmacodynamic (PK-PD) model for data generation; see Appendix D.3 for details. The observation



(a)

	Number of trajectories	
Model	2000	3708
MB	0.255±0.008	0.252±0.005
RSSM	0.332±0.014	0.329±0.008
TWDIDP2	0.262±0.008	0.259±0.006
TWD	**0.254±0.008**	**0.251±0.006**

(b)

Figure 4: (a) : The estimated policy value for four target policies in real-world dataset. (b) : Average root MSE and its standard error in the results for predicting immediate reward and next observation. The results are aggregated over 20 runs.

is two-dimensional, including the tumor volume $V(t)$ and the chemotherapy drug concentration $C(t)$. The action space $\mathcal{A}$ includes two treatments: radiotherapy $A^r(t)$ and chemotherapy $A^c(t)$. We evaluate two target policies: a random policy (denoted by 'A') and a individualized policy that is tailored to the patients' conditions (denoted by 'B'). As shown in Figure 3(b), TWD consistently achieves the lowest LMSE and is empirically unbiased in all cases, demonstrating the effectiveness of our method. In contrast, other methods yield biased estimates, resulting in significantly higher LMSEs. Importantly, the performance of these alternative methods does not show significant improvement with an increase in the total number of trajectories. These results underscore the applicability of our method in a more realistic scenario.

5.2 Real-world example: MIMIC-III database

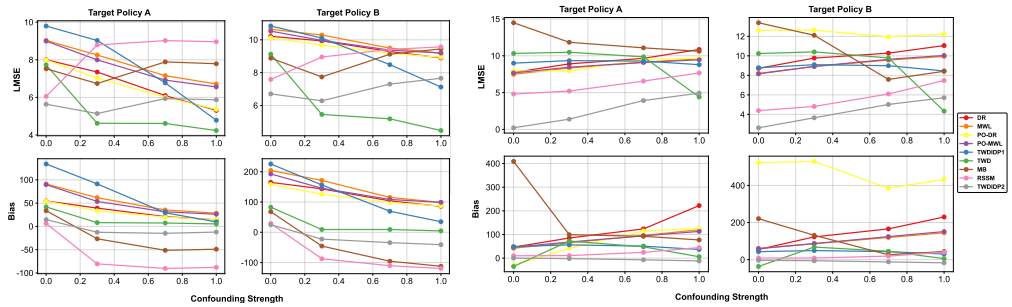
In this section, we apply the proposed two-way deconfounder to the medical information mart for intensive care (MIMIC-III) database [Johnson et al., 2016]. We extract 3,707 patients with trajectories up to 20 timesteps. Following the analysis of Zhou et al. [2023b], we define a 5 × 5 action space and set the reward to the difference between current SOFA score and next SOFA score, so a lower reward indicates a higher risk of mortality. We also extract 12 covariates as the observation; further details are described in Appendix D.4. The dataset is likely non-stationary and lacks patient's personal information. Therefore, it might be reasonable to employ TWUC to model the unmeasured confounders [Bian et al., 2023].

Given that this is a real dataset, we do not have access to the true value of the target policy. To compare TWD against other baselines, we employ two approaches. The first approach uses 90% of data for training, and the remaining 10% for evaluating an algorithm's prediction error for the reward and next observation. Notice that this approach is applicable to evaluate model-based methods only. We report all mean squared prediction errors in Figure 4(b). It can be seen that TWD results in the lowest prediction errors in all cases, demonstrating its effectiveness to infer unobserved confounders.

The second approach assesses each algorithm ability in distinguishing between tailored individualized policies and other random, non-individualized policy. An effective OPE algorithm should consistently rank an individualized policy as the superior policy. In what follows, four policies are evaluated using TWD and other baseline methods: a randomized policy, a non-individualized high dose policy, a non-individualized low dose policy and an tailored individualized policy. As shown in Figure 4(a), TWD and MB can effectively distinguish the individualized policy from other policies, with the individualized policy consistently achieving the highest estimated value. However, other methods lead to strange conclusions. For example, the result from TWDIDP2 suggest that all policies achieve similar values. Consequently, results from MWL,DR, PO-DR and RSSM suggest that the high dose policy is better than the tailored individualized policy. Additionally,TWDIDP1 performs specially poor, rendering it unsuitable for display alongside other methods in this figure.While these results require further validation by medical professionals, they highlight the potential of the proposed method in real-world medical applications.

5.3 Sensitivity analysis

In this section, we investigate how TWD performs when the proposed TWUC assumption is violated. Specifically, we incorporate unmeasured confounders $Z_{i,t}$s that are both trajectory- and time-specific



(a) The simulated dynamic process (b) Tumor Growth

Figure 5: Sensitivity analysis for the simulated dynamic process and tumor growth experiment.

in the reward function and behavior policy and we introduce a sensitivity parameter Γ to quantify the extent to which the proposed TWUC assumption is violated. When $\Gamma = 1$, the proposed TWUC assumption holds; as Γ decreases towards zero, the assumption is increasingly violated. We vary $\Gamma = \{0.0, 0.3, 0.7, 1.0\}$ in the experiments, and fix the number of the trajectory to 1000. See further details in Appendix D.5.

We focus on the two simulated examples. As shown in Figure 5(a), in the simulated dynamic process, the performance of the proposed methods remains relatively stable as long as $\Gamma > 0$. However, when $\Gamma = 0$ – where the TWUC is completely violated – TWD loses its superiority. Meanwhile, as shown in Figure 5(b), in the tumor growth example, the performance of TWD is very sensitive to Γ, and TWD performs better than other methods only if $\Gamma = 1.0$.

5.4 Ablation study

We conduct an ablation study to compare TWD against the following variants:

(1) ***TWD with transition-only loss function*** (TWD-TO): This variant employs the proposed TWUC assumption, but removes the cross-entropy loss from the objective function. Consequently, it solely uses the transition model to learn the two-way embedding vectors, without modeling the behavior policy during training.

(2) ***TWD without neural tensor network*** (TWD-MLP): In this variant, the neural tensor network is replaced with an MLP.

(3) ***One-way deconfounder without individual embedding*** (OWD-NI): This variant removes the individual embedding vector, operating under the one-way unmeasured confounding assumption.

(4) ***One-way deconfounder without time embedding*** (OWD-NT): This variant removes the time embedding vector.

We fix the number of trajectories to 1000 and report the LMSEs of various estimators in Table 1. It can be seen that: (i) OWD-NI and OWD-NT significantly underperform TWD due to their reliance on the one-way unmeasured confounding assumption. (ii) TWD consistently outperforms TWD-MLP, due to the neural tensor network's ability to capture intricate interactions between the trajectory-specific and the time-specific unmeasured confounders. (iii) In the simulated dynamic process, TWD achieves better performance than TWD-TO. This could be attributed to our proposed joint learning strategy, which simultaneously estimates both the transition function and the behavior policy and enhances the model's capability to infer unmeasured confounders. However, TWD performs worse than TWD-TO in the tumor growth example.

Table 1: Ablation study for variants of TWD

	Environments			
	DP		TG	
Model	A	B	A	B
TWD	**4.23**	**4.47**	4.42	4.33
TWD-TO	4.72	5.14	**3.58**	**3.39**
TWD-MLP	4.34	5.26	4.56	4.44
OWD-NI	7.50	8.59	6.80	6.92
OWD-NT	6.78	8.92	6.26	6.31

[1] DP: the simulated dynamic process, TG: the tumor growth example, A: Target policy A, B: Target policy B.

In summary, these results demonstrate the effectiveness of the proposed two-way unmeasured confounding model and our joint learning strategy, along with the crucial role of the neural tensor network in capturing complex interactions, all contributing to TWD's superior performance.

References

Theodore Wilbur Anderson and Cheng Hsiao. Formulation and estimation of dynamic models using panel data. *Journal of econometrics*, 18(1):47–82, 1982.

Dmitry Arkhangelsky and Guido W Imbens. Doubly robust identification for causal panel data models. *The Econometrics Journal*, 25(3):649–674, 2022.

Susan Athey and Guido W Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79, 2022.

Badi Hani Baltagi and Badi H Baltagi. *Econometric analysis of panel data*, volume 4. Springer, 2008.

