

A RELATED WORK

Our work builds upon the literature on the partial identification of causal effects, sensitivity analysis, and robust reinforcement learning from offline data.

Partial Identification and Sensitivity Analysis Seminal work of [Manski \(1990\)](#) developed the first bounds on causal effects in non-identifiable settings using observational data in the single-stage treatment model with contextual information (i.e., a contextual bandit model). These bounds were then expanded to the instrumental variable setting ([Balke & Pearl, 1997](#); [Imbens & Angrist, 1994](#)) partially identify counterfactual probabilities of causation ([Tian & Pearl, 2000](#)). More recently, ([Zhang & Bareinboim, 2021](#)) improved the bounds for applicability to continuous outcomes. ([Zhang et al., 2022](#)) established a general framework for estimating bounds on interventional and counterfactual effects. While [Zhang et al. \(2022\)](#) develop informative bounds using both observational and experimental data, they focus on general counterfactual queries by discretizing the exogenous latent space, formulating bounds as polynomial programs over this discretization and a Bayesian framework to approximately estimate bounds using MCMC.

Sensitivity analysis attempts to provide intervals on causal effects by assuming the level of confounding, for example, via models such as Marginal Sensitivity analysis, which considers deviations in the propensity score in relation to the estimated propensity ([Rosenbaum, 2005](#); [Richardson et al., 2014](#); [Todem et al., 2010](#); [Vansteelandt et al., 2006](#); [Kallus & Zhou, 2018](#); [Kallus et al., 2019](#); [Namkoong et al., 2020](#); [Jesson et al., 2022](#); [Bruns-Smith & Zhou, 2023](#); [Kausik et al., 2024](#)). Other approaches explore additional parametric assumptions about the structural functions, including linearity ([Cinelli et al., 2019](#)) and Lipschitz continuity ([Khan et al., 2023](#)). Our work does not rely on additional functional constraints on the underlying system dynamics. Instead, we focus on the settings of standard discrete Markov Decision Processes (MDPs) with an infinite horizon. We develop robust off-policy evaluation algorithms to estimate closed-form bounds over the discounted cumulative rewards of candidate policies from offline observational data contaminated with unobserved confounding bias.

Robust Reinforcement Learning Unlike planning in a standard MDP, robust reinforcement learning does not assume the parametrization of the transition probability function in the underlying model to be precisely determined. Instead, it is contained in a set of model parameters which is called the uncertainty set ([Iyengar, 2005](#); [Nilim & El Ghaoui, 2005](#); [Xu & Mannor, 2010](#); [Wiesemann et al., 2013](#); [Yu & Xu, 2015](#); [Mannor et al., 2016](#); [Petrik & Russel, 2019](#)). The goal of the agent is to learn a robust policy that performs the best under the worst possible case in the uncertainty set. Similar problems have been studied under the rubrics of safe policy learning ([Thomas et al., 2015](#); [Ghavamzadeh et al., 2016](#)) or pessimistic reinforcement learning ([Shi et al., 2022](#)).¹

Robust RL algorithms with provable guarantees have been proposed in tabular settings or under the assumptions of linear functions ([Lim et al., 2013](#); [Tamar et al., 2014](#); [Roy et al., 2017](#); [Badrinath & Kalathil, 2021](#); [Wang & Zou, 2021](#)). Combined with the computational framework of deep learning, robust RL algorithms have been extended to complex, high-dimensional domains ([Pinto et al., 2017](#); [Zhang et al., 2020](#)). More recently, ([Panaganti et al., 2022](#)) proposed Robust Fitted Q-Iteration (RFQI) to learn the best possible robust policy from offline data with theoretical guarantees on the performance of the learned policy. Our work differs from robust RL methods since it does not require a pre-specified uncertainty set of model parameters. Instead, we construct the ignorance region over the underlying system dynamics from the confounded observational data using partial causal identification. Based on the learned uncertainty set, we then derived closed-form bounds over the value functions of the target policy. *To the best of our knowledge, this is the first work that develops off-policy algorithms using eligibility traces to obtain evaluations of candidate policies from biased offline data, possibly contaminated with unmeasured confounding or no-overlap, with provable guarantees on the convergence of learned evaluations.*

¹Indeed, the idea of planning over a convex set of model parameters have been explored in online reinforcement learning. ([Strehl & Littman, 2008](#)) utilized an extended dynamic programming to learn an optimistic policy over a confidence set of models to balance the trade-off between exploration and exploitation.

B PROOFS

This section provides proof of the main theoretical results provided in the paper.

