Leveraging diverse offline data in POMDPs with unobserved confounders

Oussama Azizi
Delft University of Technology
o.azizi@tudelft.nl

Philip Boeken
University of Amsterdam
p.a.boeken@uva.nl

Onno Zoeter
Booking.com
onno.zoeter@booking.com

Frans A. Oliehoek
Delft University of Technology
f.a.oliehoek@tudelft.nl

Matthijs T. J. Spaan
Delft University of Technology
m.t.j.spaan@tudelft.nl

Abstract

In many Reinforcement Learning (RL) applications, offline data is readily available before an algorithm is deployed. Often, however, data-collection policies have had access to information that is not recorded in the dataset, requiring the RL agent to take unobserved confounders into account. We focus on the setting where the confounders are i.i.d. and, without additional assumptions on the strength of the confounding, we derive tight bounds for the causal effects of the actions on the observations and reward. In particular, we show that these bounds are tight when we leverage multiple datasets collected from diverse behavioral policies. We incorporate these bounds into Posterior Sampling for Reinforcement Learning (PSRL) and demonstrate their efficacy experimentally.

1 Introduction

A barrier to the adoption of reinforcement learning (RL) in the real world is that many methods learn from scratch. To overcome this, a possible solution is methods that can use ('offline') data, which is collected by the currently deployed control strategies, to 'warm start' the RL algorithm. Unfortunately, in many real-world applications, such as medicine and recommendation systems, the data-collection policies (called *behavioral policies*)—certainly when it is executed by humans—have access to information that is not recorded in the dataset. Such *unobserved (or 'latent') confounders* introduce spurious correlations [Ortega et al., 2021], which can mislead a naive learner and can potentially lead to performance that is arbitrarily worse than the performance that the data-collection policies realized. Although a simple approach to avoid being misled is to simply ignore the past data, this means learning from scratch which is just not feasible in many realistic applications.

To overcome this problem, work on offline RL has considered latent confounding [Bruns-Smith, 2021, Namkoong et al., 2020, Bennett et al., 2021]. However, the offline RL setting does not consider that further learning (in the 'online' phase) is possible. This further learning is considered in the so-called hybrid (i.e., offline-online) setting [Gasse et al., 2023, Wang et al., 2021, Zhang and Bareinboim, 2019] that we focus on. The key difficulty here is to incorporate the prior data without being misled, and come up with a good exploration strategy that then quickly hones in on the optimal policy. Specifically, we focus on Posterior Sampling for Reinforcement Learning (PSRL) [Osband et al., 2013], which has many desirable properties including near-optimal and worst-case regret bounds [Osband and Van Roy, 2017], and how it can be warm-started with offline data in presence of latent confounding.

Using PSRL in settings with latent confounded offline data has been explored in the context of Multi-Armed Bandits [Zhang and Bareinboim, 2017], but extending this to the sequential setting

17th European Workshop on Reinforcement Learning (EWRL 2024).

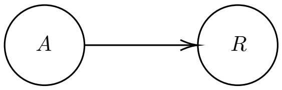


Figure 1: Stochastic MAB: Fully observable and A causes R

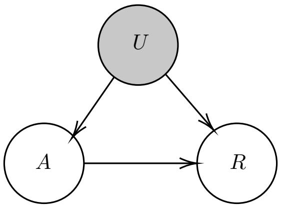


Figure 2: Partially observable MAB (Offline learning under latent confounding), where A and R are observed and not U

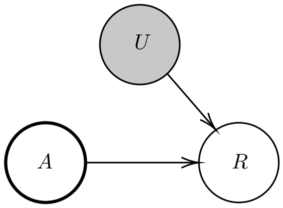


Figure 3: Partially observable MAB (Online learning), where A is controlled

is non-trivial and the focus of the current paper. We provide a definition of POMDPs with i.i.d. unobserved confounders to model the sequential problem without needing to resort to structural causal models, and use it to propose Causally-Truncated-PSRL (CT-PSRL), which can warm start from offline data. To do this, we derive causal bounds that are expressed in terms of observational quantities that can be estimated from the offline data, thus proposing a non-trivial extensions of the bounds employed by Zhang and Bareinboim [2017] in the bandit setting. Motivated by real-world applications, we show that the overlap of the causal bounds from different behavioral policies is tight when they uniformly cover the action space. Finally, we demonstrate experimentally that this uniform coverage improve the sample efficiency of CT-PSRL thus making a step towards lowering the barrier of deploying RL algorithms in real-world settings.

2 Background

2.1 Interventions and confounding

We start with a multi-armed bandit (MAB) problem depicted by the causal graph in Figure 1, where A is the action, R the reward that depends on the action A. Suppose an agent desires to evaluate its policy π. There are two settings to do so:

Interventions: The agent evaluates π by directly interacting with the environment and taking actions sampled from its own policy π and evaluate the probability of $R = r$, which can be denoted using Pearl's do notation [Pearl, 1995] as $P(R = r|\text{do}(\pi)) = \sum_a P(R = r|A = a)\pi(A = a)$. This distribution is called the interventional distribution and this corresponds to the On-policy setting.