Helmut Bartsch, Heike Dally, Odilia Popanda, Angela Risch, and Peter Schmezer. Genetic risk profiles for cancer susceptibility and therapy response. *Cancer Prevention*, pages 19–36, 2007.

Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *Operations Research*, 2023.

Zeyu Bian, Chengchun Shi, Zhengling Qi, and Lan Wang. Off-policy evaluation in doubly inhomogeneous environments. *arXiv preprint arXiv:2306.08719*, 2023.

Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*, pages 884–895. PMLR, 2020.

David Bruns-Smith and Angela Zhou. Robust fitted-q-evaluation and iteration under sequentially exogenous unobserved confounders. *arXiv preprint arXiv:2302.00662*, 2023.

Brantly Callaway and Sonia Karami. Treatment effects in interactive fixed effects models with a small number of time periods. *Journal of Econometrics*, 233(1):184–208, 2023.

George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021.

Yash Chandak, Scott Niekum, Bruno da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *Advances in Neural Information Processing Systems*, 34: 27475–27490, 2021.

Bin Chen and Yongmiao Hong. Testing for the markov property in time series. *Econometric Theory*, 28(1):130–178, 2012.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. *CoRR*, abs/1905.00360, 2019.

Xiaohong Chen and Zhengling Qi. On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. In *International Conference on Machine Learning*, pages 3558–3582. PMLR, 2022.

Yutian Chen, Liyuan Xu, Caglar Gulcehre, Tom Le Paine, Arthur Gretton, Nando De Freitas, and Arnaud Doucet. On instrumental variable regression for deep offline policy evaluation. *Journal of Machine Learning Research*, 23(302):1–40, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, 53:004912412210995, 05 2022. doi: 10.1177/00491241221099552.

Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Coindice: Off-policy confidence interval estimation. *Advances in neural information processing systems*, 33: 9398–9411, 2020.

Alexander D'Amour. Comment: Reflections on the deconfounder, 2019.

Clément De Chaisemartin and Xavier d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020.

Raaz Dwivedi, Katherine Tian, Sabina Tomkins, Predrag Klasnja, Susan Murphy, and Devavrat Shah. Counterfactual inference for sequential experiments. *arXiv preprint arXiv:2202.06891*, 2022.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for dynamics and control*, pages 486–489. PMLR, 2020.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Joachim Freyberger. Non-parametric panel data models with interactive fixed effects. *The Review of Economic Studies*, 85(3):1824–1851, 2018.

Zuyue Fu, Zhengling Qi, Zhaoran Wang, Zhuoran Yang, Yanxun Xu, and Michael R Kosorok. Offline reinforcement learning with instrumental variables in confounded markov decision processes. *arXiv preprint arXiv:2209.08666*, 2022.

Changran Geng, Harald Paganetti, and Clemens Grassberger. Prediction of treatment response for combined chemo-and radiation therapy for non-small cell lung cancer patients using a biomathematical model. *Scientific reports*, 7(1):13542, 2017.

Zvi Griliches. Issues in assessing the contribution of research and development to productivity growth. *The bell journal of economics*, pages 92–116, 1979.

F. Richard Guo, Anton Rask Lundborg, and Qingyuan Zhao. Confounder selection: Objectives and approaches, 2023.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019b.

Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvari, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR, 2021.

Tobias Hatt and Stefan Feuerriegel. Sequential deconfounding for causal inference with unobserved confounders. *arXiv preprint arXiv:2104.09323*, 2021.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.

Mao Hong, Zhengling Qi, and Yanxun Xu. A policy gradient method for confounded pomdps. *arXiv preprint arXiv:2305.17083*, 2023.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pages 263–272. Ieee, 2008.

Kosuke Imai and In Song Kim. On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, 29(3):405–415, 2021.

Alex Irpan, Kanishka Rao, Konstantinos Bousmalis, Chris Harris, Julian Ibarz, and Sergey Levine. Off-policy evaluation via off-policy classification. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 5437–5448, 2019.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.

Nan Jiang. A note on loss functions and error compounding in model-based reinforcement learning. *arXiv preprint arXiv:2404.09946*, 2024.

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

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 70(6):3282–3302, 2022.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in neural information processing systems*, 33:22293–22304, 2020.

Chinmaya Kausik, Yangyi Lu, Kevin Tan, Maggie Makar, Yixin Wang, and Ambuj Tewari. Offline policy evaluation and optimization under confounding. *arXiv preprint arXiv:2211.16583*, 2022.

Chinmaya Kausik, Kevin Tan, and Ambuj Tewari. Learning mixtures of markov chains and mdps. In *International Conference on Machine Learning*, pages 15970–16017. PMLR, 2023.

Jin Li, Ye Luo, and Xiaowei Zhang. Causal reinforcement learning: An instrumental variable approach. *arXiv preprint arXiv:2103.04021*, 2021.

Luofeng Liao, Zuyue Fu, Zhuoran Yang, Yixin Wang, Mladen Kolar, and Zhaoran Wang. Instrumental variable value iteration for causal offline reinforcement learning. *arXiv preprint arXiv:2102.09907*, 2021a.

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

Peng Liao, Zhengling Qi, Runzhe Wan, Predrag Klasnja, and Susan A Murphy. Batch policy learning in average reward markov decision processes. *Annals of statistics*, 50(6):3364, 2022.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1264–1274. Curran Associates, Inc., 2020.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*, 2018.

Miao Lu, Yifei Min, Zhaoran Wang, and Zhuoran Yang. Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. In *The Eleventh International Conference on Learning Representations*, 2023.

Edward McFowland III and Cosma Rohilla Shalizi. Estimating causal peer influence in homophilous social networks by inferring latent locations. *Journal of the American Statistical Association*, 118 (541):707–718, 2023.

Rui Miao, Zhengling Qi, and Xiaoke Zhang. Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. *Advances in Neural Information Processing Systems*, 35:593–606, 2022.

Yair Mundlak. Empirical production function free of management bias. *Journal of Farm Economics*, 43(1):44–56, 1961.

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

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. A novel embedding model for knowledge base completion based on convolutional neural network. *arXiv preprint arXiv:1712.02121*, 2017.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2015.

Sven Nyholm and Jilles Smids. Automated cars meet human drivers: responsible human-robot coordination and the ethics of mixed traffic. *Ethics and Information Technology*, 22:335–344, 2020.

Elizabeth L. Ogburn, Ilya Shpitser, and Eric J.Tchetgen Tchetgen. Comment on "blessings of multiple causes". *Journal of the American Statistical Association*, 114(528):1611–1615, October 2019. ISSN 0162-1459. doi: 10.1080/01621459.2019.1689139.

Elizabeth L Ogburn, Ilya Shpitser, and Eric J Tchetgen Tchetgen. Counterexamples to" the blessings of multiple causes" by wang and blei. *arXiv preprint arXiv:2001.06555*, 2020.

Severi Rissanen and Pekka Marttinen. A critical look at the consistency of causal estimation with deep latent variable models. *Advances in Neural Information Processing Systems*, 34:4207–4217, 2021.

Pedro HC Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122, 2020.

Matthew Schlegel, Wesley Chung, Daniel Graves, Jian Qian, and Martha White. Importance resampling for off-policy prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4), August 2020. ISSN 0090-5364. doi: 10.1214/19-aos1875.

Abhin Shah, Raaz Dwivedi, Devavrat Shah, and Gregory W Wornell. On counterfactual inference with unobserved confounding. *arXiv preprint arXiv:2211.08209*, 2022.

Chengchun Shi, Runzhe Wan, Rui Song, Wenbin Lu, and Ling Leng. Does the markov decision process fit the data: Testing for the markov property in sequential decision making. In *International Conference on Machine Learning*, pages 8807–8817. PMLR, 2020.

Chengchun Shi, Runzhe Wan, Victor Chernozhukov, and Rui Song. Deeply-debiased off-policy interval estimation. In *International conference on machine learning*, pages 9580–9591. PMLR, 2021.

Chengchun Shi, Masatoshi Uehara, Jiawei Huang, and Nan Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *International Conference on Machine Learning*, pages 20057–20094. PMLR, 2022a.

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

Chengchun Shi, Jin Zhu, Shen Ye, Shikai Luo, Hongtu Zhu, and Rui Song. Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, pages 1–12, 2022c.