Theorem 1 (Causal Bellman Equation). *For an MDP environment M with reward $Y_t \in [a, b] \subseteq \mathbb{R}$, for any policy $\pi(x | s)$, its state value function $V_\pi(s) \in [V_\pi(s), \bar{V}_\pi(s)]$ for every state $s \in \mathcal{S}$, where bounds V_π, \bar{V}_π are solutions given by the following dynamic programs,*

$$\langle V_\pi(s), \bar{V}_\pi(s) \rangle = \sum_x P(x | s) \left(\pi(x | s) \left(\tilde{\mathcal{R}}(s, x) + \gamma \sum_{s', x'} \tilde{\mathcal{T}}(s, x, s') \langle V_\pi(s'), \bar{V}_\pi(s') \rangle \right) \right) \quad (10)$$

$$+ \pi(\neg x | s) \left(\langle a, b \rangle + \gamma \left\langle \min_{s'} V_\pi(s'), \max_{s'} \bar{V}_\pi(s') \right\rangle \right) \quad (11)$$

Proof. Following the Bellman equation (Bellman, 1966), the state value function at state $s \in \mathcal{S}$ is given by

$$V_\pi(s) = \sum_x \pi(x | s) \left(\mathcal{R}(s, x) + \gamma \sum_{s'} \mathcal{T}(s, x, s') V_\pi(s') \right) \quad (20)$$

Among the above quantities, the reward function \mathcal{R} is bounded from the observational distribution (Manski, 1990) as follows,

$$\tilde{\mathcal{R}}(s, x) P(x | s) + aP(\neg x | s) \leq \mathcal{R}(s, x) \leq \tilde{\mathcal{R}}(s, x) P(x | s) + bP(\neg x | s) \quad (21)$$

where $\tilde{\mathcal{R}}$ is the nominal reward function computed from the observational distribution and is defined in Eq. (9). Replacing the reward function \mathcal{R} in Eq. (20) with the above lower bound gives

$$V_\pi(s) \geq \sum_x \pi(x | s) \left(\tilde{\mathcal{R}}(s, x) P(x | s) + aP(\neg x | s) + \gamma \sum_{s'} \mathcal{T}(s, x, s') V_\pi(s') \right) + \sum_x b\pi(x | s)P(\neg x | s) \quad (22)$$

Similarly, the transition distribution \mathcal{T} can be bounded from the observational distribution (Manski, 1990),

$$\tilde{\mathcal{T}}(s, x, s') P(x | s) \leq \mathcal{T}(s, x, s') \leq \tilde{\mathcal{T}}(s, x, s') P(x | s) + P(\neg x | s) \quad (23)$$

and $\tilde{\mathcal{T}}$ is the nominal transition distribution computed from the observational distribution defined in Eq. (9). Minimizing the lower bound in Eq. (22) subject to the above observational constraints in Eq. (23) and $\sum_{s'} \mathcal{T}(s, x, s') = 1$ gives the following lower bound:

$$V_\pi(s) \geq \sum_x \pi(x | s) P(x | s) \left(\tilde{\mathcal{R}}(s, x) + aP(\neg x | s) + \gamma \sum_{s'} \tilde{\mathcal{T}}(s, x, s') V_\pi(s') \right) + \sum_x \pi(x | s) P(\neg x | s) \left(b + \min_{s'} V_\pi(s') \right) \quad (24)$$

The above lower bound is achieved by setting the worst-case transition probability $\mathcal{T}(s, x, s^*) = P(\neg x | s)$ for state $s^* = \arg \min_{s'} V_\pi(s')$ and $\mathcal{T}(s, x, s') = \tilde{\mathcal{T}}(s, x, s') P(x | s)$ for all the other state $s' \neq s^*$. Note that the second term of the above inequality could be further written as:

$$\sum_x \pi(x | s) P(\neg x | s) \left(a + \min_{s'} V_\pi(s') \right) \quad (25)$$

$$= \sum_x \pi(x | s) (1 - P(x | s)) \left(a + \min_{s'} V_\pi(s') \right) \quad (26)$$

$$= \sum_x \pi(x | s) \left(a + \min_{s'} V_\pi(s') \right) - \sum_x \pi(x | s) P(x | s) \left(a + \min_{s'} V_\pi(s') \right) \quad (27)$$

$$= \sum_x P(x | s) \left(a + \min_{s'} V_\pi(s') \right) - \sum_x \pi(x | s) P(x | s) \left(a + \min_{s'} V_\pi(s') \right) \quad (28)$$

The last step holds since for any constant real value C , $\sum_x \pi(x | s)C = \sum_x P(x | s)C$. The above equation can be further written as

$$\sum_x \pi(x | s)P(\neg x | s) \left(a + \min_{s'} V_\pi(s') \right) = \sum_x \pi(\neg x | s)P(x | s) \left(a + \min_{s'} V_\pi(s') \right) \quad (29)$$

Replacing the second term in Eq. (24) gives

$$\begin{aligned} V_\pi(s) \geq \sum_x \pi(x | s)P(x | s) \left(\tilde{\mathcal{R}}(s, x) + bP(\neg x | s) + \gamma \sum_{s'} \tilde{\mathcal{T}}(s, x, s') V_\pi(s') \right) \\ + \sum_x \pi(\neg x | s)P(x | s) \left(a + \min_{s'} V_\pi(s') \right) \end{aligned} \quad (30)$$