Observations: Without having access to the environment, the agent can alternatively observe another policy π_β taking actions in the environment, and leverage these observations to estimate the interventional distribution. This corresponds to the Off-policy setting. In this case, the distribution of the observed rewards is $P(R = r|\pi_\beta) = \sum_a P(R = r|A = a)\pi_\beta(A = a)$. The agent can then adapt this distribution for example using Importance Sampling for this purpose.

Latent confounding: In the Off-policy setting, latent confounding corresponds to the case where the policy π_β leverages some unobserved information U, i.e. the latent confounder, as illustrated in Figure 2. In this case, the observational distribution of the reward is $P(R|\pi_\beta) = \sum_u \sum_a p(U = u)\pi_\beta(A = a|U)P(R|A, U)$. Since U is unobserved, the agent cannot correct for the bias introduced by the term $\pi_\beta(A|U)$ to estimate the interventional distribution $P(R|do(\pi)) = \sum_u \sum_a P(R|A, U)\pi(A = a)P(U)$ that is induced by the causal graph in Figure 3.[1]

2.2 Posterior Sampling for Reinforcement Learning

We consider a finite-horizon Markov Decision Process (MDP) [Puterman, 1994] defined as follows:

Definition 1. *A finite-horizon MDP is a tuple $(\mathcal{S}, \mathcal{A}, P, r, H, \rho)$, where $\mathcal{S}$, $\mathcal{A}$ and are the state space and the action space respectively. $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{A})$ and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are the state transition and reward function respectively, and H and ρ is the horizon and the initial state distribution.*

[1] A clear and concise example of the pitfalls of latent confounding can be found in the work by Ortega et al. [2021].

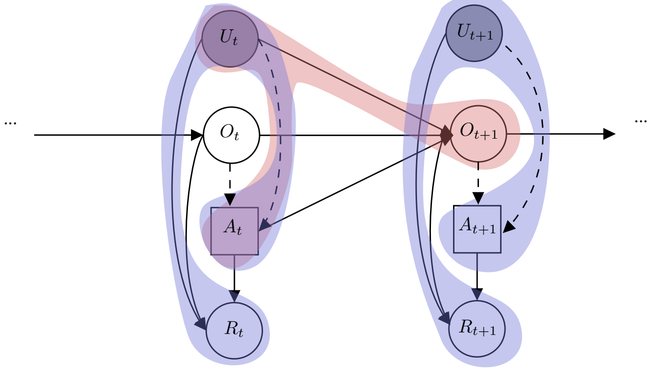


Figure 4: POMDP with i.i.d confounders: Dashed arrows represent dependencies in the offline data. Rewards (in **blue**) and next observations (in **red**) are confounded at each timestep t by the unobserved i.i.d. confounder U_t.

An MDP can be formulated as a Causal Bayesian Network or a Structural Causal Model [Buesing et al., 2019], since it allows us to evaluate the causal effect, in terms of the next states and the rewards of actions given the current state. To emphasize this, we will write that the transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ specifies the interventional distribution $\Pr(s'|s, \text{do}(a))$.

The value of any arbitrary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ in an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H, \rho)$ is defined as the scalar $V^{\pi}_{\mathcal{M}} = \mathbb{E}\Big[\sum_{t=0}^{H-1} R_t\Big]$, where at each timestep t, $R_t = r(S_t, A_t)$, $A_t \sim \pi(.|S_t)$, $S_{t+1} \sim P(.|S_t, A_t)$, and $S_0 \sim \rho(.)$. The goal is minimize the cumulative regret $\mathcal{R}(\pi, \mathcal{M})$, where $\mathcal{R}(\mathcal{M}, \pi) = V^{\pi^\star}_{\mathcal{M}} - V^{\pi}_{\mathcal{M}}$ with $\pi^\star = \arg\max_\pi V^{\pi}_{\mathcal{M}}$ being the optimal policy.

Given an MDP $\mathcal{M}$ sampled from a prior distribution f, PSRL [Osband et al., 2013] adopt a Bayesian approach for exploration by incorporating uncertainty to minimize the Bayesian regret $\mathcal{BR}(H, \pi) = \mathbb{E}_{\mathcal{M}\sim f}\big[\mathcal{R}(\mathcal{M}, \pi)\big]$. The PSRL policy π^{PS} is a sequence of episodic policies $\pi^{\text{PS}} = \{\pi_k\}_{k=0,..,E-1}$ constructed as follows. At the beginning of each episode $0 \leq k \leq E-1$, the PSRL policy has a prior f_k on the true parameters of the MDP with $f_0 = f$. The agent samples an MDP M_k from f_k and derives the optimal policy π_k with respect to M_k using a dynamic programming algorithm such as Value Iteration [Bellman, 1957]. The policy π_k interacts with the environment during the episode k and collects data $\mathcal{D}_{k+1}$ used to update the posterior f_{k+1}. With τ being the length of an episode, Osband et al. [2013] prove that the Bayesian regret of π^{PS} satisfy the following upper-bound:

$$\mathcal{BR}(H, \pi^{\text{PS}}) = O(\tau|\mathcal{S}|\sqrt{|\mathcal{A}|H\log(|\mathcal{S}||\mathcal{A}|H)}). \tag{1}$$

3 Problem formulation

We formulate our problem as a POMDP and show that the marginal problem is an MDP.