Kang Shuai, LAn Liu, Yangbo He, and Wei Li. Mediation pathway selection with unmeasured mediator-outcome confounding. *arXiv preprint arXiv:2311.16793*, 2023.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. *Advances in neural information processing systems*, 26, 2013.

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

Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*, 2019.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.

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

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

Dustin Tran and David M Blei. Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*, 2017.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline RL: PAC bounds and posterior sampling under partial coverage. *CoRR*, abs/2107.06226, 2021.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

Victor Veitch, Yixin Wang, and David Blei. Using embeddings to correct for unobserved confounding in networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR, 2020.

Hao Wang, Xingjian Shi, and Dit-Yan Yeung. Relational deep learning: A deep latent variable model for link prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Jiayi Wang, Zhengling Qi, and Chengchun Shi. Blessing from experts: Super reinforcement learning in confounded environments. *arXiv preprint arXiv:2209.15448*, 2022.

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

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. The deconfounded recommender: A causal inference approach to recommendation. *arXiv preprint arXiv:1808.06581*, 2018.

Chuhan Xie, Wenhao Yang, and Zhihua Zhang. Semiparametrically efficient off-policy evaluation in linear markov decision processes. In *International Conference on Machine Learning*, pages 38227–38257. PMLR, 2023.

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

Yang Xu, Jin Zhu, Chengchun Shi, Shikai Luo, and Rui Song. An instrumental variable approach to confounded off-policy evaluation. In *International Conference on Machine Learning*, pages 38848–38880. PMLR, 2023.

Mengxin Yu, Zhuoran Yang, and Jianqing Fan. Strategic decision-making in the presence of information asymmetry: Provably efficient rl with algorithmic instruments. *arXiv preprint arXiv:2208.11040*, 2022.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

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

Linying Zhang, Yixin Wang, Anna Ostropolets, Jami J Mulgrave, David M Blei, and George Hripcsak. The medical deconfounder: assessing treatment effects with electronic health records. In *Machine Learning for Healthcare Conference*, pages 490–512. PMLR, 2019.

Wenzhuo Zhou, Yuhan Li, Ruoqing Zhu, and Annie Qu. Distributional shift-aware off-policy interval estimation: A unified error quantification framework. *arXiv preprint arXiv:2309.13278*, 2023a.

Yunzhe Zhou, Zhengling Qi, Chengchun Shi, and Lexin Li. Optimizing pessimism in dynamic treatment regimes: A bayesian learning approach. In *International Conference on Artificial Intelligence and Statistics*, pages 6704–6721. PMLR, 2023b.

Yunzhe Zhou, Chengchun Shi, Lexin Li, and Qiwei Yao. Testing for the markov property in time series via deep conditional generative learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1204–1222, 2023c.

A Discussion

A.1 Critiques of Wang and Blei [2019]

In this section, we discuss the critiques of the paper Wang and Blei [2019] and illustrate how our proposal addresses these criticisms.

Specifically, Wang and Blei [2019] imposed the "consistency of substitute confounders" assumption which requires that the unmeasured confounders Z_i can be consistently estimated from the causes A_i. However, this assumption indeed invalidates the derivations of Theorems 6-8 of Wang and Blei [2019], resulting in the inconsistency of the algorithm. Specifically, as commented by D'Amour [2019], if the event $A = a$ provides a perfect measurement of Z such that there is some function $\widehat{z}(A)$ such that $\widehat{z}(a) = Z$, then the overlap condition fails. As a result, the ATE cannot be consistently identified. Ogburn et al. [2019] expressed similar concerns towards the algorithm.

Unlike Wang and Blei [2019], our algorithm does not require such an assumption. Under the RL setting, the proposed two-way unmeasured confounding assumption effectively limits the number of unmeasured confounders to $\mathcal{O}(N) + \mathcal{O}(T)$, which facilitates their consistent estimation when both the number of trajectories N and the number of decision points per trajectory T grow to infinity, avoiding the need for the unmeasured confounders to be deterministic functions of the actions.

A.2 Limitations

Although the proposed two-way unmeasured confounding assumption is more flexible than the existing one-way unmeasured confounding, it might not adequately handle more complex scenarios involving latent confounders that are both trajectory- and time-specific, or policy-dependent. In practice, this assumption can be partially evaluated by testing whether including the estimated two-way confounders in the observation leads to a Markovian system [see e.g., Chen and Hong, 2012, Shi et al., 2020, Zhou et al., 2023c].

A.3 Future work

Our proposal requires the existence of an observation which, along with the two-way confounders, blocks all backdoor paths between the treatment and both the immediate reward and future observations. Naïvely using pre-treatment variables as observations might lead to biased estimators in certain causal models like the M-graph Cinelli et al. [2022]. While confounder selection algorithms [Guo et al., 2023] are designed to address the problem, they mainly focus on the bandit setting. Our future work is to extend these algorithms to RL to further improve the practical applicability of our method.

B Proofs

B.1 Notations

We first introduce list the notations that will be used in our proof.

- $R(a,o,u,w) := \sum_{r,o'} r\mathcal{P}(r,o'|a,o,u,w)$, which denotes the expectation of reward given observation o, action a and latent factors (u,w), and $\widehat{R}$ denotes its estimator.
- $P(o'|a,o,u,w) := \sum_r \mathcal{P}(r,o'|a,o,u,w)$ denotes the observation transition function, which calculates the probability of transitioning from observation o to observation o' given the action a and latent factors (u,w), and $\widehat{P}$ denotes its estimator.
- $\rho_0(\bullet)$ denotes the probability mass function of $(O_{i,1},U_i)$.
- $d_{\bar{\mathcal{D}}}$ denotes the data distribution of quadruples (a,o,u,w), given by $d_{\bar{\mathcal{D}}}(a,o,u,w) = T^{-1}\sum_{t=1}^{T}\mathbb{P}(O_{1,t}=o,A_{1,t}=a,U_1=u,W_t=w)$.
- d^π denotes the distribution of quadruples (a,o,u,w) in a trajectory of horizon T generated under π, i.e., $d^\pi(a,o,u,w) = T^{-1}\sum_{t=1}^{T}\mathbb{P}^\pi(O_{1,t}=o,A_{1,t}=a,U_1=u,W_t=w)$.
- $\mathrm{TV}(\widehat{\mathcal{P}}(\bullet|a,o,u,w),\mathcal{P}(\bullet|a,o,u,w)) := \sum_{r,o'}|\widehat{\mathcal{P}}(r,o'|a,o,u,w)-\mathcal{P}(r,o'|a,o,u,w)|/2$ denotes the total variation between $\widehat{\mathcal{P}}(\bullet|a,o,u,w)$ and $\mathcal{P}(\bullet|a,o,u,w)$.
- $\|\widehat{\mathcal{P}}-\mathcal{P}\|_{d_{\bar{\mathcal{D}}}} := \sqrt{\mathbb{E}_{(A,O,U,W)\sim d_{\bar{\mathcal{D}}}}\mathrm{TV}^2(\widehat{\mathcal{P}}(\bullet|A,O,U,W),\mathcal{P}(\bullet|A,O,U,W))}$ is used to measure the average estimation error of $\widehat{\mathcal{P}}$ on the given offline dataset.

Additionally, we will use a generic constant c in the proof, whose value is allowed to vary from place to place.

B.2 Proof of Proposition 1

Proof. Notice that the MSE of the predicted reward-next-observation pair is equal to the sum of the MSE of the predicted immediate reward and that of the predicted next observation. Consequently, it suffices to show the MSE of the predicted immediate reward remains constant.

Toward that end, we consider the following linear model under unconstrained unmeasured confounding,

$$R_{i,t} = (O_{i,t},A_{i,t})\zeta + Z_{i,t} + \varepsilon_{i,t},$$

where $Z_{i,t}$s are one-dimensional, $\zeta = (\zeta_1,\zeta_2)^\top = \mathbb{R}^2$ denotes the coefficient vector, and $\varepsilon_{i,t}$ are mean zero i.i.d. measurement errors with each $\varepsilon_{i,t}$ being independent of the set of variables $\mathcal{H}_{i,t} = \{(O_{i',t'},A_{i',t'},Z_{i',t'}) : i' \neq i \text{ or } t' \leq t\}$.