After a few simplifications, we obtain

$$\begin{aligned} V_\pi(s) \geq P(x | s) \left(\pi(x | s) \left(\tilde{\mathcal{R}}(s, x) + \gamma \sum_{s', x'} \tilde{\mathcal{T}}(s, x, s') V_\pi(s') \right) \right. \\ \left. + \pi(\neg x | s) \left(a + \gamma \min_{s'} V_\pi(s') \right) \right) \end{aligned} \quad (31)$$

Finally, minimizing the value function V_π subject to the above inequality gives the lower bound \underline{V}_π . The upper bound \overline{V}_π over the state value function could be similarly derived. \square

Theorem 2 (Causal Bellman Equation). *For an MDP environment M with reward signals $Y_t \in [a, b] \subseteq \mathbb{R}$, for any policy $\pi(x | s)$, its state-action value function $Q_\pi \in [Q_\pi(s, x), \overline{Q}_\pi(s, x)]$ for any state-action pair $(s, x) \in \mathcal{S} \times \mathcal{X}$, where bounds $\underline{Q}_\pi, \overline{Q}_\pi$ are given by as follows,*

$$\langle Q_\pi(s, x), \overline{Q}_\pi(s, x) \rangle = P(x | s) \left(\tilde{\mathcal{R}}(s, x) + \gamma \sum_{s', x'} \tilde{\mathcal{T}}(s, x, s') \langle V_\pi(s'), \overline{V}_\pi(s') \rangle \right) \quad (12)$$

$$+ P(\neg x | s) \left(\langle a, b \rangle + \gamma \langle \min_{s'} \underline{V}_\pi(s'), \max_{s'} \overline{V}_\pi(s') \rangle \right) \quad (13)$$

Proof. Applying Bellman equation (Bellman, 1966) allows us to iteratively write the state-action value function for any state-action pair $(s, x) \in \mathcal{S} \times \mathcal{X}$ as

$$Q_\pi(s, x) = \mathcal{R}(s, x) + \gamma \sum_{s'} \mathcal{T}(s, x, s') V_\pi(s') \quad (32)$$

where the reward function \mathcal{R} is bounded from the observational distribution (Manski, 1990) following Eq. (21). Replacing the reward function \mathcal{R} in the above equation with the corresponding lower bound gives

$$Q_\pi(s, x) \geq P(x | s) \left(\tilde{\mathcal{R}}(s, x) + \gamma \sum_{s'} \mathcal{T}(s, x, s') V_\pi(s') \right) + aP(\neg x | s) \quad (33)$$

Similarly, the transition distribution \mathcal{T} can be bounded from the observational distribution (Manski, 1990) following Eq. (23). Minimizing the lower bound in Eq. (33) subject to the above observational constraints in Eq. (23) and $\sum_{s'} \mathcal{T}(s, x, s') = 1$ gives the following solution:

$$Q_\pi(s, x) \geq P(x | s) \left(\tilde{\mathcal{R}}(s, x) + \gamma \sum_{s'} \tilde{\mathcal{T}}(s, x, s') V_\pi(s') \right) + P(\neg x | s) \left(a + \min_{s'} V_\pi(s') \right) \quad (34)$$

This lower bound is achieved by setting the worst-case transition probability $\mathcal{T}(s, x, s^*) = P(\neg x | s)$ for state $s^* = \arg \min_{s'} V_\pi(s')$ and $\mathcal{T}(s, x, s') = \tilde{\mathcal{T}}(s, x, s')P(x | s)$ for all the other state $s' \neq s^*$. Finally, notice that $V_\pi(s)$ is a function of $Q_\pi(s, x)$ and is given by $V_\pi(s) = \sum_x \pi(x | s)Q_\pi(s, x)$. Minimizing the state-action value function Q_π subject to the above inequality leads to the lower bound \underline{Q}_π . The upper bound \overline{Q}_π could be similarly derived. \square

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}_π : $w = a$ and $s^* = \arg \min_s V_t(s)$; (2) Upper Bound \overline{V}_π : $w = b$ and $s^* = \arg \max_s V_t(s)$. Then, Alg. [1](#) with offline updating converges with probability 1 to lower bound \underline{V}_π and upper bound \overline{V}_π , respectively, under the usual step-size conditions on α .

Proof. We will focus on the convergence of lower bound $\underline{V}_\pi(s)$; the proof for the upper bound $\overline{V}_\pi(s)$ follows analogously. The proof is structured in two stages. First, we consider the truncated lower bound estimates corresponding to Eq. [\(14\)](#), which sums the adjusted rewards obtained from the environment for only n steps, then uses the current estimate of the value function lower bound to approximate the remaining value:

$$\underline{R}_t^{(n)} = \sum_{k=0}^{n-1} \gamma^k \left(\pi_{t+k} y_{t+k} + \neg \pi_{t+k} (b + \gamma \min_{s'} V(s')) \right) \prod_{i=t}^{t+k-1} \pi_i + \gamma^n V(s_{t+n}) \prod_{i=t}^{t+k-1} \pi_i \quad (35)$$

We need to show that $\underline{R}_t^{(n)} - \underline{V}_\pi$ is a contraction mapping in the max norm. If this is true for any n , then by applying the general convergence theorem, the n -step return converges to \underline{V}_π . Then any convex combination will also converge to \underline{V}_π . For example, any combination using a λ parameter in the style of eligibility traces will converge to \underline{V}_π .