Definition 2. *(POMDP with i.i.d unobserved confounders) A* POMDP with i.i.d. unobserved confounders *is POMDP $\langle \mathcal{S}, \mathcal{A}, P, \mathcal{O}, \Omega, r, H, \rho\rangle$ where the state is factored as $\mathcal{S} = \mathcal{O} \times \mathcal{U}$, where $\mathcal{O}$ is the space of observations and $\mathcal{U}$ is the space of unobserved confounders. The transition model is factored as $P = P_u \times P_o$ and initial state distribution as $\rho = P_u \times \rho_o$ where ρ_o is the initial observation distribution. The observation model Ω is simply the projection of the state on the observation space $\mathcal{O}$. Hence, the state at time t, $S_t = (U_t, O_t)$ is generated by sampling the observation O_t from $P_o(.|O_{t-1}, U_{t-1}, do(A_{t-1}))$, and sampling the unobserved confounder U_t from $P_u(.)$, with $O_0 \sim \rho_o$. Furthermore we assume that the observation and action spaces are discrete and finite $\mathcal{O} = \{o_1, .., o_N\}$, $N = |\mathcal{O}|$ and $\mathcal{A} = \{a_1, .., a_K\}$, $K = |\mathcal{A}|$ and the reward is bounded: $r : \mathcal{S} \times \mathcal{A} \to [r_{min}, r_{max}]$ for some $r_{min} < r_{max}$.*

Figure 4 illustrates our setting, which corresponds to the setting of Zhang and Bareinboim [2019], with the difference that their work is stated in terms of the Structural Causal Model formulation [Pearl, 2000].

We assume that a stationary behavioral policy $\pi_\beta : \mathcal{S} \to \Delta(\mathcal{A})$ conditioning on both O_t and U_t while interacting with $\mathcal{M}$, to collect a dataset $\mathcal{D} = \{h_H^e\}_{e=1}^M$, consisting of M episodes of length H, where each episode $h_H^e = \{(o_t^e, a_t^e, r_t^e)\}_{t=0,..,H-1}$. That is the episodes consist of tuples of observations actions and rewards, **without** the unobserved confounders.

One is required generally in POMDPs to condition on the full history of observations or have access to the transition and observation models [Kaelbling et al., 1998, Spaan, 2012] to achieve optimal behavior. However, we show in the following proposition that under this specific structure of POMDPs, the marginalized problem is reduced to an MDP:

Proposition 3. *(Reduction to an MDP) Let r_m be the marginal reward function, i.e.:*

$$\forall o, a \in \mathcal{O} \times \mathcal{A} \ : r_m(o,a) = \mathbb{E}_u[R|o, do(a)] = \sum_{u \in \mathcal{S}_u} r(u,o,a)P_u(u),$$

and P_m the marginal transition model, i.e.:

$$\forall o, a, o' \in \mathcal{O} \times \mathcal{A} \times \mathcal{O} \ : P_m(o'|o, do(a)) = \sum_{u \in \mathcal{S}_u} P(o'|u,o,a)P_u(u),$$

The marginalized problem $\mathcal{M}_m = \langle \mathcal{O}, \mathcal{A}, P_m, r_m, H, \rho_o \rangle$ is a Markov Decision Process.

Proof. From Figure 4, it is clear that O_{t+1} is d-separated from $O_{t-1}, A_{t-1}, ..., O_0$ given $do(A_t)$ and O_t □

We denote by $\Pi : \{\pi : \mathcal{O} \to \mathcal{A}\}$ the set of memoryless deterministic policies. Our goal is to find a policy which maximizes the cumulative return $\pi^\star = \arg\max_{\pi \in \Pi} \mathbb{E}_{\mathcal{M}_m}[V^\pi(o)]$. Because the observation and the action spaces are finite, the policy $\pi^\star$ is guaranteed to be optimal as proven by Puterman [1994, Proposition 4.4.3].

4 Incorporating the causal natural bounds in PSRL

In this section, we introduce the natural bounds on causal effect and provide insights on their informativeness. These bounds induce a partial identification set on both the reward function and transition model of $\mathcal{M}_m$, and are used to derive CT-PSRL.