In vector and matrix notations, we define $Y = (R_{1,1},\ldots,R_{1,T},\ldots R_{N,2},\ldots,R_{N,T})^\top$ as the (NT)-dimensional outcome vector, $\beta = (\zeta_1,\zeta_2,Z_{1,1},\ldots,Z_{1,T},\ldots,Z_{N,1},\ldots,z_{N,T})^\top$ as the $(NT+2)$-dimensional coefficient vector, $\mathcal{E} = (\varepsilon_{1,1},\ldots,\varepsilon_{1,T},\ldots,\varepsilon_{N,1},\ldots,\varepsilon_{N,T})^\top \in \mathbb{R}^{NT}$ as the set of residuals and

$$X = \begin{pmatrix} O_{1,1} & A_{1,1} & 1 & 0 & \cdots & 0 \\ O_{1,2} & A_{1,2} & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \ddots & \cdots \\ O_{N,T} & A_{N,T} & 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{NT\times(NT+2)}.$$

Thus, we can formulate the linear model as $Y = X\beta + \mathcal{E}$. Since the LSE $\widehat{\beta}$ is an unbiased estimator, its predictor $\widehat{Y} = X\widehat{\beta}$'s MSE can be calculated as follows,

$$\begin{aligned}\text{MSE}(\widehat{Y}) &= \frac{1}{NT}\mathbb{E}\|\widehat{Y} - \mathbb{E}(Y|X\beta)\|_2^2 = \frac{1}{NT}\mathbb{E}\|X(X^\top X)^- X^\top \mathcal{E}\|_2^2 \\ &= \frac{1}{NT}\mathbb{E}[\mathcal{E}^\top X(X^\top X)^- X^\top \mathcal{E}] = \frac{1}{NT}\mathbb{E}\Big\{\text{trace}[X(X^\top X)^- X^\top \mathcal{E}\mathcal{E}^\top]\Big\} \\ &= \frac{\sigma^2}{NT}\mathbb{E}\Big\{\text{trace}[X(X^\top X)^- X^\top]\Big\} = \sigma^2,\end{aligned} \tag{2}$$

where the conditional expectation $\mathbb{E}(Y|X\beta)$ is taken in a componentwise manner, $(X^\top X)^-$ denotes the generalized inverse of $X^\top X$ and $\sigma^2 = \text{Var}(\varepsilon_{i,t})$. Here, the second last equation holds due to the independence between $\varepsilon_{i,t}$ are $\mathcal{H}_{i,t}$. Consequently, the MSE is a fixed constant. This completes the proof. □

B.3 Proof of Proposition 2

Proof. Similar to the proof of Proposition 1, it suffices to prove that the MSE of the predicted immediate reward remains a constant. Toward that end, we assume the immediate reward satisfies the following linear two-way unmeasured confounding assumption,

$$R_{i,t} = (O_{i,t}, A_{i,t})\zeta + U_i + W_t + \varepsilon_{i,t}.$$

However, one-way unmeasured confounding assumes only the existence of trajectory-specific unmeasured confounders and ignores time-specific confounders, leading to the following model

$$R_{i,t} = (O_{i,t}, A_{i,t})\zeta + H_i + \varepsilon_{i,t}.$$

Because we are more concerned with estimating fixed effects and also to simplify subsequent symbols, we consider a scenario that is easier to estimate. Let $\widehat{\zeta}$ denote the LSE of ζ based on the one-way model. Since $X^\top \widehat{Y} = X^\top X(X^\top X)^{-1}X^\top Y = X^\top Y$, the LSE of H_i satisfies

$$\widehat{H}_i = \frac{1}{T}\sum_{t=1}^{T}[R_{i,t} - (O_{i,t}, A_{i,t})\widehat{\zeta}] = U_i + \frac{1}{T}\sum_{t=1}^{T}[W_t - (O_{i,t}, A_{i,t})(\widehat{\zeta} - \zeta)].$$

Therefore, the MSE of the predicted $\widehat{R}_{i,t} = (O_{i,t}, A_{i,t})\widehat{\zeta} + \widehat{H}_i$ is given by

$$\frac{1}{NT}\sum_{i,t}\mathbb{E}[(O_{i,t}, A_{i,t})\widehat{\zeta} + \widehat{H}_i - (O_{i,t}, A_{i,t})\zeta - U_i - W_t]^2 = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(W_t - \frac{1}{T}\sum_{t'=1}^{T}W_{t'})^2.$$

For a stationary and ergodic time series $\{W_t\}_t$, $T^{-1}\sum_{t'=1}^{T}W_{t'}$ is to converge to $\mathbb{E}(W_t)$ as $T \to \infty$ and thus, the above equation is to converge to the variance of W_t. Consequently, unless each W_t is degenerate, the MSE will not decay to zero. This completes the proof. □

B.4 Proof of Proposition 3

Proof. To simplify the proof, we only show the MSE of predicted immediate reward decays to zero as both N and T grow to infinity. The two-way model can be similarly expressed as $Y = X\beta + \mathcal{E}$, where $Y = (R_{1,1}, \ldots, R_{1,T}, \ldots R_{N,1}, \ldots, R_{N,T})^\top \in \mathbb{R}^{NT}$, $\beta = (\zeta_1, \zeta_2, U_1, \ldots, U_N, W_1, \ldots, W_T)^\top \in \mathbb{R}^{N+T+2}$, $\mathcal{E} = (\varepsilon_{1,1}, \ldots, \varepsilon_{1,T}, \ldots \varepsilon_{N,1}, \ldots, \varepsilon_{N,T})^\top \in \mathbb{R}^{NT}$, and

$$X = \begin{pmatrix} O_1 & A_1 & \mathbf{1}_T & \mathbf{0}_T & \cdots & \mathbf{0}_T & I_T \\ O_2 & A_2 & \mathbf{0}_T & \mathbf{1}_T & \cdots & \mathbf{0}_T & I_T \\ \cdots & \cdots & \cdots & \cdots & \ddots & \cdots & \cdots \\ O_N & A_N & \mathbf{0}_T & \mathbf{0}_T & \cdots & \mathbf{1}_T & I_T \end{pmatrix} \in \mathbb{R}^{NT\times(N+T+2)},$$

where $\mathbf{1}_T = (1, \ldots, 1)^\top \in \mathbb{R}^T$, $\mathbf{0}_T = 0 \cdot \mathbf{1}_T$, and I_T represents the $T \times T$ identity matrix.

Similar to Equation (2), the MSE of the predicted reward is equal to

$$\text{MSE}(\widehat{Y}) = \frac{1}{NT}\mathbb{E}[\mathcal{E}^\top X(X^\top X)^{-1}X^\top \mathcal{E}] = \sigma^2\frac{\text{trace}(X^\top X)}{NT} = \frac{\sigma^2(N+T+2)}{NT},$$

which decays to zero as both N and T increase to infinity. The proof is hence completed. □

B.5 Proof of Theorem 1

To simplify the proof, we analyze a variant of our estimator computed via sample-splitting and cross-fitting. Specifically, we begin by dividing the entire dataset $\mathcal{D}$ into two subsets $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$, each containing $N/2$ trajectories. Next, we separately apply the proposed two-way deconfounder algorithm detailed in Section 2 to the two data subsets to learn the transition functions and latent confounders. Denote them by $\widehat{\mathcal{P}}^{(j)}$, $\{\widehat{U}_i^{(j)}\}_i$ and $\{\widehat{W}_t^{(j)}\}_t$, respectively, for $j = 1, 2$. Finally, we construct two model-based estimators, given by

$$\widehat{\eta}^{(1)} = \frac{2}{NT} \sum_{i \in \mathcal{D}^{(1)}} \sum_{t=1}^{T} \widehat{\mathbb{E}}^{(2)} \Big(R_{i,t} \mid O_{i,1}, \widehat{U}_i^{(1)}, \Big\{ \widehat{W}_{t'}^{(1)} \Big\}_{t'=1}^{t} \Big), \tag{3}$$

and

$$\widehat{\eta}^{(2)} = \frac{2}{NT} \sum_{i \in \mathcal{D}^{(2)}} \sum_{t=1}^{T} \widehat{\mathbb{E}}^{(1)} \Big(R_{i,t} \mid O_{i,1}, \widehat{U}_i^{(2)}, \Big\{ \widehat{W}_{t'}^{(2)} \Big\}_{t'=1}^{t} \Big), \tag{4}$$

where $\widehat{\mathbb{E}}^{(j)}$ denotes the expectation that uses the transition function $\mathcal{P}^{(j)}$ to approximate $\mathbb{E}^{\pi}$, and the summation $\sum_{i \in \mathcal{D}^{(j)}}$ is carried over all trajectories in $\mathcal{D}^{(j)}$. We average them construct our final estimator $\widehat{\eta}^{\pi} = (\widehat{\eta}^{(1)} + \widehat{\eta}^{(2)})/2$.