The expected value of the adjusted return with regard to the observational distribution for state s can be expressed as follows [2](#):

$$\mathbb{E} \left[\underline{R}_t^{(n)} \mid S_t = s \right] \quad (36)$$

$$= \sum_{k=1}^n \sum_{\bar{s}_{1:k}, \bar{x}_{1:k}, \bar{y}_{1:k}} P(\bar{s}_{1:k}, \bar{x}_{1:k}, \bar{y}_{1:k}) \gamma^{k-1} \left(\pi_k y_k + \neg \pi_k (b + \min_{s'} V(s')) \right) \prod_{i=1}^{k-1} \pi_i \quad (37)$$

$$+ \sum_{\bar{s}_{1:n}, \bar{x}_{1:n}} P(\bar{s}_{1:n}, \bar{x}_{1:n}) \gamma^n V(s_n) \prod_{i=1}^{n-1} \pi_i \quad (38)$$

$$= \sum_{k=1}^n \gamma^{k-1} \sum_{\bar{s}_{1:k}, \bar{x}_{1:k}} \prod_{i=1}^{k-1} \tilde{T}(s_i, x_i, s_{i+1}) P(x_i \mid s_i) \pi(x_i \mid s_i) \quad (39)$$

$$\cdot \left(\pi(x_k \mid s_k) \tilde{\mathcal{R}}(s_k, x_k) + \neg \pi(x_k \mid s_k) (b + \gamma \min_{s'} V(s')) \right) \quad (40)$$

$$+ \gamma^n \sum_{\bar{s}_{1:n}, \bar{x}_{1:n}} \prod_{i=1}^{n-1} \tilde{T}(s_i, x_i, s_{i+1}) P(x_i \mid s_i) \pi(x_i \mid s_i) V(s_n) \quad (41)$$

By applying the extended Bellman equation for the lower bound \underline{V}_π iteratively n times, we obtain:

$$\underline{V}_\pi(s) = \sum_{k=1}^n \sum_{\bar{s}_{1:k}, \bar{x}_{1:k}} \gamma^{k-1} \prod_{i=1}^{k-1} \tilde{T}(s_i, x_i, s_{i+1}) P(x_i \mid s_i) \pi(x_i \mid s_i) \quad (42)$$

$$\cdot \left(\pi(x_k \mid s_k) \tilde{\mathcal{R}}(s_k, x_k) + \neg \pi(x_k \mid s_k) (b + \gamma \min_{s'} \underline{V}_\pi(s')) \right) \quad (43)$$

$$+ \gamma^n \sum_{\bar{s}_{1:n}, \bar{x}_{1:n}} \prod_{i=1}^{n-1} \tilde{T}(s_i, x_i, s_{i+1}) P(x_i \mid s_i) \pi(x_i \mid s_i) \underline{V}_\pi(s_n) \quad (44)$$

Therefore,

$$\max_s \left| \mathbb{E} \left[\underline{R}_t^{(n)} \mid S_t = s \right] - \underline{V}_\pi(s) \right| \leq \gamma \max_s |V(s) - \underline{V}_\pi(s)| \quad (45)$$

This means that any n -step return is a contraction in the max norm, and therefore, by applying [\(Jaakkola et al., 1994, Theorem 1\)](#), it converges to $\underline{V}_\pi(s)$.

²We abuse notation a bit and ignore the expected value operator $\mathbb{E}[\cdot]$ outside.

In the second stage, we show that by applying the updates of Alg. 1 for n successive steps, we perform the same update by using the n -step adjusted return $\underline{R}_t^{(n)}$. The eligibility trace for state s can be written as, for $t_n \in \mathbf{t}(s)$,

$$e_t(s) = \gamma^{t-t_n} \prod_{i=t_n+1}^t \pi_i. \quad (46)$$

We have

$$\sum_{k=1}^n e_{t+k-1}(s) \delta_{t+k-1}(s) \quad (47)$$

$$= \sum_{k=1}^n \gamma^{k-1} \prod_{i=t+1}^{t+k-1} \pi_i \left(\pi_{t+k}(y_{t+k} + \gamma V(s_{t+k})) + \pi_{t+k} \left(b + \gamma \min_{s'} V(s') \right) \right) \quad (48)$$

$$- V(s_{t+k-1}) \quad (49)$$

$$= \sum_{k=0}^{n-1} \gamma^k \left(\pi_{t+k} y_{t+k} + \neg \pi_{t+k} \left(b + \gamma \min_{s'} V(s') \right) \right) \prod_{i=t}^{t+k-1} \pi_i + \gamma^n V(s_{t+n}) \prod_{i=t}^{t+k-1} \pi_i \quad (50)$$

$$- V(s_t) \quad (51)$$

$$= \underline{R}_t^{(n)} - V(s_t) \quad (52)$$

Since C-TD (λ) is equivalent to applying a convex mixture of n -step updates, and each update converges to correct lower bounds \underline{V}_π for the state value functions, Alg. 1 converges to correct lower bounds as well. \square