4.1 The natural bounds on causal effects

First, we introduce the natural bounds for the reward and transition model in the following Theorem:

Theorem 4. *The transition model and reward function of the marginal MDP $\mathcal{M}_m$ satisfy the following causal bounds for all $o, o' \in \mathcal{O}$ and $a \in \mathcal{A}$:*

$$\underbrace{Pr(o'|o,a)\pi_{\beta m}(a|o)}_{\alpha_p(o,a,o')} \leq P_m(o'|o, \mathrm{do}(a)) \leq \underbrace{1 - (1 - Pr(o'|o,a))\pi_{\beta m}(a|o)}_{\beta_p(o,a,o')}, \textit{ and,} \tag{2}$$

$$\begin{aligned} &\underbrace{\mathbb{E}(R|o,a))\pi_{\beta m}(a|o) + r_{min}(1 - \pi_{\beta m}(a|o))}_{\alpha_r(o,a)} \leq r_m(o,a) \\ &\leq \underbrace{\mathbb{E}(R|o,a)\pi_{\beta m}(a|o) + r_{max}\ (1 - \pi_{\beta m}(a|o))}_{\beta_r(o,a)}, \end{aligned} \tag{3}$$

where $\pi_{\beta m}$ is the marginal behavioral policy, $\pi_{\beta m}(a|o) = \sum_u \pi_\beta(a|o,u)P_u(u)$. We denote furthermore the width of these bounds by $\delta_p(o', a, o) = \beta_p(o', a, o) - \alpha_p(o', a, o)$ and $\delta_r(o,a) = \beta_r(o,a) - \alpha_r(o,a)$.

The proof is provided in the supplement and relies on the *natural bounds* of the tradition of Manski [1990] to the POMDP setting [Robins, 1989, Balke and Pearl, 1997, Manski and Nagin, 1998, Manski and Pepper, 2013], as is also applied to contextual bandits by Zhang and Bareinboim [2017]. We note that no such bounds exist when the action space is uncountable.

Corollary 5. *For binary rewards, the bounds* (3) *simplify to the same form as in Eq. 2:*

$$P(R=1|o,a)\pi_{\beta m}(a|o) \leq r_m(o,a) \leq 1-(1-P(R=1|o,a))\pi_{\beta m}(a|o).$$

The causal bounds in Theorem 4 involve only observational quantities, hence the behavioral policy π_β. Most importantly, they can be estimated from the offline data $\mathcal{D}$ using their empirical distributions that are guaranteed to be asymptotically consistent, hence converging to their true value when $|\mathcal{D}| \to \infty$.

4.2 On the informativeness of the natural bounds on causal effect

In this section we discuss the informativeness of the bounds in terms of the tightness of their widths. First, we observe the following:

Proposition 6. *(Tightness of the natural bounds) The width of the causal bounds satisfy the following for all o', a, o:*

$$\delta_p(o',a,o) = \frac{\delta_r(o,a)}{r_{max}-r_{min}} = 1-\pi_{\beta m}(a|o)$$

and we have $\sum_a \delta_p(o',a,o) = \sum_a \frac{\delta_r(o,a)}{r_{max}-r_{min}} = K-1$.

We refer the reader to Appendix A.2 for the proof. We deduce from Proposition 6 that the width of the causal bounds is independent of the dynamics of $\mathcal{M}_m$ and only depends on the concentration of the marginal behavioral policy $\pi_{\beta m}$. In addition, the sum of the widths is a constant, independent of the behavioral policy, and scales linearly with size of the action space K, which intuitively means that the bounds cannot be tight for all actions. However, we establish in the following propositions that we can construct tighter causal bounds in the case of multiple behavioral policies, which is motivated by real-world problems:

Proposition 7. *(Natural bounds with multiple behavioral policies) Assume we are provided with L datasets collected by L behavioral policies $\{\pi_j\}_{j=1,..,L}$ with reward and transition bounds $\{(\alpha_r^j(o,a), \beta_r^j(o,a))\}_{j=1,..,L}$ and $\{(\alpha_p^j(o,a,o'), \beta_p^j(o,a,o'))\}_{j=1,..,L}$ for every $o, o' \in \mathcal{O}$ and $a \in \mathcal{A}$. Let $\overline{\alpha_r}(o,a) = \max_j \alpha_r^j(o,a)$, $\overline{\alpha_p}(o,a,o') = \max_j \alpha_r^j(o,a,o')$, $\underline{\beta_r}(o,a) = \min_j \alpha_r^j(o,a)$, and $\underline{\beta_p}(o,a,o') = \min_j \beta_r^j(o,a,o')$. Then $\mathcal{M}_m$ satisfy the following inequalities:*

$$\overline{\alpha_p}(o,a,o') \leq P_m(o'|o, \mathrm{do}(a)) \leq \underline{\beta_p}(o,a,o')$$
$$\overline{\alpha_r}(o,a) \leq r_m(o,a) \leq \underline{\beta_r}(o,a),$$

for all $o, o' \in \mathcal{O}$ and $a \in \mathcal{A}$. Furthermore, we denote their widths by $\underline{\delta_p}(o',a,o) = \underline{\beta_p}(o',a,o) - \overline{\alpha_p}(o',a,o)$ and $\underline{\delta_r}(o,a) = \underline{\beta_r}(o,a) - \overline{\alpha_r}(o,a)$.