Notice that in both (3) and (4), the transition function used to conduct the model-based estimator is independent of the initial observation and the estimated latent confounders. This avoids imposing additional VC-class type conditions on the transition network. We remark that sample splitting is widely used in statistics and machine learning [see e.g., Chernozhukov et al., 2018, Shi et al., 2021, Kallus and Uehara, 2022].

Proof. We focus on bounding $\mathbb{E}|\widehat{\eta}^{(1)} - \eta^{\pi}|$. The same upper bound applies to $\mathbb{E}|\widehat{\eta}^{(2)} - \eta^{\pi}|$ and $\mathbb{E}|\widehat{\eta}^{\pi} - \eta^{\pi}|$, thanks to the triangle inequality. To ease notation, we will remove the superscripts in $\widehat{\mathbb{E}}^{(2)}$, $\widehat{\mathcal{P}}^{(2)}$, $\widehat{U}_i^{(1)}$, $\widehat{W}_t^{(1)}$, and present them as $\widehat{\mathbb{E}}$, $\widehat{\mathcal{P}}$, $\widehat{U}_i$, $\widehat{W}_t$. Due to the use of cross-fitting, $\widehat{\mathbb{E}}$ and $\widehat{\mathcal{P}}$ are independent of $\widehat{U}_i$s and $\widehat{W}_t$s. The proof is divided into two steps. In the first step, we upper bound $\mathbb{E}|\widehat{\eta}^{(1)} - \overline{\eta}^{(1)}|$ where

$$\overline{\eta}^{(1)} = \frac{2}{NT} \sum_{i \in \mathcal{D}^{(1)}} \sum_{t=1}^{T} \mathbb{E}^{\pi}(R_{i,t} | O_{i,1}, U_i, \{W_{t'}\}_{t'=1}^{t}).$$

Next, in the second step, we upper bound $\mathbb{E}|\overline{\eta}^{(1)} - \eta^{\pi}|$. Finally, combining these two bounds leads to the upper bound for $\mathbb{E}|\widehat{\eta}^{(1)} - \eta^{\pi}|$.

Step 1. Recall that

$$\widehat{\eta}^{(1)} = \frac{2}{NT} \sum_{i \in \mathcal{D}^{(1)}} \sum_{t=1}^{T} \widehat{\mathbb{E}}(R_{i,t} | O_{i,1}, \widehat{U}_i, \{\widehat{W}_{t'}\}_{t'=1}^{t}).$$

Notice that the difference $\overline{\eta}^{(1)} - \widehat{\eta}^{(1)}$ can be decomposed into the sum of

$$\begin{aligned} & \frac{2}{NT} \sum_{i \in \mathcal{D}^{(1)}} \sum_{t=1}^{T} \Big[\mathbb{E}^{\pi}(R_{i,t} | O_{i,1}, U_i, \{W_{t'}\}_{t'=1}^{t}) - \widehat{\mathbb{E}}(R_{i,t} | O_{i,1}, U_i, \{W_{t'}\}_{t'=1}^{t}) \Big] \\ + \quad & \frac{2}{NT} \sum_{i \in \mathcal{D}^{(1)}} \sum_{t=1}^{T} \Big[\widehat{\mathbb{E}}^{\pi}(R_{i,t} | O_{i,1}, U_i, \{W_{t'}\}_{t'=1}^{t}) - \widehat{\mathbb{E}}(R_{i,t} | O_{i,1}, \widehat{U}_i, \{\widehat{W}_{t'}\}_{t'=1}^{t}) \Big] \\ = \quad & I_1 + I_2. \end{aligned}$$

We first study I_1. Since $O_{i,1}$s and U_is are independent of W_ts, the first line forms a sum of i.i.d. random variables. Thus, it converges to

$$\begin{aligned} I_1^* := \frac{1}{T} \sum_{t=1}^{T} \sum_{o,u} \Big[\mathbb{E}^{\pi}(R_{i,t} | O_{i,1} = o, U_i = u, \{W_{t'}\}_{t'=1}^{t}) - \widehat{\mathbb{E}}(R_{i,t} | O_{i,1} = o, U_i = u, \{W_{t'}\}_{t'=1}^{t}) \Big] \\ \times \rho_0(o, u) = \frac{1}{T} \sum_{t=1}^{T} \sum_{o,u} I_{1,t}^*(o, u, \{W_{t'}\}_{t'=1}^{t}) \rho_0(o, u), \end{aligned}$$

with the expected error $\mathbb{E}|I_1 - I_1^*|$ upper bounded by $cR_{\max}/\sqrt{N}$ for some constant $c > 0$ by Cauchy-Schwarz inequality.

It remains to study I_1^*, or equivalently, $I_{1,t}^*$ for each t. To begin with, consider the case where $t = 1$. It follows that

$$\begin{aligned}|I_{1,1}^*(o,u,W_1)| &= \sum_a \pi(a|o)[R(a,o,u,W_1) - \widehat{R}(a,o,u,W_1)] \\ &= \sum_{a,r,o'} \pi(a|o) r[\mathcal{P}(r,o'|a,o,u,W_1) - \widehat{\mathcal{P}}(r,o'|a,o,u,W_1)] \\ &\leq R_{\max} \sum_a \pi(a|o) \text{TV}(\widehat{\mathcal{P}}(\bullet|a,o,u,W_1), \mathcal{P}(\bullet|a,o,u,W_1))\end{aligned}$$

When $t = 2$, one can show that

$$|I_{1,2}^*(o_1,u,\{W_t\}_{t=1}^2)| \leq R_{\max} \sum_{a_1} \pi(a_1|o_1) \text{TV}(\widehat{\mathcal{P}}(\bullet|a_1,o_1,u,W_1), \mathcal{P}(\bullet|a_1,o_1,u,W_1))$$

$$+R_{\max} \sum_{a_2,o_2,a_1} \pi(a_2|o_2)\pi(a_1|o_1) \text{TV}(\widehat{\mathcal{P}}(\bullet|a_2,o_2,u,W_2), \mathcal{P}(\bullet|a_2,o_2,u,W_2)) P(o_2|a_1,o_1,u,W_1).$$

Using the same argument, one can show that

$$\begin{aligned}|I_{1,t}^*(o_1,u,\{W_{t'}\}_{t'=1}^t)| \leq R_{\max} \sum_{t'=1}^t \sum_{a_{t'},o_{t'},\cdots,a_1} \pi(a_t'|o_t') \prod_{j=1}^{t'-1} \Big[\pi(a_j|o_j) P(o_{j+1}|a_j,o_j,u,W_j)\Big] \\ \times \text{TV}(\widehat{\mathcal{P}}(\bullet|a_{t'},o_{t'},u,W_{t'}), \mathcal{P}(\bullet|a_{t'},o_{t'},u,W_{t'})).\end{aligned}$$

As such, we obtain

$$\begin{aligned}&\sum_{t=1}^T \sum_{o,u} \mathbb{E}|I_{1,t}^*(o,u,\{W_{t'}\}_{t'=1}^t)|\rho_0(o,u) \\ \leq\ & R_{\max} \sum_{t=1}^T \mathbb{E}^\pi[\text{TV}(\widehat{\mathcal{P}}(\bullet|A_{1,t},O_{1,t},U_1,W_t), \mathcal{P}(\bullet|A_{1,t},O_{1,t},U_1,W_t))].\end{aligned}$$

Using the change of measure theorem, the right-hand-side can be upper bounded by

$$CTR_{\max}\mathbb{E}[\text{TV}(\widehat{\mathcal{P}}(\bullet|A_{1,t},O_{1,t},U_1,W_t), \mathcal{P}(\bullet|A_{1,t},O_{1,t},U_1,W_t))],$$

where C denotes the single-policy concentration coefficient defined in Assumption 1. By Cauchy-Schwarz inequality, the above expression can be further bounded from above by $CTR_{\max}\varepsilon_{\mathcal{P}}$.