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}_π : $w = a$ and $s^* = \arg \min_s \sum_{x'} \pi(x' | s) Q_t(s, x')$; (2) Upper Bound \overline{Q}_π : $w = b$ and $s^* = \arg \max_s \sum_{x'} \pi(x' | s) Q_t(s, x')$. Then, Alg. 2 with offline updating converges with probability 1 to lower bound \underline{Q}_π and upper bound \overline{Q}_π , respectively, under the usual step-size conditions on α .

Proof. We will focus on the convergence of lower bound $\underline{Q}_\pi(s, x)$; the proof for the upper bound $\overline{Q}_\pi(s, x)$ follows analogously. This proof is structured in two stages. Let Q_n denote the n -step tree backup estimator defined in Eq. (19). First we show that $\mathbb{E}[Q_n(s, x)] - \underline{Q}_\pi(s, x)$ is a contraction using a proof by induction.

Let Q be the current estimate of the lower bound for the value function. For $n = 1$,

$$\max_{s,x} |\mathbb{E}[Q_1(s, x)] - \underline{Q}_\pi(s, x)| \quad (53)$$

$$= \max_{s,x} \left| P(x | s) \left(\tilde{\mathcal{R}}(s, x) + \gamma \sum_{s', x'} \tilde{\mathcal{T}}(s, x, s') \sum_{x'} \pi(x' | s') Q(s', x') \right) \right. \quad (54)$$

$$\left. + P(\neg x | s) \left(b + \gamma \min_{s'} \sum_{x'} \pi(x' | s') Q(s', x') \right) \right) \quad (55)$$

$$- P(x | s) \left(\tilde{\mathcal{R}}(s, x) + \gamma \sum_{s', x'} \tilde{\mathcal{T}}(s, x, s') \sum_{x'} \pi(x' | s') \underline{Q}_\pi(s', x') \right) \quad (56)$$

$$\left. - P(\neg x | s) \left(b + \gamma \min_{s'} \sum_{x'} \pi(x' | s') \underline{Q}_\pi(s', x') \right) \right| \quad (57)$$

$$\leq \gamma \max_{s,x} |Q(s, x) - \underline{Q}_\pi(s, x)| \quad (58)$$

For the induction step, we assume that

$$\max_{s,x} |\mathbb{E}[Q_n(s, x)] - \underline{Q}_\pi(s, x)| \leq \gamma \max_{s,x} |Q(s, x) - \underline{Q}_\pi(s, x)| \quad (59)$$

Next we want to show that the same holds for $Q_{n+1}(s, x)$. We can rewrite $Q_{n+1}(s, x)$ as follows,

$$Q_{n+1}(s, x) = \mathbb{1}_{x_t=x} \left(y_t + \sum_{x'} \left(\mathbb{1}_{x' \neq x} \pi(x' | s_{t+1}) Q(s_{t+1}, x') + \mathbb{1}_{x'=x} Q_n(s_{t+1}, x) \right) \right) \quad (60)$$

$$+ \mathbb{1}_{x_t \neq x} \left(w + \sum_{x'} \pi(x' | s^*) Q(s^*, x') \right) \quad (61)$$

We must have

$$\max_{s,x} |\mathbb{E}[Q_{n+1}(s, x)] - \underline{Q}_\pi(s, x)| \quad (62)$$

$$= \max_{s,x} \left| P(x | s) \left(\tilde{\mathcal{R}}(s, x) + \gamma \sum_{s', x'} \tilde{\mathcal{T}}(s, x, s') \sum_{x'} \pi(x' | s') \right. \right. \quad (63)$$

$$\left. \mathbb{1}_{x' \neq x} Q(s', x') + \mathbb{1}_{x'=x} \mathbb{E}[Q_n(s', x)] \right) \quad (64)$$

$$+ P(\neg x | s) \left(b + \gamma \min_{s'} \sum_{x'} \pi(x' | s') Q(s', x') \right) \quad (65)$$

$$- P(x | s) \left(\tilde{\mathcal{R}}(s, x) + \gamma \sum_{s', x'} \tilde{\mathcal{T}}(s, x, s') \sum_{x'} \pi(x' | s') \underline{Q}_\pi(s', x') \right) \quad (66)$$

$$- P(\neg x | s) \left(b + \gamma \min_{s'} \sum_{x'} \pi(x' | s') \underline{Q}_\pi(s', x') \right) \quad (67)$$