The proof is available in Appendix A.3. In the Multi-Armed Bandits problem, we show in the following proposition that when the behavioral policies satisfy certain conditions, the causal bounds derived from Proposition 7 are tight:

Proposition 8. *Consider a Multi-Armed Bandits problem, with 2 actions $\{a_0, a_1\}$ and two confounders $\mathcal{U} = \{u_0, u_1\}$, with $p(u_0) = p(u_1) = \frac{1}{2}$ and no observations. With binary rewards $R|A = a_j, U = u_i \sim Bern(\mu_{i,j})$, $\mu_{0,0} = \mu_{1,1} = \frac{1}{2} + \delta$ and $\mu_{0,1} = \mu_{1,0} = \frac{1}{2} - \delta$, for some $0 < \delta < \frac{1}{2}$.*

Then, for all $0 < \epsilon < 1$, there exist behavioral policies π_0 and π_1 with reward bounds $(\alpha_r^0(a), \beta_r^0(a))$ and $(\alpha_r^1(a), \beta_r^1(a))$ respectively, such that $\sum_a \underline{\delta}_r(a) \leq 2\epsilon$.

A proof is provided in Appendix A.4. Proposition 8 is a sufficient condition to get tight bounds in case we are provided with a set of offline data collected by different behavioral policies. The idea of the proof relies on the concentration of behavioral policies around disjoint sets of actions while keeping a uniform coverage of the actions. In this case, the overlap of the causal bounds with respect to each behavioral policy guarantees tightness for all actions. This comes evidently at the cost of a larger behavioral policy regret due to the uniform coverage, and a number of behavioral policies that scale with the size of the action space. We provide a visualization for this observation in one of our experimental domains in Figure 5.

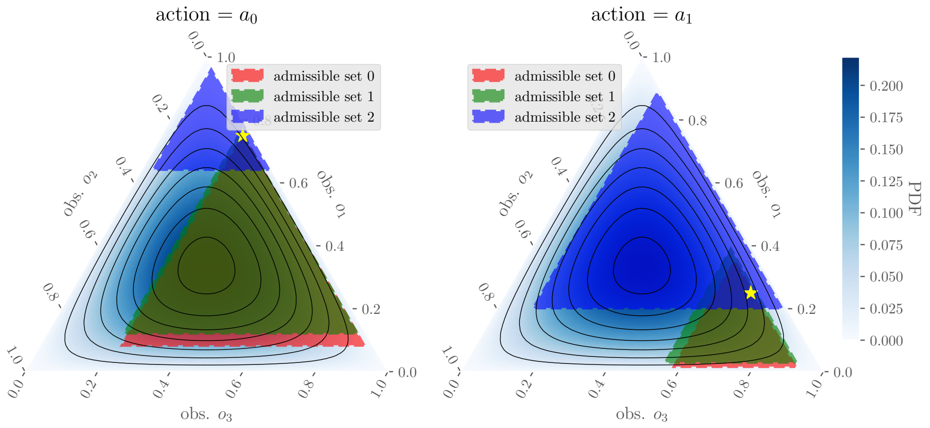


Figure 5: Uniform vs. near-optimal coverage in Switching Riverswim: Yellow stars depict the true parameters $P_m(.|o_1, do(a_0))$ and $P_m(.|o_1, do(a_1))$. Admissible sets 0 and 1 are induced by two near-optimal π_β, while admissible set 2 is induced by a suboptimal π_β. The intersections of 0 and 2, and 1 and 2 are the tightest around the true parameters thanks to the uniform coverage.

4.3 The Causally-Truncated PSRL algorithm (CT-PSRL)

In this section, we explain how the causal bounds are used to improve sample efficiency of PSRL in the online phase. We define the set of admissible transition models induced by the causal lower and upper bounds $\alpha_p(o, a, o')$ and $\beta_p(o, a, o')$ in Equation (2):

$$\mathcal{P}(\alpha_p, \beta_p) = \{P : \mathcal{O} \times \mathcal{A} \to \Delta(\mathcal{O}) : \ \forall o, a, o' : \ \alpha_p(o, a, o') \leq P(o'|o, a) \leq \beta_p(o, a, o')\}.$$

Furthermore, we define the set of admissible reward functions induced by the causal lower and upper bounds $\alpha_r(o, a)$ and $\beta_r(o, a)$ in Equation (3):

$$\mathcal{R}(\alpha_r, \beta_r) = \{r : \mathcal{O} \times \mathcal{A} \to [r_{min}, r_{max}] : \ \forall o, a : \ \alpha_r(o, a) \leq r(o, a) \leq \beta_r(o, a)\}.$$