Using similar arguments, we can upper bound $|I_2|$ by

$$\frac{R_{\max}}{N} \sum_{i=1}^N \sum_{t=1}^T \max_{a,o} [\text{TV}(\widehat{\mathcal{P}}(\bullet|a,o,\widehat{U}_i,\widehat{W}_t), \widehat{\mathcal{P}}(\bullet|a,o,U_i,W_t))].$$

As neural networks are Lipschitz continuous functions of their parameters, the above total variation norm is proportional to $\|(\widehat{U}_i^\top, \widehat{W}_t^\top)^\top - (U_i^\top, W_t^\top)\|_2$. Consequently, under Assumption 4, the above expression is upper bounded by $cTR_{\max}\varepsilon_{U,W}$ for some constant $c > 0$. This completes the proof of Step 1.

Step 2. According to the DGP described in Section 2, $(U_1, O_{1,1})$, $\cdots$, $(U_n, O_{n,1})$ are i.i.d., η^π, and are independent of $\{W_t\}_t$. η^π can thus be represented by $T^{-1}\sum_{t=1}^T \mathbb{E}^\pi(R_{1,t})$. We further decompose $|\overline{\eta}^{(1)} - \eta^\pi|$ into the following two components:

$$\begin{aligned}|\overline{\eta}^{(1)} - \eta^\pi| =& \Big|\frac{2}{NT} \sum_{i\in\mathcal{D}^{(1)}} \sum_{t=1}^T \mathbb{E}^\pi(R_{i,t}|O_{i,1},U_i,\{W_{t'}\}_{t'=1}^t) - \frac{1}{T}\sum_{t=1}^T \mathbb{E}^\pi(R_{1,t})\Big| \\ \leq& \Big|\frac{2}{NT} \sum_{i\in\mathcal{D}^{(1)}} \sum_{t=1}^T \mathbb{E}^\pi(R_{i,t}|O_{i,1},U_i,\{W_{t'}\}_{t'=1}^t) - \frac{1}{T}\sum_{t=1}^T \mathbb{E}^\pi(R_{1,t}|\{W_{t'}\}_{t'=1}^t)\Big| \\ &+ \Big|\frac{1}{T}\sum_{t=1}^T \mathbb{E}^\pi(R_{1,t}|\{W_{t'}\}_{t'=1}^t) - \frac{1}{T}\sum_{t=1}^T \mathbb{E}^\pi(R_{1,t})\Big| = I_3 + I_4.\end{aligned}$$

Conditional on $\{W_t\}_t$, I_3 corresponds to a sum of i.i.d. random variables. An application of Cauchy-Schwarz inequality implies that $\mathbb{E}I_3 \leq \sqrt{\mathbb{E}I_3^2} \leq cR_{\max}/\sqrt{N}$, for some constant $c > 0$. Similarly, under Assumption 5, we have

$$\mathbb{E}I_4 \leq \sqrt{\mathbb{E}I_4^2} \leq \frac{R_{\max}}{T}\sqrt{\sum_{1\leq t_1,t_2\leq T}\rho(t_1,t_2)} = O(R_{\max}T^{\alpha-1}).$$

The proof is hence completed by combining the results in both steps. □

C Implementation.

C.1 Details for loss function

The final loss is defined as follows:

$$\begin{aligned} L\left(\mathcal{D};\{u_i\}_i,\{w_t\}_t\right) =& (1-\alpha)\cdot\frac{1}{2}\sum_{i,t}\left\{\left[\widehat{\mu}-\varphi_{i,t}\right]^\top\widehat{\Sigma}^{-1}\left[\widehat{\mu}-\varphi_{i,t}\right]+\log\det\widehat{\Sigma}\right\} \\ &+\alpha\cdot\sum_{i,t}\mathrm{CrossEntropy}\left(a_{i,t},\widehat{\pi}_b(a_{i,t}|o_{i,t},u_i,w_t)\right), \end{aligned}$$

where loss weighting $\alpha \in (0,1)$ is a hyperparameter, $\varphi_{i,t} = \left(r_{i,t}, o_{i,t+1}^\top\right)^\top$, and $(a_{i,t}, o_{i,t}, r_{i,t}, o_{i,t+1})$ is sampled from the offline dataset $\mathcal{D}$, $\widehat{\Sigma} = \mathrm{diag}(\widehat{\sigma})$, $\det\widehat{\Sigma}$ represents the determinant of $\widehat{\Sigma}$, and $\mathrm{CrossEntropy}$ is the cross-entropy loss.

C.2 Implementation Details for Two-way Deconfounder Model.

The Two-way Deconfounder Model described in Section 3 was implemented in Pytorch and trained on an NVIDIA GeForce RTX 3090. For every task, we repeat it 20 times and the results are aggregated over 20 runs. The training time of 4 to 12 hours hours (depending on the task) on a single NVIDIA GeForce RTX 3090. Each dataset undergoes a 75/25 split for training, validation respectively. Using the validation set, we perform hyperparameter optimization using many iterations of grid search to find the optimal values for hyperparameters. The search range for each hyperparameter is described as follows, learning rate $lr \in [0.005, 0.001]$, Batch size $bs \in [2^8, 2^9, 2^{10}, 2^{11}, 2^{12}]$, weight decay $\lambda \in [0.01, 0.0001]$, two-way embedding dimension $d_{tw} \in [2^1, 2^2, 2^3]$, Loss weighting $\alpha \in [0.0, 0.3, 0.5, 0.7]$.

C.3 Implementation Details for Ablation Study.

Variants of TWD are model-based methods, model-based method with two-way embedding(TWD-TO,TWD-MLP) and model-based method with one-way embedding(TWD-NI,TWD-NT)Yu et al. [2020], Kausik et al. [2022], where the first can be applied to the two-way setting , and the second is originally designed for the one-way case. We adopt the same embedding approach for variants of TWD as our method. TWD-TO can be seen as a special case of our method, which only remove the CrossEntropy loss of the behavior policy. TWD-MLP remove the neural tensor network. However, the biggest difference between TWD-NI or TWD-NT and our method is that it use the one-way embedding based on Kausik et al. [2022] removing the embedding of trajectory-invariant or time-invariant unmeasured confounders. We provide Change details of those variants compared to TWD in the following.

- TWD-TO: Because it removes the CrossEntropy loss of the behavior policy, the loss function is as follows,

  $$L\left(\mathcal{D};\{u_i\}_i,\{w_t\}_t\right) = \sum_{i,t}[\left\{\left[\widehat{\mu}-\varphi_{i,t}\right]^\top\widehat{\Sigma}^{-1}\left[\widehat{\mu}-\varphi_{i,t}\right]+\log\det\widehat{\Sigma}\right\}],$$

  other model settings are the same as TWD.
- TWD-MLP: Because it replace the neural tensor network with a simple MLP , d-dimensional embedding vector representing the joint information of observation and latent confounders

is as follows, $e(o, u_i, w_t) = f(g(M\tau_{i,t} + b))$, where $\tau_{i,t}$ is shorthand for $[o_{i,t}^\top, u_i^\top, w_t^\top]^\top$, f is a standard linear layer, g is the ReLU activation function, $M \in \mathbb{R}^{k\times(d_o+2d)}$, d_o represents the dimension of observation o, and $b \in \mathbb{R}^k$ is the bias term.

- TWD-NI: Because it remove the individual embedding , d-dimensional embedding vector representing the joint information of observation and latent confounders is as follows, $e(o, w_t) = f(g(M[o_{i,t}^\top, w_t^\top]^\top + b))$, where f is a standard linear layer, g is the ReLU activation function, $M \in \mathbb{R}^{k\times(d_o+d)}$, d_o represents the dimension of observation o, and $b \in \mathbb{R}^k$ is the bias term.
- TWD-NT: Because it remove the time embedding , d-dimensional embedding vector representing the joint information of observation and latent confounders is as follows, $e(o, u_i) = f(g(M[o_{i,t}^\top, u_i^\top]^\top + b))$, where f is a standard linear layer, g is the ReLU activation function, $M \in \mathbb{R}^{k\times(d_o+d)}$, d_o represents the dimension of observation o, and $b \in \mathbb{R}^k$ is the bias term.