$$\leq \gamma \max_{s,x} \left| P(x | s) \gamma \sum_{s', x'} \tilde{\mathcal{T}}(s, x, s') \sum_{x'} \pi(x' | s') \mathbb{1}_{x' \neq x} (Q(s', x') - \underline{Q}_\pi(s', x')) \right. \quad (68)$$

$$\left. + \mathbb{1}_{x'=x} \mathbb{E}[(Q_n(s', x) - \underline{Q}_\pi(s', x'))] \right) \quad (69)$$

$$+ P(\neg x | s) \min_{s'} \sum_{x'} \pi(x' | s') (Q(s', x') - \underline{Q}_\pi(s', x')) \quad (70)$$

$$\leq \gamma \max_{s,x} |Q(s, x) - \underline{Q}_\pi(s, x)| \quad (71)$$

By applying (Jaakkola et al., 1994, Theorem 1), we can conclude that any n -step adjusted return converges to the correct lower bound for the state-action value function. Since all the n -step returns converge to \underline{Q}_π , any convex linear combination of n -step returns also converges to \underline{Q}_π .

For the second part of the proof, we show that C-TB (λ) with $\lambda = 1$ for n steps is equivalent to using Q_n . The eligibility trace for a state-action pair (s, x) can be rewritten as:

$$e_t(s, x) = \gamma^k \prod_{i=t+1}^{t+k-1} \pi_{i+1} \mathbb{1}_{x_i=x}. \quad (72)$$

By adding and subtracting the weighted action value $\pi_{t+k} \mathbb{1}_{x_{t+k}=x}$ for the action taken on each step from the return, and regrouping, we have

$$Q(s_t, x) + \sum_{k=1}^n \gamma^{k-1} \prod_{i=t+1}^{t+k-1} \pi_{i+1} \mathbb{1}_{x_i=x} \left(\mathbb{1}_{x_{t+k}=x} \left(y_{t+k} + \sum_{x' \neq x} \pi(x' | s_{t+k+1}) Q(s_{t+k+1}, x') \right) \right) \quad (73)$$

$$+ \mathbb{1}_{x_{t+k} \neq x} \left(w + \min_{s'} \sum_{x'} \pi(x' | s') Q(s', x') \right) - Q(s_{t+k}, x) \quad (74)$$

$$= Q(s_t, x) + \sum_{k=1}^n e_{t+k}(s_t, x) \delta_{t+k}(x) \quad (75)$$

This concludes the proof. \square

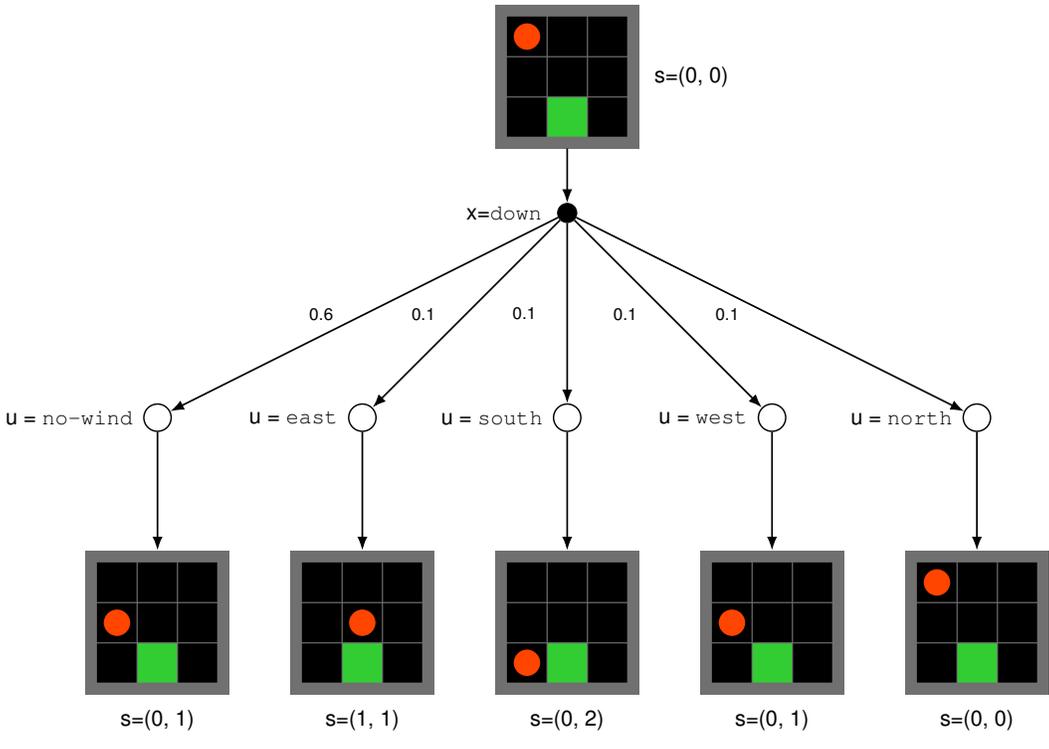


Figure 6: Trajectories sampled from the interventional transition distribution \mathcal{T} .

