Causal Eligibility Traces for Confounding Robust Off-Policy Evaluation

Junzhe Zhang[1] **Elias Bareinboim**[2]

[1]Department of Electrical Engineering and Computer Science, Syracuse University
[2]Department of Computer Science, Columbia University

Abstract

A unifying theme in Artificial Intelligence is learning an effective policy to control an agent in an unknown environment in order to optimize a certain performance measure. Off-policy methods can significantly improve sample efficiency during training, since they allow an agent to learn from observed trajectories generated by different *behavior policies*, without directly deploying *target policies* in the underlying environment. This paper studies off-policy evaluation from biased offline data where (1) *unobserved confounding* bias cannot be ruled out a priori; or (2) the observed trajectories do not *overlap* with intended behaviors of the learner, i.e., the target and behavior policies do not share a common support. Specifically, we extend Bellman's equation to derive effective closed-form bounds over value functions from the observational distribution contaminated with unobserved confounding and no overlap. Second, we propose two novel algorithms that use eligibility traces to estimate these bounds from finite observational data. Compared to other methods for robust off-policy evaluation in sequential environments, these methods are model-free and extend, for the first time, the well-celebrated temporal difference algorithms (Sutton, 1988) to biased offline data with unobserved confounding and no overlap.

1 INTRODUCTION

A typical reinforcement learning agent learns from past data, i.e., from observed trajectories of states, actions, and reward signals generated by the agent intervening in the underlying environment. This data reflects the influence of the decision-making policy used to allocate actions based on the observed state, which is called the *behavior policy*. This policy might be selected by the agent in the past or by a different demonstrator operating in the same environment. *Policy evaluation* studies the problem of evaluating the effectiveness of a candidate *target policy* from the combination of past data and theoretical assumptions about the environment. When the behavior and target policies coincide, the evaluation is called *on-policy* learning, in which the expected return of candidate policies given the agent's starting state (i.e., the value function) could be directly estimated with empirical means [Sutton and Barto, 1998]. In practice, however, the learner might have to learn about policies different from the currently deployed one that generated the data, leading to the *off-policy* learning problem.

Off-policy learning is a popular area of research, as it allows for more efficient learning by using data from different policies. Several algorithms have been proposed for off-policy evaluation from finite observations, including Q-learning [Watkins, 1989, Watkins and Dayan, 1992], importance sampling [Swaminathan and Joachims, 2015, Jiang and Li, 2016], and temporal difference [Precup et al., 2000, Munos et al., 2016]. These algorithms rely on two critical assumptions about the behavior policy. First, no unobserved confounder affects the behavior policy's selected action and the subsequent state and reward. Second, the behavior policy is stochastic, covering all intended actions the target policy selects given all observed states. When either of these assumptions does not hold, the effect of the target policy is generally not *identifiable*, i.e., the model assumptions are insufficient to uniquely determine the value function from the offline data [Pearl, 2000, Zhang and Bareinboim, 2019].

In recent times, researchers have been using partial identification methods to obtain reliable off-policy evaluation in situations where there are unobserved confounders, and the behavior and target policies have no common support [Kallus and Zhou, 2018, Zhang and Bareinboim, 2019, Kallus and Zhou, 2020, Namkoong et al., 2020, Khan et al., 2023, Bruns-Smith and Zhou, 2023, Kausik et al., 2024]. Partial identification is a well-studied problem in causal inference [Balke and Pearl, 1997, Zhang et al., 2022, Zhang

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:4933–4942.

and Bareinboim, 2021], econometrics [Imbens and Rubin, 1997, Poirier, 1998, Romano and Shaikh, 2008, Stoye, 2009, Bugni, 2010, Todem et al., 2010, Moon and Schorfheide, 2012], and dynamical systems [Bajari et al., 2007, Norets and Tang, 2014, Dickstein and Morales, 2018, Morales et al., 2019, Berry and Compiani, 2023]. It enables the derivation of informative bounds on target effects from confounded observational data. Several model-based algorithms have been proposed, which estimate the underlying system dynamics from offline data based on a combination of conditions and constraints. These include (1) the marginal sensitivity model that assumes access to a bound over the odds ratio between the nominal and actual behavioral policies [Kallus and Zhou, 2018, 2020, Namkoong et al., 2020, Khan et al., 2023, Bruns-Smith and Zhou, 2023]; (2) parametric knowledge about the system dynamics (i.e., reward function and transition distribution) are invoked under which informative bounds are derived [Kausik et al., 2024]; (3) the decision horizon is finite, i.e., the agent only determines a finite number of actions [Kallus and Zhou, 2018, Zhang and Bareinboim, 2019, Namkoong et al., 2020, Khan et al., 2023, Kausik et al., 2024]. We refer readers to the complete technical report [Zhang and Bareinboim, 2024, Appendix A] for a more detailed survey.

This paper contributes to this growing line of literature by studying model-free algorithms for robust off-policy evaluation over an infinite horizon from confounded offline data generated by a behavioral policy with no overlap support. We propose novel partial identification algorithms using eligibility traces to obtain informative bounds over the expected return of candidate policies from offline data generated from an unknown Markov decision process where the unobserved confounders exist, and overlap does not hold.

More specifically, our contributions are summarized as follows. (1) We extend the Bellman equation that permits one to derive optimal bounds over target value functions from the observational distribution generated by an unknown behavior policy. (2) We propose a novel off-policy temporal difference algorithm (`C-TD(`λ`)`) using eligibility traces to estimate bounds over the state value function from finite observations contaminated with unobserved confounding and no overlap. (3) We introduce an alternative eligibility trace algorithm following tree backup (`C-TB(`λ`)`) that obtains bounds over the state-action value function from biased observations. Finally, we evaluate our proposed algorithms using extensive simulations in synthetic environments. All proofs and details of the experiment setup are provided in the technical report [Zhang and Bareinboim, 2024].

Notations. We use capital letters to denote random variables (X), small letters for their values (x) and $\mathcal{X}$ for the domain of X. For an arbitrary set $\boldsymbol{X}$, let $|\boldsymbol{X}|$ be its cardinality. Fix indices $i, j \in \mathbb{N}$. Let $\bar{\boldsymbol{X}}_{i:j}$ stand for a sequence $\{X_i, X_{i+1}, \dots, X_j\}$. We denote by $P(\boldsymbol{X})$ a probability distribution over variables $\boldsymbol{X}$. Similarly, $P(\boldsymbol{Y} \mid \boldsymbol{X})$ represents a set of conditional distributions $P(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})$ for all realizations $\boldsymbol{x}$. We consistently use $P(\boldsymbol{x})$ as abbreviations of probabilities $P(\boldsymbol{X} = \boldsymbol{x})$; so does $P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}) = P(\boldsymbol{y} \mid \boldsymbol{x})$. Finally, $\mathbb{1}_{\boldsymbol{Z}=\boldsymbol{z}}$ is an indicator function that returns 1 if event $\boldsymbol{Z} = \boldsymbol{z}$ holds true; otherwise, it returns 0.