D Simulation details

D.1 A linear simulation setting

Let $\mathcal{O} = \mathbb{R}, \mathcal{U} = \mathbb{R}, \mathcal{W} = \mathbb{R}, \mathcal{R} = \mathbb{R}$, and $\mathcal{A} = \{1, 0\}$. For evey i and t, u_i and w_t are sampled a Normal distribution $\mathcal{N}(0, 1)$. The observed data is generated as follows,

The transition model. For each i and t, given $(O_{i,t}, A_{i,t}, U_i, W_t)$, We generate

$$O_{i,t+1} = 0.7O_{i,t} + A_{i,t} + 2U_i + 2W_t - 0.5 + e_s$$

where the random error $e_s \sim \mathcal{N}(0, 1)$

The reward model. For each i and t, given $(O_{i,t}, A_{i,t}, U_i, W_t)$, We generate

$$R_{i,t} = O_{i,t} + 3A_{i,t} + 2U_i + 2W_t + e_r$$

where the random error $e_r \sim \mathcal{N}(0, 2)$

Behavior policy. For each i and t, given $(O_{i,t}, U_i, W_t)$, We generate

$$p(A_{i,t} = 1 \mid O_{i,t}, U_i, W_t) = \text{sigmoid}(O_{i,t} + U_i + W_t)$$

Target policy. $p(A_{i,t} = 1) = 0.5$

We generate datasets respectively under the behavioral and target policy,. Under the assumptions of one-way unmeasured confounders, two-way unmeasured confounders, and unconstrained unmeasured confounders, respectively, we use the least squares method to fit the model parameters. Then, we use the learned model to perform the off-policy evaluation. The number of trajectory is $N = \{200, 300, 400, 500, 600, 700, 800\}$. The length of the horizon is fixed at 50. we calculate MSE error in fitting the observed data and MSE error in off-policy evaluation, respectively. Two-way unmeasured confounders outperformed the other two on off-policy evaluation, as shown in the Figure 1.

D.2 Simulated Dynamic Process

We first describe the detailed setting for the simulated data. Let $\mathcal{O} = \mathbb{R}^4, \mathcal{U} = \mathbb{R}, \mathcal{W} = \mathbb{R}, \mathcal{R} = \mathbb{R}$, and $\mathcal{A} = \{1, 0\}$. We let $T = 50$ and the number of data trajectories $N = \{250, 500, 1000, 1500, 2000\}$. We consider a four-dimensional variable $O_{i,t} = (O_{i,t}^1, O_{i,t}^2, O_{i,t}^3, O_{i,t}^4)$ whose initial distribution is given by $\mathcal{N}(\mathbf{0}_4, \mathbf{I}_4)$ where $\mathbf{I}_4$ denotes a four-dimensional identity matrix. The individual-invariant unmeasured confounders $\{U_i\}_i$ are sampled a Normal distribution $\mathcal{N}(0, 1)$. The time-invariant unmeasured confounders $\{W_t\}_t$ are sampled a Normal distribution $\mathcal{N}(0, 1)$. The data generating process is outlined as follows, reward function and transition function are given by $R_{i,t} = f_r(O_{i,t}, A_{i,t}, U_i, W_t, \Gamma) + e_r$, $e_r \sim \mathcal{N}(0, 1)$, and $O_{i,t+1} = f_o(O_{i,t}, A_{i,t}, U_i, W_t, \Gamma) + e_o$, $e_o \sim \mathcal{N}(\mathbf{0}_4, \mathbf{I}_4)$, the behavior policy satisfies $\mathbb{P}(A_{i,t} = 1 \mid O_{i,t}, U_i, W_t) = \text{sigmoid}(f_a(O_{i,t}, U_i, W_t, \Gamma))$. Γ is the confounding strength parameter. We set $\Gamma = 1$. $f_o(\cdot)$, $f_r(\cdot)$ and $f_a(\cdot)$ are the functions of input variable on the next observation, immediate reward, and action, respectively.

The transition model. For each i and t, given $(O_{i,t}, A_{i,t}, U_i, W_t)$ and Γ, We generate

$$O_{i,t+1} = \mu_s O_{i,t} + \Gamma \cdot \beta_o f\left(U_i, W_t\right) + \lambda_o A_{i,t} \mathbf{I}_4 + b_o + e_o$$

where $\mathbf{I}_4 = [1,1,1,1]^\top$, the random error $e_o \sim \mathcal{N}\left(\mathbf{0}_4, \mathbf{I}_4\right)$ with $\mathbf{I}_4$ denoting the 4-by-4 identity matrix,

- $\mu_o = \begin{bmatrix} 0.8 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0.8 \end{bmatrix}$,
- $\beta_o = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix}$,
- $\lambda_o = [1,1,1,1]^\top$,
- $b_o = [-0.5,-0.5,-0.5,-0.5]^\top$,
- $f_o\left(U_i, W_t\right) = [u_i - w_t, u_i + w_t, -u_i - w_t, -u_i + w_t]^\top$.

The reward model. For each i and t, given $(O_{i,t}, A_{i,t}, U_i, W_t)$ and Γ, We generate

$$R_{t,t} = \mu_r O_{t,t} + \Gamma \cdot \beta_r f\left(U_i, W_t\right) + \lambda_r A_{i,t} + e_r$$

where the random error $e_r \sim \mathcal{N}\left(0,1\right)$, $\beta_r = 3.0$, $\lambda_r = 2.5$,

- $\mu_r = [0.25, 0.25, 0.25, 0.25]^\top$,
- $f_r\left(U_i, W_t\right) = u_i \cdot w_t$.

Behavior policy. For each i and t, given $(O_{i,t}, U_i, W_t)$ and Γ, We generate

$$p\left(A_{i,t} = 1 \mid O_{i,t}, U_i, W_t\right) = \text{sigmoid}\left(\mu_a O_{i,t} - 4 + \Gamma \cdot \beta_a f_a\left(U_i, W_t\right)\right)$$

where $\beta_a = 3$,

- $\mu_a = [0.25, 0.25, 0.25, 0.25]^\top$,
- $f_r\left(U_i, W_t\right) = u_i \cdot w_t + u_i + w_t$.

Target policy.

- Target policy A: $p(A_{i,t} = 1) = 0.3$.
- Target policy B: $p(A_{i,t} = 1) = 0.5$.

D.3 Tumor Growth simulated data setup

In this section, we give brief description about Tumor Growth environment setting and experiment details. We set up 10 subsets randomly with different numbers of patients who received treatments sequentially over 60 stages. We treat each patient as a time-invariant confounding variable to add the heterogeneous effect. To be specific, we randomized patients into three groups, with each patient having a group label $S_i \in \{1,2,3\}$.

The transition model. We apply the unobserved confounding variables to the PK-PD model:

- $V(t) = (1 + \rho \log\left(\frac{K}{V(t-1)}\right) - \beta_c C(t) - \left(\alpha A^r(t) + \beta A^r(t)^2\right) + e_t) V(t-1)$. where ρ and K are model parameters sampled for each patient according to prior distributions in Geng et al. [2017] and $e_t \sim N\left(0, 0.01^2\right)$. Specifically, β_c, α and β are model parameters sampled for representing specific characteristics which affect with patient's response to chemotherapy and radiotherapy according to randomly subclassing patients into 3 different groups $S_i \in \{1,2,3\}$ in Geng et al. [2017], which are influenced by genetic factors [Bartsch et al., 2007] and can be regarded as time-invariant confounders.

- $C(t) = C(t-1)/2 + 5 \cdot A^c(t)$, which relies on the chemo treatment action and exponentially decays over time.