C EXPERIMENTAL SETUPS

In this section, we provide details on the experimental setups and additional discussion on the simulation environment. All experiments were performed on a 2021 MacBook Pro with 16GB memory, implemented in Python. The simulation environment is built upon the Gymnasium framework (Brockman et al., 2016). We plan to release the source code with the camera-ready version of the manuscript.

Windy Gridworld Our simulation builds on the Windy Gridworld environment described in Fig. 1B, where the red dot represents the agent and the green square represents the goal state. The agent’s location is represented using a vector (i, j) where $i \in \{0, 1, 2\}$ is the column index, and $j \in \{0, 1, 2\}$ is the row index. So the agent’s starting state is $(0, 0)$ and the goal state is $(1, 2)$. Fig. 7 shows the detailed state representation for each location in the gridworld.

The agent can take five actions $x \in \mathcal{X}$ - up, down, right, left, and stay-put, corresponding to vector $(0, -1), (0, 1), (1, 0), (-1, 0)$, and $(0, 0)$ respectively. Meanwhile, the agent’s movement is also affected by a wind; the wind direction $u \in \mathcal{U}$ include - north, south, east, west, and no-wind, corresponding to vector $(0, -1), (0, 1), (1, 0), (-1, 0)$, and $(0, 0)$ respectively. Table 1 summarizes the detailed parametrization for the agent’s action and the wind direction.

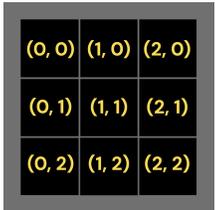


Figure 7: Agent’s state in Windy Gridworld environment.

Action x	up	down	right	left	stay-put
Wind u	north	south	east	west	no-wind
Vector v	$(0, -1)$	$(0, 1)$	$(1, 0)$	$(-1, 0)$	$(0, 0)$

Table 1: Vector representations for the agent’s action X and the wind direction U .

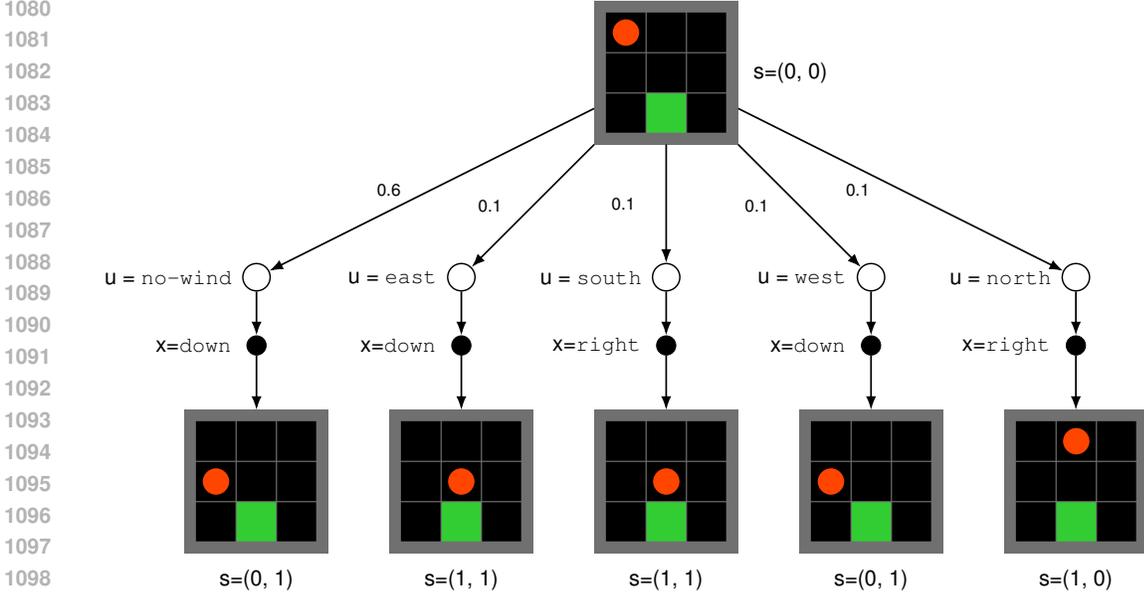


Figure 8: Trajectories sampled from the observational transition distribution $\tilde{\mathcal{T}}$ induced by a confounded behavior policy f_X .

every time step $t = 1, 2, \dots$, the wind U_t can blow in directions north, south, east, west with equal probabilities of 10%; otherwise, the weather is nice and there is no-wind. That is,

$$\forall i \in \{-1, 1\}, \quad P(U_t = (i, 0)) = P(U_t = (0, i)) = 0.1, \quad \text{and} \quad P(U_t = (0, 0)) = 0.6 \quad (76)$$