2 CHALLENGES OF CAUSAL INCONSISTENCY

We will focus on a sequential decision-making problem in the Markov Decision Process (MDP, Puterman [1994]) where the agent intervenes on a sequence of actions to optimize certain rewards/primary outcomes. The standard MDP formalism focuses on the perspective of learners who could actively intervene in the environment. Consequently, the data collected from randomized experiments are not contaminated with unobserved confounding bias, which is generally assumed away in the model. However, when considering offline data collected by passive observation, the learner may not necessarily have deliberate control over the behavioral policy generating the data. Consequently, this could lead to confounding bias in various decision-making tasks [Kallus and Zhou, 2018, Zhang et al., 2020, Kumor et al., 2021, Guo et al., 2022, Ruan et al., 2024]. In this paper, we will consider an extended family of MDPs that explicitly models the presence of unobserved confounders.

Definition 1. A Confounded Markov Decision Process (CMDP) $\mathcal{M}$ is a tuple of $\langle \mathcal{S}, \mathcal{X}, \mathcal{Y}, \mathcal{U}, \mathcal{F}, P \rangle$ where (1) $\mathcal{S}, \mathcal{X}, \mathcal{Y}$ are, respectively, the space of observed states, actions, and rewards; (2) $\mathcal{U}$ is the space of unobserved exogenous noise; (3) $\mathcal{F}$ is a set consisting of the transition function $f_S : \mathcal{S} \times \mathcal{X} \times \mathcal{U} \mapsto \mathcal{S}$, behavioral policy $f_X : \mathcal{S} \times \mathcal{U} \mapsto \mathcal{X}$, and reward function $f_Y : \mathcal{S} \times \mathcal{X} \times \mathcal{U} \mapsto \mathcal{Y}$; (4) P is an exogenous distribution over the domain $\mathcal{U}$.

Throughout this paper, we will consistently assume the state-action domain $\mathcal{X} \times \mathcal{S}$ to be finite; the reward domain $\mathcal{Y}$ is bounded in a real interval $[a, b] \subset \mathbb{R}$.

Consider a demonstrator agent interacting with a CMDP $\mathcal{M}$, generating the offline data. For every time step $t = 1, \dots, T$, the nature first draws an exogenous noise U_t from the distribution $P(\mathcal{U})$; the demonstrator then performs an action $X_t \leftarrow f_X(S_t, U_t)$, receives a subsequent reward $Y_t \leftarrow r_t(S_t, X_t, U_t)$, and moves to the next state $S_{t+1} \leftarrow f_S(S_t, X_t, U_t)$. The observed trajectories of the demonstrator (from the learner's perspective) are thus summarized as the observational distribution $P(\bar{\boldsymbol{X}}_{1:T}, \bar{\boldsymbol{S}}_{1:T}, \bar{\boldsymbol{Y}}_{1:T})$, i.e.,

$$P(\bar{\boldsymbol{x}}_{1:T}, \bar{\boldsymbol{s}}_{1:T}, \bar{\boldsymbol{y}}_{1:T}) = P(s_1) \prod_{t=1}^{T} \Bigg(\int_{\mathcal{U}} \mathbb{1}_{s_{t+1}=f_S(s_t,x_t,u_t)} \mathbb{1}_{x_t=f_X(s_t,u_t)} \mathbb{1}_{y_t=f_Y(s_t,x_t,u_t)} P(u_t) \Bigg)$$

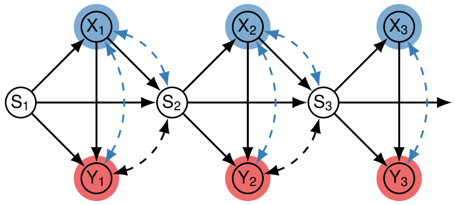


Figure 1: Causal diagram representing the data-generating mechanisms in a Markov Decision Process (MDP)

Fig. 1 shows the causal diagram $\mathcal{G}$ [Bareinboim et al., 2022] describing the generative process generating the offline data in CMDPs. More specifically, solid nodes represent observed variables X_t, S_t, Y_t, and arrows represent the functional relationships f_X, f_S, f_Y among them. By convention, exogenous variables U_t are often not explicitly shown; bi-directed arrows $X_t \longleftrightarrow Y_t$ and $X_t \longleftrightarrow S_{t+1}$ indicate the presence of an unobserved confounder (UC) U_t affecting the action, state, and reward simultaneously. These bi-directed arrows (highlighted in blue) characterize the spurious correlations among action X_t, reward Y_t, and state S_{t+1} in the offline data, violating the condition of no unmeasured confounding (NUC, Robbins [1985]). Such violations could lead to challenges in off-policy evaluation, which we will discuss for the remainder of this section.

Off-Policy Evaluation. A policy π in a CMDP $\mathcal{M}$ is a decision rule $\pi : \mathcal{S} \mapsto \mathcal{X}$ mapping from state to action. Similarly, $\pi(x_t \mid s_t)$ is a stochastic policy mapping from state space $\mathcal{S}$ to a distribution over action space $\mathcal{X}$. An intervention do(π) is an operation that replaces the behavioral policy f_X in CMDP $\mathcal{M}$ with the policy π. Let $\mathcal{M}_\pi$ be the submodel induced by intervention do(π). The interventional distribution $P_\pi(\bar{\boldsymbol{X}}_{1:T}, \bar{\boldsymbol{S}}_{1:T}, \bar{\boldsymbol{Y}}_{1:T})$ is defined as the joint distribution over observed variables in $\mathcal{M}_\pi$, i.e.,

$$P_\pi(\bar{\boldsymbol{x}}_{1:T}, \bar{\boldsymbol{s}}_{1:T}, \bar{\boldsymbol{y}}_{1:T}) = P(s_1) \prod_{t=1}^{T} \Big(\pi(x_t \mid s_t) \\ \mathcal{T}(s_t, x_t, s_{t+1}) \mathcal{R}(s_t, x_t, y_t) \Big) \tag{1}$$

where the transition distribution $\mathcal{T}$ and the reward distribution $\mathcal{R}$ are given by, for $t = 1, \dots, T$,

$$\mathcal{T}(s_t, x_t, s_{t+1}) = \int_{\mathcal{U}} \mathbb{1}_{s_{t+1} = f_S(s_t, x_t, u_t)} P(u_t) \tag{2}$$

$$\mathcal{R}(s_t, x_t, y_t) = \int_{\mathcal{U}} \mathbb{1}_{y_t = f_Y(s_t, x_t, u_t)} P(u_t) \tag{3}$$

For convenience, we write the reward function $\mathcal{R}(s, x)$ as the expected value $\sum_y y \mathcal{R}(s, x, y)$.