Together, the set of admissible transition models and set of admissible reward functions define a partial identification set of the true MDP $\mathcal{M}_m$. We incorporate these bounds into PSRL by simply constraining the support of the posteriors f_k^r and f_k^P in step 4 to match the admissible sets. With this simple modification we derive the CT-PSRL in Algorithm 1, where in step 1 and step 2, the function TRUNCATEDISTRIBUTION truncates the densities of the transition and reward priors. More formally, the function TRUNCATEDISTRIBUTION takes as input some distribution with density f defined on some measurable space $\mathcal{X}$ and lower and upper bounds α and β and outputs a distribution with density $\tilde{f}$ such that $\forall x \in \mathcal{X}$:

$$\tilde{f}(x) = \frac{1}{\int_{\mathcal{X}(\alpha,\beta)} f(x')dx'} f(x) \cdot \mathbb{1}\{x \in \mathcal{X}(\alpha, \beta)\},$$

where $\mathbb{1}\{.\}$ is the indicator function and $\mathcal{X}(\alpha, \beta) = \{x \in \mathcal{X} : \ \alpha \leq x \leq \beta\}$.

When the causal bounds are uninformative, the support of the priors are not truncated and the regret upper-bound of CT-PSRL matches the PSRL regret upper-bound in Equation (1):

Corollary 9 (Regret upper bound of CT-PSRL with uninformative causal bounds)**.** *Let $\pi^{\text{CT-PS}}$ be the policy derived from Algorithm 1. If $\alpha_p(o, a, o') = \alpha_r(o, a) = 0$ and $\beta_p(o, a, o') = \beta_r(o, a) = 1$ for all $o, o' \in \mathcal{O}$ and $a \in \mathcal{A}$, then:*

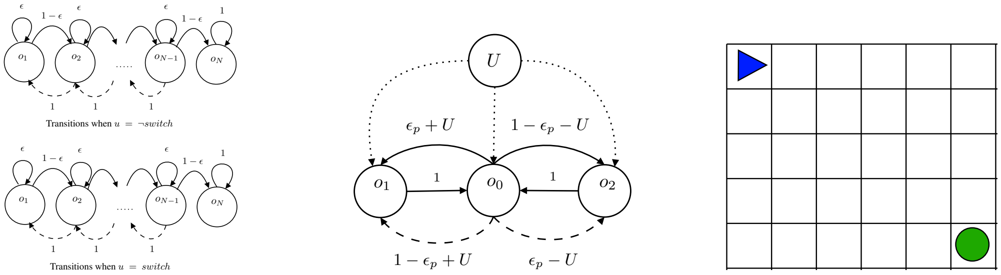
$$\mathcal{BR}(H, \pi^{\text{CT-PS}}) = O(\tau|\mathcal{O}|\sqrt{|\mathcal{A}|H \log(|\mathcal{O}||\mathcal{A}|H)}).$$

5 Experiments

In this section, we show how CT-PSRL experimentally improves sample efficiency in two relevant benchmark domains. We compare CT-PSRL to U-PSRL which is a PSRL agent with untruncated

Algorithm 1: Causally-Truncated PSRL (CT-PSRL)

```
Data: Reward bounds $(\alpha_r, \beta_r)$ and transition bounds $(\alpha_p, \beta_p)$, untruncated prior distribution of
      true MDP parameters: $f_0 = f_0^r \times f_0^P$
$\tilde{f}_0^P \leftarrow$ TRUNCATEDISTRIBUTION$(f_0^P, \alpha_p, \beta_p)$
$\tilde{f}_0^r \leftarrow$ TRUNCATEDISTRIBUTION$(f_0^r, \alpha_r, \beta_r)$
for episode $k = 0, .., E - 1$ do
    Sample $M_k = (r_k, P_k) \sim \tilde{f}_k^r(.) \times \tilde{f}_k^P(.)$
    Compute $\pi_k = \arg\max_{\pi \in \Pi} \mathbb{E}_{M_k}[V^\pi(o)]$
    $\mathcal{D} \leftarrow \{\}$
    for timestep $t = 1, ..., H$ do
        Sample and play $a_t \sim \pi_k(.|o_t)$
        Observe $r_t$ and $o_t$
        $\mathcal{D} \leftarrow \mathcal{D} \cup \{(o_t, a_t, r_t, o_{t+1})\}$
    end
    $\tilde{f}_{k+1}^P \leftarrow$ UPDATEPOSTERIORS$(\tilde{f}_k^P, \mathcal{D})$
    $\tilde{f}_{k+1}^r \leftarrow$ UPDATEPOSTERIORS$(\tilde{f}_k^r, \mathcal{D})$
end
```