At every time step t , the agent receives a constant reward $Y_t \leftarrow -1$. The next state of the agent is shifted by both its action and the wind direction through the mechanism

$$S_{t+1} \leftarrow \max \{ \min \{ S_t + X_t + U_t, (2, 2) \}, (0, 0) \}. \quad (77)$$

In other words, the agent’s next state S_{t+1} is a vector sum of the agent’s current location S_t , its action X_t , and the wind direction U_t , truncated by the board’s boundary $i = 0, 2$ and $j = 0, 2$. For instance, we show in Fig. 6 the system dynamics for the agent’s interactions with the gridworld environment at from the location $s = (0, 0)$, taking the action down ($x = (0, 1)$). In this case, when the wind is blowing towards south ($u = (0, 1)$), the agent’s location will be shifted by both the action x and the windy direction u , and moves to the bottom left corner $s' = (0, 2)$ at the next time step. Since among all wind directions, $u = \text{east}$ is the only latent state moving the agent to the center $s' = (0, 2)$, we must have the following evaluation for the interventional distribution $P_{X_t}(S_{t+1} | S_t)$,

$$P_{X_t \leftarrow (0,1)}(S_{t+1} = (0, 2) | S_t = (0, 0)) = P(U_t = (1, 0)) \quad (78)$$

$$= 0.1 \quad (79)$$

That is, the agent’s transition distribution $\mathcal{T}(s, x, s') = 0.1$ when starting from $s = (0, 1)$, taking action $x = (0, 1)$, and moving to the next state $s' = (0, 2)$.

Confounded Behavior Policy Consider now an off-policy learning task in the windy gridworld, where the agent’s goal is to evaluate the expected return of a target policy π^* described in Fig. 2a. Following such a policy π^* , the agent will consistently move towards the goal state $s = (1, 2)$ from its current location, regardless of the wind direction.

The detailed parametrization of the agent’s system dynamics in the windy gridworld remains unknown. Instead, it has access to observed trajectories generated by a behavior policy $x \leftarrow f_X(s, u)$ which could sense the wind and select an action accordingly; Fig. 9 provides a detailed description for this behavior policy. For example, when the agent is located in the top-left corner ($s = (0, 0)$) and the wind is blowing south ($s = (0, 1)$), the behavior policy $x \leftarrow f_X(s, u)$ will decide to move right ($x = (1, 0)$) so that the agent could get to the center ($s' = (1, 1)$).

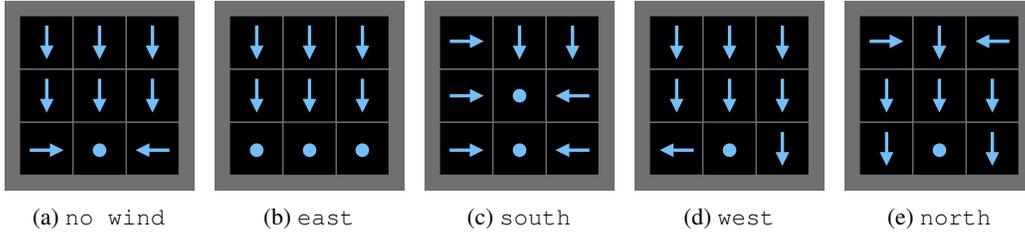


Figure 9: A confounded behavior policy f_X selecting values based on the agent’s location S and the latent wind direction U .

Consequently, the wind direction U_t becomes an unobserved confounder in the generative process for the offline observational data, affecting the allocated action X_t and the next state S_{t+1} simultaneously. The presence of unobserved confounders lead to violations of causal consistency (Def. 2). To witness, Fig. 8 shows observed trajectories in the offline data when the agent starts from state $s = (0, 0)$. When the weather is nice (no-wind) or the wind u is blowing towards east or west, the behavior policy selects action $x = \text{down}$, similar to the interventional trajectories of Fig. 6. On the other hand, when the wind is blowing towards north or south, the behavior policy selects action $x = \text{right}$, moving the agent towards the center of the board. Among all the possible next state in the observational data, we find that the agent will never reach the bottom left corner $s = (0, 2)$. This means that when evaluating the observational distribution $P(S_{t+1} | S_t, X_t)$, we must have

$$P(S_{t+1} = (0, 2) | S_t = (0, 0), X_t = (0, 1)) = 0 \quad (80)$$

In other words, the nominal transition distribution $\tilde{T}(s, x, s') = 0$ when one observes the agent starting from $s = (0, 1)$, taking action $x = (0, 1)$, and moving to the next state $s' = (0, 2)$. Comparing the evaluations in Eqs. (79) and (80), we find that $P_{x_t}(s_{t+1} | s_t) \neq P(s_{t+1} | s_t, x_t)$, that is, causal consistency (Def. 2) does not hold between the agent’s system dynamics in windy gridworld and the observational distribution generated by the confounded behavior policy in Fig. 9.