Fix a discounted factor $\gamma \in [0, 1]$. A common objective for an agent is to optimize its cumulative return $R_t = \sum_{i=0}^{\infty} \gamma^i Y_{t+i}$. In analysis, we often evaluate the state value function $V_\pi(s)$, which is the expected return given the agent's starting state $S_t = s$. That is, $V_\pi(s) = \mathbb{E}_\pi [R_t \mid S_t = s]$. A similar state-action value function $Q_\pi(s, x)$ is defined as the expected return starting from state s, taking action x and thereafter following policy π, i.e., $Q_\pi(s, x) = \mathbb{E}_{X_t \leftarrow x, \pi} [R_t \mid S_t = s]$. One could recursively evaluate the value function of any state s using the *Bellman Equation* [Bellman, 1966] given by,

$$V_\pi(s) = \sum_x \pi(x \mid s) \Big(\mathcal{R}(s, x) + \gamma \sum_{s'} \mathcal{T}(s, x, s') V_\pi(s') \Big) \tag{4}$$

Similarly, an analogous equation for the state-action value function is given by

$$Q_\pi(s, x) = \mathcal{R}(s, x) + \gamma \sum_{s'} \mathcal{T}(s, x, s') V_\pi(s') \tag{5}$$

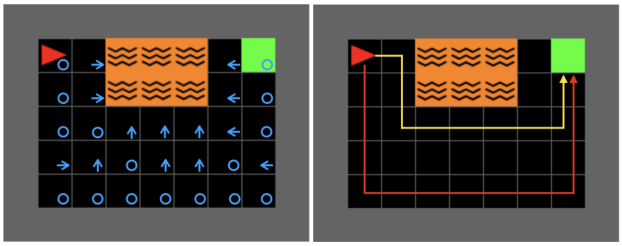
In off-policy evaluation, the agent (i.e., learner) attempts to estimate the effects of a candidate policy $\pi(x|s)$ from the observational data generated by the behavior policy f_X (demonstrator). Standard off-policy methods focus on the identifiable setting where the transition distribution $\mathcal{T}$ and reward function $\mathcal{R}$ remain consistent in both the interventional P_π and observational distribution P. Formally,

Definition 2 (Causal Consistency)**.** For a CMDP $\mathcal{M}$, the Causal Consistency holds if the following statement holds, for every time step $t = 1, 2, \dots,$

$$\begin{aligned} \mathcal{T}(s_t, x_t, s_{t+1}) &= P(s_{t+1} \mid s_t, x_t), \\ \mathcal{R}(s_t, x_t, y_t) &= P(y_t \mid s_t, x_t) \end{aligned} \tag{6}$$

When Causal Consistency holds, the learner could recover the parametrization of the transition distribution $\mathcal{T}$ and reward function $\mathcal{R}$ from the observational data, following the identification formula in Eq. (6). Several off-policy algorithms have been proposed to estimate the effect of candidate policies from finite observations under causal consistency [Watkins, 1989, Watkins and Dayan, 1992, Swaminathan and Joachims, 2015, Jiang and Li, 2016, Precup et al., 2000, Munos et al., 2016]. There exist graphical criteria in the literature [Pearl and Robins, 1995, Shpitser et al., 2010, Perković et al., 2015] to evaluate whether causal consistency (Def. 2) holds from causal knowledge of the environment, including the celebrated *backdoor* criterion [Pearl, 2000, Def. 3.3.1].

However, in many practical applications, causal consistency could be fragile and does not necessarily hold due to some violations in the generative process. These include: (1) there exists an unobserved confounder affecting the action X_t and subsequent outcomes Y_t, S_{t+1} simultaneously (blue, dashed arrows in Fig. 1); (2) there is no overlap in the support between the target and behavior policies, i.e., the propensity score $P(x_t \mid s_t) = 0$ for some state-action pair s_t, x_t. When either of these violations occurs, applying standard off-policy methods may fail to recover the expected return of the target policy, leading to estimation bias. The following example illustrates such challenges.



(a) (b)

Figure 2: A windy gridworld environment where the red arrow represents the agent and green square is the goal state; the agent can take five actions - `up`, `down`, `right`, `left`, and `stay-put`; the wind can blow in five directions - `north`, `south`, `west`, `east`, and `no-wind`. The agent attempts to reach the goal without stepping into lava.

Example 1. Consider a Windy Gridworld described in Fig. 2, which we adapted from one of Deepmind's AI safety Gridworlds [Leike et al., 2017]. The red triangle represents the agent, and the green square represents the goal state. The agent can take five actions X_t - `up`, `down`, `right`, `left`, and `stay-put`. The agent receives a constant reward $Y_t \leftarrow 1$ if it reaches the goal state. On the other hand, the task fails, and the agent receives no reward ($Y_t \leftarrow 0$) if it steps into the lava (orange tiles) on its way.

Additionally, the agent's movement is affected by the wind direction U_t, which could take five values at each time-step, including - `east`, `south`, `west`, `north`, and `no-wind`. The distribution of the wind direction depends on the agent's location. As an example, Fig. 2a shows samples of wind directions for every position at a single time step. In general, the wind tempts to push the agent toward the lava; the closer the agent gets to the lava, the stronger the wind becomes. If the agent decides to move (i.e., $X_t \leftarrow$ `up`, `down`, `right`, `left`), its next state of the agent is shifted by both its action and the wind direction through the mechanism $S_{t+1} \leftarrow S_t + X_t + U_t$. Otherwise, the agent will stay put ($X_t \leftarrow$ `stay-put`) at its current position, regardless of the wind direction, i.e., $S_{t+1} \leftarrow S_t$.

Figs. 3a to 3b shows the value function estimation obtained by standard off-policy methods, including temporal difference with importance sampling [Precup et al., 2000], and tree backup [Sutton and Barto, 1998]. For comparison, we also include in Fig. 3c the actual value function computed from the ground-truth model using value iteration. The simulation reveals that standard off-policy evaluation deviates from the ground truth return. In this case, the demonstrator will only move if there is no wind, which makes the shorter path appear less risky than it actually is. The wind direction U_t is thus an unobserved confounder affecting both the action X_t and next state S_{t+1} in the offline data, violating causal consistency. See the complete technical report [Zhang and Bareinboim, 2024, Appendix C] for additional discussions on the windy Gridworld environment.