(a) *Switching RiverSwim*

(b) *C-WinModel*

(c) *C-GridWorld*

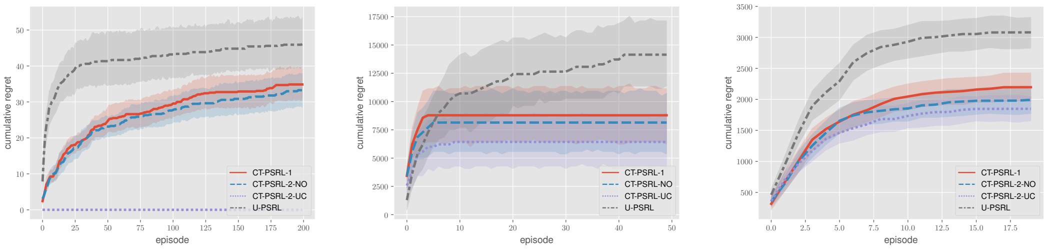
Figure 6: Environments illustrations: *Switching Riverswim* (Fig. 6a): Two possible environment configurations depending on the values of the confounder. Agent starts from the left and needs to reach right. *C-WinModel* (Fig. 6b): The agent needs to reach o_0 through o_1, continuous and dashed arrows represent transitions under actions a_0 and a_1. *C-GridWorld* Fig. 6c: The agent needs to reach the diagonally opposed corner starting from the upper-left corner.

uniform priorsẆe model the posteriors as Dirichlet distributions for both the transition and the reward since the rewards take finite values in all domains and we implement the truncation using rejection sampling. Furthermore, we create three variants of CT-PSRL: CT-PSRL-1, CT-PSRL-2-NO and CT-PSRL-3-UC where the number suffix indicate the number of datasets collected by different behavioral policies used to compute the bounds, and the type of coverage provided by the behavioral policies: "NO" stands for near-optimal and "UC" or uniform-coverage. The details of the behavioral policies for each domain as well as a detailed domain description are provided in Appendix B.

Switching RiverSwim: We construct a variant of the standard RiverSwim environment [Strehl and Littman, 2008] with switching directions as depicted in Figure 6a. The environment consists of N observations, two confounders, and two actions. The agent starts at the far left node and has to reach far left node. Depending on the value of the confounder, the effect of actions is switched.

C-WinModel: C-ModelWin [Bennett et al., 2021] depicted in Figure 6b is a confounded variant of the standard ModelWin introduced by Thomas and Brunskill [2016]. The environment consists of 3 observations, 2 actions and 2 confounders.

C-GridWorld: C-GridWorld [Bennett et al., 2021] is a 2-dimensional grid where the goal is to reach the top-right corner starting from the bottom-left. Similar to C-WinModel, there are 2 confounders and the observations consist of the position of the agent and the actions correspond to the 4 directions of movement.



(a) *Switching RiverSwim*

(b) *C-WinModel*

(c) *C-Gridworld*

Figure 7: Cumulative regrets in Switching RiverSwim (Fig. 6a), C-WinModel (Fig. 6b) and, C-Gridworld (Fig. 6c) of CT-PSRL leveraging one offline data CT-PSRL-1, CT-PSRL with two datasets with uniform coverage CT-PSRL-2-UC, CT-PSRL with two datasets with near-optimal coverage CT-PSRL-2-NO, and PSRL with untruncated uniform priors U-PSRL.

We run experiments across 5 random seeds in all environments and we visualize the cumulative regrets in Figure 7. We observe that CT-PSRL-1 outperforms U-PSRL. The advantage of CT-PSRL lays in the first few episodes when the tails of the posterior are long and the truncation provides a significant regret reduction. Clearly, the benefits of such truncation depend on the tightness of the causal bounds, which in turn depend on the behavioral policy. We additionally observe that incorporating additional bounds collected by a secondary behavioral policy improves considerably the performance. Finally, in both environments, CT-PSRL-2-UC outperforms CT-PSRL-2-NO, thanks to the tight bounds guaranteed by the uniform coverage. In the Switching RiverSwim domain, the bounds are tight enough to recover the optimal policy before the exploration phase.

6 Related work

Treating latent confounding has been addressed in various works in both the Causal Inference and RL literature, both in the pure offline [Bruns-Smith, 2021, Namkoong et al., 2020, Bennett et al., 2021] and hybrid (i.e., offline-online) setting [Zhang and Bareinboim, 2017, Gasse et al., 2023, Wang et al., 2021, Zhang and Bareinboim, 2019]. Furthermore, various structural assumptions have been considered: one-step confounding [Namkoong et al., 2020], i.i.d. confounders [Bruns-Smith, 2021, Zhang and Bareinboim, 2019, Bennett et al., 2021], and persistent confounding [Pace et al., 2023, Tennenholtz et al., 2022]. In general, the goal is to mainly tackle two problems: deriving worst-case bounds in off-policy evaluation [Namkoong et al., 2020, Bruns-Smith, 2021] and off-policy improvement [Pace et al., 2023, Gasse et al., 2023, Wang et al., 2021, Zhang and Bareinboim, 2019].