The reward model. The reward model consists of three parts:

- *pathological-health reward*: A negative reward R_n for penalising the patient if tumor size is large or accepts too much drugs and radio: $1.5\exp(-V(t))$.
- *side-effect reward*: A positive reward R_p for accepting treatments when take action at time step t: $\exp(-(A^r(t) + A^c(t))^2)$
- *confounded reward*: A reward R_{TW}about the two-way confounder parameter:$\Gamma \cdot (4 \cdot sigmoid(S_i - 2) - \sin(0.1\pi t))$.
- *total reward*: $R = R_n + R_P + R_{TW} + e_r, e_r \sim \mathcal{N}(0, 1)$

Behavior policy. We introduce two-way unmeasured confounders to treatment assignment policy. For each i and t, We generate,

$$A^c(t) \sim Bernoulli(\text{sigmoid}(\frac{\gamma_c}{D_{\max}}(D(t) - \theta_c) + \Gamma \cdot (3 \cdot sigmoid(S_i - 2) - 0.75 \cdot \sin(0.1\pi t))))$$

$$A^r(t) \sim Bernoulli(\text{sigmoid}(\frac{\gamma_r}{D_{\max}}(D(t) - \theta_r) + \Gamma \cdot (3 \cdot sigmoid(S_i - 2) - 0.75 \cdot \sin(0.1\pi t))))$$

where $D(t)$ is tumor diameters, $D_{\max}$is the half maximum size of the tumor, θ_c and θ_r are fixed such that $\theta_c = \theta_r = D_{\max}/2$ and γ_c and γ_r are also fixed such that $\gamma_c = \gamma_r = 10.0$.

Target policy.

- Target policy A: $A^c(t), A^r(t) \sim Bernoulli(0.05)$.
- Target policy B: $A^c(t), A^r(t) \sim Bernoulli(p)$, where p is related to the conditions of patients:

$$p = \begin{cases} 0.2, & if\ Y_t > 88, \\ 0.1, & if\ 44 < Y_t \leq 88, \\ 0.05, & if\ 5 < Y_t \leq 44, \\ 0.01, & otherwize \end{cases} \tag{5}$$

D.4 A real-world example: MIMIC-III.

We show how the Two-way Deconfounder can be applied to a real-world medical setting using the Medical Information Mart for Intensive Care (MIMIC-III) database [Johnson et al., 2016]. We extract 3,707 patients with trajectories up to 20 timesteps. Following the analysis of Zhou et al. [2023b], we define a 5 × 5 action space by discretizing both vasopressors fluid A^v and intravenous A^i fluid interventions into 5 levels. We define the reward $R_{i,t} = SOFA_{i,t} - SOFA_{i,t+1}$ as the difference between current SOFA score and next SOFA score, so a lower reward indicates a higher risk of mortality. We extract 12 patient covariates as observation space, including Glasgow coma scale,Systolic blood pressure, Weight, SOFA, SIRS, Fraction of inspired oxygen, Blood Urea Nitrogen, Creatinine, Serum Glutamic-Oxaloacetic Transaminase, Partial Pressure of Oxygen , Total Bilirubin, Platelet Count. Considering the complexity of the real data, we incorporate the normalization step before modeling them.

Target policy.

- the randomized policy: $P(A^v = 0) = P(A^v = 1) = p(A^v = 2) = P(A^v = 3) = p(A^v = 4) = 0.2$, $P(A^i = 0) = P(A^i = 1) = p(A^i = 2) = P(A^i = 3) = p(A^i = 4) = 0.2$.
- the high dose polic: $P(A^v = 3) = p(A^v = 4) = 0.5$, $P(A^i = 3) = p(A^i = 4) = 0.5$.
- the low dose polic: $P(A^v = 0) = 0.4, P(A^v = 1) = p(A^v = 2) = 0.3$, $P(A^i = 0) = 0.4, P(A^i = 1) = p(A^i = 2) = 0.3$.
- the tailored policy:
  - if normalized SOFA score $\geq$ 0.95, $p(A^v = 4) = p(A^v = 3) = 0.3$, $p(A^v = 2) = 0.2$ and $P(A^v = 0) = P(A^v = 1) = 0$; $p(A^i = 4) = p(A^i = 3) = 0.3$, $p(A^i = 2) = 0.2$ and $P(A^i = 0) = P(A^i = 1) = 0$.

- if $0.95 >$ normalized SOFA score ≥ 0.7, $p(A^v = 4) = p(A^v = 3) = 0.1, p(A^v = 2) = 0.2, p(A^v = 1) = p(A^v = 0) = 0.3$;$p(A^i = 4) = p(A^i = 3) = 0.1, p(A^i = 2) = 0.2, p(A^i = 1) = p(A^i = 0) = 0.3$.
- if normalized SOFA score < 0.7, $p(A^v = 0) = 0.8, p(A^v = 1) = 0.2 and p(A^v = 2) = p(A^v = 1) = p(A^v = 0) = 0$; $p(A^i = 0) = 0.8, p(A^i = 1) = 0.2 and p(A^i = 2) = p(A^i = 1) = p(A^i = 0) = 0$.

D.5 Sensitivity and Limitation analysis

In three simulated datasets, given $Z_{i,t} \sim N\left(0, \sigma_z^2\right)$, the modified reward model and behavior policy according to the additive noise model are as follows,

Simulated Dynamic Process.

- The reward model: For each i and t, given $(O_{i,t}, A_{i,t}, U_i, W_t)$ and Γ, We generate

$$R_{t,t} = \mu_r O_{t,t} + \Gamma \cdot \beta_r f\left(U_i, W_t\right) + (1 - \Gamma) \cdot Z_{i,t} + \lambda_r A_{i,t} + e_r$$

  where the random error $e_r \sim \mathcal{N}(0, 1)$, $\beta_r = 3.0$, $\lambda_r = 2.5$,
  - $\mu_r = [0.25, 0.25, 0.25, 0.25]^\top$,
  - $f_r\left(U_i, W_t\right) = u_i \cdot w_t$.
- Behavior policy: For each i and t, given $(O_{i,t}, U_i, W_t)$ and Γ, We generate

$$p\left(A_{i,t} = 1 \mid O_{i,t}, U_i, W_t\right) = \text{sigmoid}\left(\mu_a O_{i,t} - 4 + \Gamma \cdot \beta_a f_a\left(U_i, W_t\right) + (1 - \Gamma) \cdot Z_{i,t}\right)$$

  where $\beta_a = 3$,
  - $\mu_a = [0.25, 0.25, 0.25, 0.25]^\top$,
  - $f_r\left(U_i, W_t\right) = u_i \cdot w_t + u_i + w_t$.

Tumor Growth.

- The reward model: $R = R_n + R_P + R_{TW} + (1 - \Gamma) \cdot Z_{1,t} + e_r, e_r \sim \mathcal{N}(0, 1)$
- Behavior policy: For each i and t, We generate,

$$\begin{aligned} A^c(t) \sim & Bernoulli(\text{sigmoid}(\frac{\gamma_c}{D_{\max}}\left(D(t) - \theta_c\right) + (1 - \Gamma) \cdot Z_{i,t} \\ & + \Gamma \cdot (3 \cdot sigmoid(S_i - 2) - 0.75 \cdot \sin(0.1\pi t)))) \\ A^r(t) \sim & Bernoulli(\text{sigmoid}(\frac{\gamma_r}{D_{\max}}\left(D(t) - \theta_r\right) + (1 - \Gamma) \cdot Z_{i,t} \\ & \Gamma \cdot (3 \cdot sigmoid(S_i - 2) - 0.75 \cdot \sin(0.1\pi t)))) \end{aligned}$$

  where $D(t)$ is tumor diameters, $D_{\max}$is the half maximum size of the tumor, θ_c and θ_r are fixed such that $\theta_c = \theta_r = D_{\max}/2$ and γ_c and γ_r are also fixed such that $\gamma_c = \gamma_r = 10.0$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: At the last part of the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions are provided in Section 2 and Section 4, while the specific proofs of theoretical results are shown in Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The information needed to reproduce the main experimental results are shown in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data is subject to an agreement and cannot be shared due to its sensitivity, but it is publicly available at `https://physionet.org/content/mimiciii/1.4/`. The code is provided in Section 5.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the details about experiment in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide these information in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide these information in Appendix C.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This paper merely proposes a new method for off-policy evaluation, and there are no negative social impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.