2.1 PARTIAL CAUSAL IDENTIFICATION IN CONFOUNDED MDPS

This section will introduce partial identification methods for off-policy evaluation that is robust to the unobserved confounding and no overlap. For every time step $t = 1, 2, \ldots$, let the reward Y_t be bounded in a real interval $[a, b]$. By applying a similar bounding strategy in [Manski, 1990, Zhang and Bareinboim, 2019, Joshi et al., 2024], we derive the following bounds over the transition probability distribution $\mathcal{T}$ for every realization $(s, x, s') \in \mathcal{S} \times \mathcal{X} \times \mathcal{S}$,

$$\begin{aligned} \mathcal{T}(s,x,s') &\geq \widetilde{\mathcal{T}}(s,x,s')\, P(x \mid s) \\ \mathcal{T}(s,x,s') &\leq \widetilde{\mathcal{T}}(s,x,s')\, P(x \mid s) + P(\neg x \mid s) \end{aligned} \tag{7}$$

where $P(x \mid s) = P(X_t = x \mid S_t = s)$ and $P(\neg x \mid s) = 1 - P(x \mid s)$; and $\widetilde{\mathcal{T}}$ is the nominal transition distribution computed from the observational distribution as $\widetilde{\mathcal{T}}(s,x,s') = P(S_{t+1} = s' \mid S_t = s, X_t = x)$. Similarly, one could also derive the following bound over the reward function $\mathcal{R}$ for every state-action pair $(s, x) \in \mathcal{S} \times \mathcal{X}$,

$$\begin{aligned} \mathcal{R}(s,x) &\geq \widetilde{\mathcal{R}}(s,x)\, P(x \mid s) + aP(\neg x \mid s), \\ \mathcal{R}(s,x) &\leq \widetilde{\mathcal{R}}(s,x)\, P(x \mid s) + bP(\neg x \mid s) \end{aligned} \tag{8}$$

where $\widetilde{\mathcal{R}}$ is the nominal reward function given by $\widetilde{\mathcal{R}}(s,x) = \mathbb{E}[Y_t \mid S_t = s, X_t = x]$.

To bound the value function $V_\pi(s)$ at state s induced by a candidate policy π, one could minimize/maximize the optimization program using the Bellman's equation in Eq. (4) as the objective function, subject to constraints in Eqs. (7) and (8). Interestingly, this optimization problem is equivalent to a linear program; solving it leads to the following *extended Bellman equation*.

Theorem 1 (Causal Bellman Equation). *For a CMDP $\mathcal{M}$ with reward domain $\mathcal{Y} = [a, b]$, for any policy $\pi(x \mid s)$, its state value function $V_\pi(s) \in \left[\underline{V_\pi}(s), \overline{V_\pi}(s)\right]$ for every state $s \in \mathcal{S}$, where the lower bound $\underline{V_\pi}$ is a solution given by the following dynamic program,*

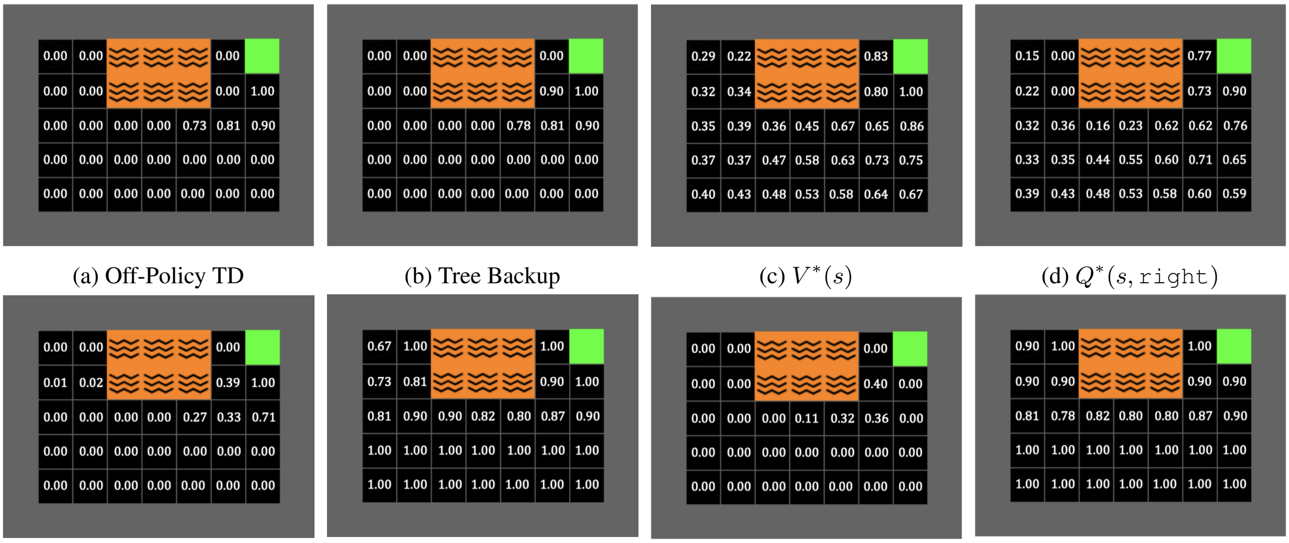
$$\underline{V_\pi}(s) = \sum_x P(x \mid s)\Bigg(\pi(\neg x \mid s)\Big(a + \gamma \min_{s'} \underline{V_\pi}(s')\Big) \tag{9}$$

$$+\pi(x \mid s)\Big(\widetilde{\mathcal{R}}(s,x) + \gamma \sum_{s',x'} \widetilde{\mathcal{T}}(s,x,s')\, \underline{V_\pi}(s')\Big)\Bigg) \tag{10}$$

Similarly, the upper bound $\overline{V_\pi}$ is a solution given by

$$\overline{V_\pi}(s) = \sum_x P(x \mid s)\Bigg(\pi(\neg x \mid s)\Big(b + \gamma \max_{s'} \overline{V_\pi}(s')\Big) \tag{11}$$

$$+\pi(x \mid s)\Big(\widetilde{\mathcal{R}}(s,x) + \gamma \sum_{s',x'} \widetilde{\mathcal{T}}(s,x,s')\, \overline{V_\pi}(s')\Big)\Bigg) \tag{12}$$



(a) Off-Policy TD

(b) Tree Backup

(c) $V^*(s)$

(d) $Q^*(s, \texttt{right})$

(e) $\underline{V^*}(s)$

(f) $\overline{V^*}(s)$

(g) $\underline{Q^*}(s, \texttt{right})$

(h) $\overline{Q^*}(s, \texttt{right})$

Figure 3: (a - b) Value function estimation obtained by standard off-policy methods; (c - d) The ground-truth value function computed from the underlying model; (e - h) Lower and upper bounds on the value functions obtained by causally enhanced off-policy algorithms using eligibility traces (C-TD(λ) and C-TB(λ))

Thm. 1 can be seen as an extension of the Bellman equation using the confounded observational distribution with no overlap. For instance, in the lower bound $\underline{V_\pi}(s)$, Eq. (9) could be thought as a regularizing term measuring the uncertainty due to unobserved confounding; Eq. (10) follows the standard iterative step in Bellman equation in Eq. (4), measuring the expected return when the target policy's action coincides with the observed action selected by the behavior policy. Finally, both terms are weighted by the nominal propensity score $P(x \mid s)$. The same derivation also applies to the upper bound $\overline{V_\pi}(s)$. An analogous extended Bellman equation bounding the state-action value function from the observational distribution can also be derived as follows.

Theorem 2 (Causal Bellman Equation)**.** *For a CMDP $\mathcal{M}$ with reward domain $\mathcal{Y} = [a, b]$, for any policy $\pi(x \mid s)$, its state-action value function $Q_\pi \in \left[\underline{Q_\pi}(s, x), \overline{Q_\pi}(s, x)\right]$ for any state-action pair $(s, x) \in \mathcal{S} \times \mathcal{X}$, where bounds $\underline{Q_\pi}$ is a solution given by the following dynamic program,*

$$\underline{Q_\pi}(s, x) = P(\neg x \mid s)\Big(a + \gamma \min_{s'} \underline{V_\pi}(s')\Big) \tag{13}$$

$$+ P(x \mid s)\Big(\widetilde{\mathcal{R}}(s, x) + \gamma \sum_{s'} \widetilde{\mathcal{T}}(s, x, s')\, \underline{V_\pi}(s')\Big) \tag{14}$$

Similarly, the upper bound $\overline{Q_\pi}$ is a solution given by

$$\overline{Q_\pi}(s, x) = P(\neg x \mid s)\Big(b + \gamma \max_{s'} \overline{V_\pi}(s')\Big) \tag{15}$$

$$+ P(x \mid s)\Big(\widetilde{\mathcal{R}}(s, x) + \gamma \sum_{s'} \widetilde{\mathcal{T}}(s, x, s')\, \overline{V_\pi}(s')\Big) \tag{16}$$

Among the bounds in Thm. 2, Eq. (13) is a regularized term accounting for uncertainties when the intervention do(x) is not observed in the offline data; Eq. (14) is the standard iterative step of the Bellman equation in Eq. (5), weighted by the score $P(x \mid s)$. Since Thms. 1 and 2 are closed-form solutions of optimization programs and the observational constraints in Eqs. (7) and (8) are tight, the extended Bellman's equation bounds are optimal from offline data and Markov property. This means that these bounds cannot be further improved without additional assumptions.

3 CONFOUNDING ROBUST ELIGIBILITY TRACES

The extended causal Bellman equations described so far require one to have precise estimations for the full models of the nominal transition distribution $\widetilde{\mathcal{T}}$, reward function $\widetilde{\mathcal{R}}$, and the propensity score $P(x \mid s)$. However, in practice, the detailed parameterizations of these probability models are generally assumed to be unknown. The learner must recover them from finite samples drawn from the confounded observational distribution.

This section will introduce novel model-free algorithms, using eligibility traces [Sutton, 1988], to bound value functions from finite observational samples. We consider the episodic framework, where the agent interacts with the environment for repeated episodes $n = 1, 2, 3, \dots$; each episode contains a finite number of time steps $t = 1, 2, \dots, T_n$. At each episode, the environment starts at state s_1 following the initial distribution $P(S_1)$. At each time step t, taking the observed state s_t of the environment as input, the behavior policy selects an action x_t. In response to inter-

Algorithm 1 Causal Temporal Difference (C-TD(λ))

```
Require: Observational data $\mathcal{D}$ and a policy $\pi(x \mid s)$.
Update the eligibility traces for all state $s \in \mathcal{S}$,
    $e_t(s) = \gamma\lambda\pi(x_{t-1} \mid s_{t-1})e_{t-1}(s) + \mathbb{1}_{s=s_t}$
    where $\lambda \in [0, 1]$ is an eligibility trace decay factor.
Compute the temporal difference error
    $\delta_t = \pi(x_t \mid s_t)\,(y_t + \gamma V_t(s_{t+1})) + \pi(\neg x_t \mid s_t)\,(w + \gamma V_t(s^*)) - V_t(s_t)$
Update the value function $V_{t+1}(s) \leftarrow V_t(s) + \alpha e_t(s)\delta_t$
    for all state $s \in \mathcal{S}$.
```

vention do(x_t), the environment produces a subsequent reward y_t and moves to the next observed state s_{t+1}. If the next state s_{t+1} is *terminal*, the episode terminates at time step $T_n = t + 1$; the learner receives observational data $\{\bar{\boldsymbol{x}}_{1:T_n-1}, \bar{\boldsymbol{s}}_{1:T_n}, \bar{\boldsymbol{y}}_{1:T_n-1}\}$.

3.1 CAUSAL TEMPORAL DIFFERENCE

We first introduce a novel augmentation procedure on the celebrated temporal difference (TD, [Sutton, 1988, Precup et al., 2000]) that allows one to estimate the bounds over state value functions, which we call the *causal temporal difference* (C-TD). Fig. 4 shows the backup diagram illustrating the idea of our proposed algorithm. Similar to the standard off-policy TD, our algorithm will update the estimation of state value functions $\underline{V_\pi}, \overline{V_\pi}$ using the sampled trajectories of transitions in the observational data. It could use a finite number of n-step trajectories or the entire trajectory. Different from the standard off-policy TD, our proposed algorithm does not weight each step of the transition using importance sampling (or equivalently, inverse propensity weighting) since the true behavior policy f_X (propensity score) is not recoverable from the observational data. Instead, C-TD weights each transition using the target policy π and adjusts for the misalignment between the target and behavior policies using an overestimation/underestimation of value function at state s^*. Such s^* is set as the best-case state associated with the highest value in our current estimation when computing upper bounds and the worst-case state estimate for lower bounds.

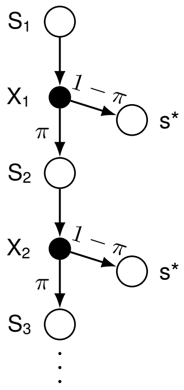


Figure 4: Backup diagram for C-TD(λ).

To formally introduce the estimation algorithm, we first introduce some necessary notations. Let $\boldsymbol{N}(s)$ denote the set of indices of episodes containing a state $s \in \mathcal{S}$, and let $\boldsymbol{t}_n(s)$ be the collection of time steps in the n-th episode such that for every $t \in \boldsymbol{t}_n(s)$, $s_t = s$. For any time step t, let $\pi_t = \pi(x_t \mid s_t)$ and $\neg\pi_t = 1 - \pi(x_t \mid s_t)$. We iteratively define the estimator for bounds over the state value function $V_\pi(s)$ as follows, for any state $s \in \mathcal{S}$,

$$\widehat{V_\pi}(s) = \frac{1}{N} \sum_{n \in \boldsymbol{N}(s)} \sum_{t \in \boldsymbol{t}_n(s)} \sum_{k=0}^{T_n - t} \gamma^k \Big(\pi_{t+k} y_{t+k} + \neg\pi_{t+k}\big(w + \gamma V(s^*)\big)\Big) \prod_{i=t}^{t+k-1} \pi_i \tag{17}$$

Among the above equations, N represents the total number of occurrences for the even $s_t = s$ in the observational data. we set parameters $w = a$ and $V(s^*) = \min_s V(s)$ when estimating the lower bound $\underline{V_\pi}(s)$; parameters $w = b$ and $V(s^*) = \max_s V(s)$ for the upper bound $\overline{V_\pi}(s)$.

An eligibility-trace version of our proposed estimation strategy is described Alg. 1. The algorithm keeps track of eligibility traces for every state in a similar manner to standard off-policy temporal difference algorithms. The main difference is that here the eligibility trace is multiplied by the target policy $\pi(x_{t-1} \mid s_{t-1})$ and a decay-rate λ, not including the nominal propensity score $P(x_{t-1} \mid s_{t-1})$. When computing the temporal difference error, the algorithm adjusts for the misalignment between the target and behavior policies by adding a regularized term $w + \gamma V_t(s^*)$, weighted by the probability $1 - \pi(x_t \mid s_t)$. We describe in Alg. 1 a version of C-TD(λ) using *online update*. This means that the bound estimates are updated at every time step. The *offline* version of the algorithm will use the same temporal difference error and eligibility traces. However, the update only occurs at the end of each episode; the increments and decrements are accumulated on the side, and the value function estimates do not change during the episode.

Theorem 3. *For any behavior policy, for any choice of* $\lambda \in [0, 1]$ *that does not depend on the actions chosen at each state, let parameters* w *and* s^* *be defined as follows: (1) Lower Bound* $\underline{V_\pi}$*:* $w = a$ *and* $s^* = \arg\min_s V_t(s)$*; (2) Upper Bound* $\overline{V_\pi}$*:* $w = b$ *and* $s^* = \arg\max_s V_t(s)$. *Then, Alg. 1 with offline updating converges with probability* 1 *to lower bound* $\underline{V_\pi}$ *and upper bound* $\overline{V_\pi}$*, respectively, under the usual step-size conditions on* α.

The proof of Thm. 3 first shows a contraction property for estimates $\widehat{V_\pi}$, and then follows the general convergence theorem in [Jaakkola et al., 1994].

3.2 CAUSAL TREE BACKUP

The algorithm described so far focuses on the estimation of the state value functions. We next introduce a novel al-

Algorithm 2 Causal Tree-Backup (`C-TB(`λ`)`)

```
Require: Observational data $\mathcal{D}$ and a policy $\pi(x|s)$.
Update the eligibility traces for all state-action pairs $s, x \in \mathcal{S} \times \mathcal{X}$:
    $e_t(s,x) = \gamma\lambda\pi(x_t \mid s_t)\mathbb{1}_{x_{t-1}=x}e_{t-1}(s,x) + \mathbb{1}_{s\neq s_t}$
    where $\lambda \in [0,1]$ is an eligibility trace decay factor.
Compute the temporal difference error for every action $x \in \mathcal{X}$. More specifically, if $x = x_t$,
    $\delta_t(x) = y_t + \gamma \sum_{x'} \pi(x \mid s_{t+1}) Q_t(s_{t+1}, x') - Q_t(s_t, x)$
    Otherwise,
    $\delta_t(x) = w + \gamma \sum_{x'} \pi(x' \mid s^*) Q_t(s^*, x') - Q_t(s_t, x)$
Update the action-value function $Q_{t+1}(s,x) \leftarrow Q_t(s,x) + \alpha e_t(s,x)\delta_t(x)$ for all $s, x \in \mathcal{S} \times \mathcal{X}$.
```

gorithm to bound the state-action value function Q_π from finite samples of the observational distribution.

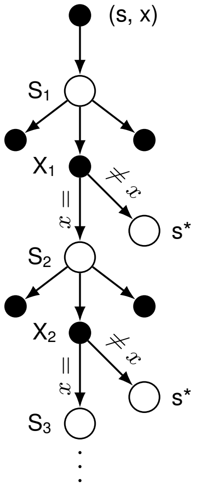
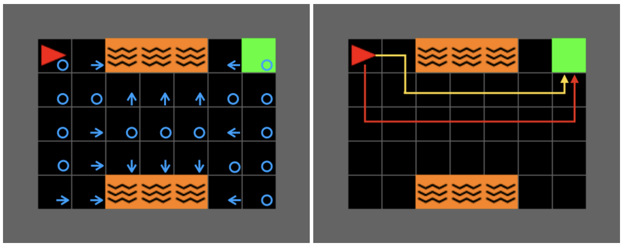


Figure 5: Backup diagram for `C-TB(`λ`)`.

Our algorithm is based on an augmentation on the standard tree backup (`TB` [Precup et al., 2000]), which we call the *causal tree backup* (`C-TB(`λ`)`). The main idea of this new algorithm is illustrated in the backup diagram of Fig. 5. Similar to the standard tree backup, our algorithm updates the value estimates for the action selected by the behavior policy at each time step based on the subsequent reward and the current estimation for the value of the next state. The algorithm then forms a new estimate for the target value function, using the old value estimates for the actions not observed in the observational data and the new estimated value for t-he action taken by the behavior policy. On the other hand, the main differences include the following. (1) Eligibility traces will not only be weighted by the target policy $\pi(x_t \mid s_t)$ using the observed trajectories, but also an indicator function $\mathbb{1}_{x_{t-1}=x}$ returning 1 if the previous action x_{t-1} coincides with the target action x. (2) When the behavior policy takes the same action $x_t = x$ as the target action, the update follows standard `TB` and uses the next sampled state s_t; when the sampled action $x_t \neq x$ differs from the target, our algorithm updates, instead, using the value function associated with the next worst-case or best-case state s^*, corresponding to the estimation of the lower and upper bounds respectively. The n-step causal tree-backup estimator is defined as



(a) (b)

Figure 6: An alternative windy gridworld environment where the lava is placed at both the top and bottom. The wind is the weakest at the center row of the map.

$$\begin{aligned}
\widehat{Q_\pi}(s,x) = {} & \frac{1}{N}\sum_{n\in \boldsymbol{N}(s)}\sum_{t\in \boldsymbol{t}_n(s)} \gamma^n Q(s_{t+n},x_{t+n}) \\
& \cdot \prod_{i=t}^{t+n-1}\pi_{i+1}\mathbb{1}_{x_i=x} + \sum_{k=t}^{t+n}\gamma^{k-t+1}\prod_{i=t}^{t+k-1}\pi_{i+1}\mathbb{1}_{x_i=x} \\
& \cdot \left(\mathbb{1}_{x_k\neq x}\left(w+\sum_{x'}\pi(x'\mid s^*)Q(s^*,x')\right) + \mathbb{1}_{x_k=x}\right. \\
& \left. \cdot \left(y_k + \sum_{x'\neq x}\pi(x'\mid s_{k+1})Q(s_{k+1},x')\right)\right)
\end{aligned} \tag{18}$$

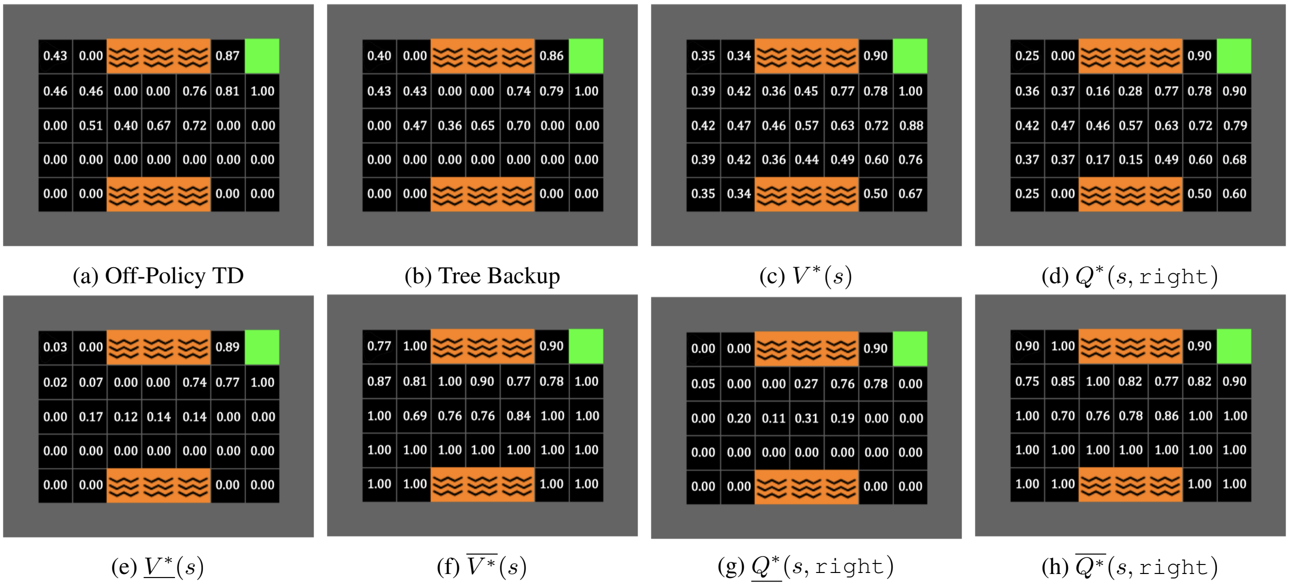
The above tree backup estimator also has a simple incremental implementation using eligibility traces. An online version of this implementation is shown in Fig. 5.

Theorem 4. *For any behavior policy, for any choice of* $\lambda \in [0,1]$ *that does not depend on the actions chosen at each state, let parameters* w *and* s^* *be defined as follows: (1) Lower Bound* $\underline{Q_\pi}$*:* $w = a$ *and* $s^* = \arg\min_s \sum_{x'} \pi(x' \mid s)Q_t(s,x')$*; (2) Upper Bound* $\overline{Q_\pi}$*:* $w = b$ *and* $s^* = \arg\max_s \sum_{x'} \pi(x' \mid s)Q_t(s,x')$*. Then, Alg. 2 with offline updating converges with probability* 1 *to lower bound* $\underline{Q_\pi}$ *and upper bound* $\overline{Q_\pi}$*, respectively, under the usual step-size conditions on* α*.*

The proof of the above theorem relies on a contraction property on the estimates $\widehat{Q}_\pi$ and follows from the general convergence theorem in [Jaakkola et al., 1994].

4 EXPERIMENTS

We demonstrate our algorithms in different variations of the Windy Gridworlds adapted from adapted from Deepmind's AI safety Gridworlds [Leike et al., 2017]. We found that simulation results support our findings, and the proposed causal eligibility trace algorithms consistently obtain informative bounds over target value functions. All experiments



(a) Off-Policy TD

0.43	0.00				0.87	
0.46	0.46	0.00	0.00	0.76	0.81	1.00
0.00	0.51	0.40	0.67	0.72	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00				0.00	0.00

(b) Tree Backup

0.40	0.00				0.86	
0.43	0.43	0.00	0.00	0.74	0.79	1.00
0.00	0.47	0.36	0.65	0.70	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00				0.00	0.00

(c) $V^*(s)$

0.35	0.34				0.90	
0.39	0.42	0.36	0.45	0.77	0.78	1.00
0.42	0.47	0.46	0.57	0.63	0.72	0.88
0.39	0.42	0.36	0.44	0.49	0.60	0.76
0.35	0.34				0.50	0.67

(d) $Q^*(s, \texttt{right})$

0.25	0.00				0.90	
0.36	0.37	0.16	0.28	0.77	0.78	0.90
0.42	0.47	0.46	0.57	0.63	0.72	0.79
0.37	0.37	0.17	0.15	0.49	0.60	0.68
0.25	0.00				0.50	0.60

(e) $\underline{V^*}(s)$

0.03	0.00				0.89	
0.02	0.07	0.00	0.00	0.74	0.77	1.00
0.00	0.17	0.12	0.14	0.14	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00				0.00	0.00

(f) $\overline{V^*}(s)$

0.77	1.00				0.90	
0.87	0.81	1.00	0.90	0.77	0.78	1.00
1.00	0.69	0.76	0.76	0.84	1.00	1.00
1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.00	1.00				1.00	1.00

(g) $\underline{Q^*}(s, \texttt{right})$

0.00	0.00				0.90	
0.05	0.00	0.00	0.27	0.76	0.78	0.00
0.00	0.20	0.11	0.31	0.19	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00				0.00	0.00

(h) $\overline{Q^*}(s, \texttt{right})$

0.90	1.00				0.90	
0.75	0.85	1.00	0.82	0.77	0.82	0.90
1.00	0.70	0.76	0.78	0.86	1.00	1.00
1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.00	1.00				1.00	1.00

Figure 7: Estimations of value functions obtained by (a, b) standard off-policy methods, (c, d) value interaction in the ground-truth model, and (e, h) causally enhanced off-policy algorithms using eligibility traces (`C-TD(`λ`)` and `C-TB(`λ`)`). The offline data are generated by a confounded behavior policy determining the agent's actions based on the wind direction.

use 5×10^4 offline observational samples, meaning that error bars are not significant, hence, not explicitly shown; the decay factor $\lambda = 0.5$ and discount factor $\gamma = 0.9$. For details on the experimental setup, we refer readers to the complete technical report [Zhang and Bareinboim, 2024].

Experiment 1. Consider again the learning setting in Example 1 where the demonstrator, following the behavior policy, decides whether to stay put and where to move based on the agent's state and the wind direction. Consequently, the offline data is contaminated with the unobserved confounding bias. We apply `C-TD(`λ`)` to derive bounds over the optimal value function $V^*(s)$ and provide them in Figs. 3e and 3f. The analysis reveals the derived bounds are consistent, containing the target value function in Fig. 3c.

Additionally, we compute the optimal state-action value function $Q^*(s, x)$ for action $x = \texttt{right}$ and provide it in Fig. 3d. We then estimate its bounds using `C-TB(`λ`)` from offline data; the bounding results are shown in Figs. 3d and 3e. By inspection, one can see our proposed algorithms are robust against the causal inconsistency in the offline data and consistently recover the informative bounds containing the actual value functions in the ground-truth model.

Experiment 2. We now consider an alternative Windy Gridworld described in Fig. 6a where the lava is placed at both top and bottom. Without sensing the wind, a preferable policy for the agent is to move along the center of the map where the wind strength is weak (highlighted in red in Fig. 6b). At the same time, the demonstrator takes the shortest path (orange in Fig. 6b) along the lava since it can sense the wind and take safe actions. Similar to the previous setting, the presence of wind direction becomes an unobserved confounder in the offline data, making the shorter route appear safer than it actually is.

We apply standard off-policy algorithms to evaluate the effect of the target policy π^* and provide their evaluations in Figs. 7a to 7b. We also compute bounds over the target value functions using our proposed algorithms, `C-TD(`λ`)` and `C-TB(`λ`)`, and provide their evaluations in Figs. 7e to 7f and Figs. 7g to 7h respectively. Comparing the bounds with the ground-truth value functions in Figs. 7c and 7d, we found that `C-TD(`λ`)` and `C-TB(`λ`)` can consistently obtain informative bounds. As expected, standard off-policy methods are not robust against causal inconsistency and deviate significantly from the target value functions.

5 CONCLUSION

This paper investigates off-policy evaluation in Markov Decision Processes from offline data collected by a different behavior policy, where unobserved confounding bias and no-overlap cannot be ruled out *a priori*. This leads to violations of causal consistency (Def. 2), which could pose significant challenges to standard off-policy algorithms. We first extend the celebrated Bellman's equation to derive informative bounds over values functions from the observational data, which are robust against bias due to the presence of unobserved confounding and no-overlap. Based on these extended equations, we propose two novel model-free off-policy algorithms using eligibility traces – one based on the standard temporal difference (`C-TD(`λ`)`), and the other based on the tree-backup (`C-TB(`λ`)`). These algorithms permit us to bound value functions from finite observations. Our simulation results show that standard off-policy RL algorithms cannot recover the actual value functions of a target policy when UCs and no-overlap generally exist. On the other hand, our proposed algorithms permit us to derive robust evaluations of the target value functions from imperfect offline observations.

ACKNOWLEDGEMENTS

This research was supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and the Alfred P. Sloan Foundation.

References

Patrick Bajari, C Lanier Benkard, and Jonathan Levin. Estimating dynamic models of imperfect competition. *Econometrica*, 75(5):1331–1370, 2007.

A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1172–1176, September 1997.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl*, pages 507–556. 2022.

Richard Bellman. Dynamic programming. *Science*, 153 (3731):34–37, 1966.

Steven T Berry and Giovanni Compiani. An instrumental variable approach to dynamic models. *The Review of Economic Studies*, 90(4):1724–1758, 2023.

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

Federico A Bugni. Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica*, 78(2):735–753, 2010.

Michael J Dickstein and Eduardo Morales. What do exporters know? *The Quarterly Journal of Economics*, 133 (4):1753–1801, 2018.

Hongyi Guo, Qi Cai, Yufeng Zhang, Zhuoran Yang, and Zhaoran Wang. Provably efficient offline reinforcement learning for partially observable markov decision processes. In *International Conference on Machine Learning*, pages 8016–8038. PMLR, 2022.

Guido W Imbens and Donald B Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The annals of statistics*, pages 305–327, 1997.

Tommi Jaakkola, Satinder Singh, and Michael Jordan. Reinforcement learning algorithm for partially observable markov decision problems. *Advances in neural information processing systems*, 7, 1994.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 652–661, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v48/jiang16.html`.

Shalmali Joshi, Junzhe Zhang, and Elias Bareinboim. Towards safe policy learning under partial identifiability: A causal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13004–13012, 2024.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9289–9299, 2018.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22293–22304. Curran Associates, Inc., 2020.

Chinmaya Kausik, Yangyi Lu, Kevin Tan, Maggie Makar, Yixin Wang, and Ambuj Tewari. Offline policy evaluation and optimization under confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 1459–1467. PMLR, 2024.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: partial identification through smoothness. *arXiv preprint arXiv:2305.11812*, 2023.

Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential causal imitation learning with unobserved confounders. *Advances in Neural Information Processing Systems*, 2021.

Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

C.F. Manski. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80: 319–323, 1990.

Hyungsik Roger Moon and Frank Schorfheide. Bayesian and frequentist inference in partially identified models. *Econometrica*, 80(2):755–782, 2012.

Eduardo Morales, Gloria Sheu, and Andrés Zahler. Extended gravity. *The Review of economic studies*, 86(6): 2668–2712, 2019.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1054–1062, 2016.

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.

Andriy Norets and Xun Tang. Semiparametric inference in dynamic binary choice models. *Review of Economic Studies*, 81(3):1229–1262, 2014.

J. Pearl and J.M. Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, pages 444–453. Morgan Kaufmann, San Francisco, 1995.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H Maathuis. A complete generalized adjustment criterion. *arXiv preprint arXiv:1507.01524*, 2015.

Dale J Poirier. Revising beliefs in nonidentified models. *Econometric theory*, 14(4):483–509, 1998.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759–766, 2000.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

Herbert Robbins. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1985.

Joseph P Romano and Azeem M Shaikh. Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, 138(9):2786–2807, 2008.

Kangrui Ruan, Junzhe Zhang, Xuan Di, and Elias Bareinboim. Causal imitation for markov decision processes: A partial identification approach. *Advances in neural information processing systems*, 2024.

I. Shpitser, T.J. VanderWeele, and J.M. Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 527–536. AUAI, Corvallis, OR, 2010.

Jörg Stoye. More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315, 2009.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

D Todem, J Fine, and L Peng. A global sensitivity test for evaluating statistical hypotheses with nonidentifiable models. *Biometrics*, 66(2):558–566, 2010.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge England, 1989.

J. Zhang and E. Bareinboim. Eligibility traces for confounding robust off-policy evaluation. Technical Report R-105, Causal Artificial Intelligence Lab, Columbia University, May 2024.

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, pages 13401–13411, 2019.

Junzhe Zhang and Elias Bareinboim. Bounding causal effects on continuous outcomes. In *Proceedings of the 35nd AAAI Conference on Artificial Intelligence*, 2021.

Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. *Advances in Neural Information Processing Systems*, 33, 2020.

Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In *International Conference on Machine Learning*, pages 26548–26558. PMLR, 2022.