In the hybrid setting, one is required to incorporate the prior data into an exploration strategy. Exploration approaches can be grouped into two main categories, namely methods that rely on the Optimism in the Face of Uncertainty (OFU) framework, and the Posterior Sampling framework. Posterior Sampling for Reinforcement Learning (PSRL) [Osband et al., 2013] has been shown to hold many advantages over traditional approaches based on optimism [Osband and Van Roy, 2017]. Particularly, to avoid the computational untractability of efficient optimization on ellipsoidal confidence sets [Russo and Van Roy, 2014, Dani et al., 2008], the existing OFU approaches involve computing rectangular confidence sets that provide rather loose bounds on the true confidence sets [Osband and Van Roy, 2017]. In contrast, PSRL focuses on solving one problem instance.

Closest to our setting are works by Zhang and Bareinboim [2017, 2019] and by Gasse et al. [2023]. In the bandits case, Zhang and Bareinboim [2017] leverage causal bounds on the rewards to derive a variant of Thompson Sampling [Thompson, 1933] and kl-UCB [Garivier and Cappé, 2011] by truncating the posteriors and upper-confidence bounds. In the sequential setting, Zhang and Bareinboim [2019] derive an Upper Confidence Bound algorithm, based on similar causal bounds on the cumulative return and transition model. Our setting is equivalent to the one considered by Zhang and Bareinboim [2019], but their formulation is stated as an SCM. However, our causal bounds are not directly comparable to their bounds since they rely on the full history of observations rather than the action-observation pairs. Our work can be seen as an extension of the B-TS algorithm [Zhang and Bareinboim, 2017] to the sequential case.

Gasse et al. [2023] focus on the general POMDP case. In order to leverage offline data, they model the hybrid setting as an augmented POMDP by modelling the data regime in the causal graph. To construct a belief over the POMDP parameters, they treat the problem as a Maximum Likelihood Estimation [Mandl, 1974], and leverage history dependent causal bounds to truncate the belief. These bounds coincide with ours in the i.i.d confounders case. However, the Maximum Likelihood Estimation (MLE) is well-known not to be suited for exploration problems since it requires a strict identifiability assumption [Mandl, 1974]. When this assumption is violated, Borkar and Varaiya [1979] show that only closed-loop identification is guaranteed, and the MLE policy is not necessarily optimal under the true POMDP parameters.

7 Conclusions

Taking advantage of existing data is an important open question in reinforcement learning. In certain real-world settings, however, the decision maker (i.e., policy) that collected the data can base its decisions on additional information that has not been recorded. It is a well-known issue in offline RL that ignoring these so-called unobserved confounders can lead to poor performance. The core contribution of our work is that by leveraging offline data from multiple behavioral policies we can tightly bound the causal effect of interventions. We then showed that these bounds are even more informative when the behavioral policies uniformly cover the action space. We propose Causally-Truncated PSRL, an extension of the PSRL algorithm that exploits these bounds by truncating the posterior distributions. We demonstrate how CT-PSRL improves sample efficiency in three relevant domains, focusing on the effect of having datasets generated by different behavioral policies.

As future work, we would like to extend the theoretical result on the tightness of the bounds to the sequential case with an arbitrary number of behavioral policies. A limitation of our algorithm is that it relies on exact causal bounds which only hold when the offline data is sufficiently large. We would like to adapt our algorithm to incorporate inaccurate estimates of the causal bounds from an arbitrary number of offline data samples. Furthermore, we would like to provide theoretical guarantees of CT-PSRL in the form of tight regret-upper bounds and sample complexity bounds that depend on the width of the natural bounds. Finally, we would like to explore broader settings than the i.i.d. confounders case, which is a very promising direction for designing causal RL agents.

Acknowledgments

This research was performed as part of the Mercury Machine Learning Lab, a collaboration between Booking.com, the University of Amsterdam and Delft University of Technology.

References

A. Balke and J. Pearl. Bounds on Treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

A. Bennett, N. Kallus, L. Li, and A. Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1999–2007. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/bennett21a.html`.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5), 2021.

V. Borkar and P. Varaiya. Adaptive control of Markov chains, I: Finite parameter set. *IEEE Transactions on Automatic Control*, 24(6):953–957, 1979. doi: 10.1109/TAC.1979.1102191.

D. A. Bruns-Smith. Model-free and model-based policy evaluation when causality is uncertain. In *International Conference on Machine Learning*, pages 1116–1126. PMLR, 2021.

L. Buesing, T. Weber, Y. Zwols, N. Heess, S. Racaniere, A. Guez, and J.-B. Lespiau. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=BJG0voC9YQ`.

M. Chevalier-Boisvert, L. Willems, and S. Pal. Minimalistic gridworld environment for Openai gym. `https://github.com/maximecb/gym-minigrid`, 2018.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, volume 2, page 3, 2008.

P. Forré and J. M. Mooij. A Mathematical Introduction to Causality. 2023.

A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

M. Gasse, D. Grasset, G. Gaudron, and P.-Y. Oudeyer. Using confounded data in latent model-based reinforcement learning. *Transactions on Machine Learning Research*, 2023.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

P. Mandl. Estimation and control in markov chains. *Advances in Applied Probability*, 6(1):40–60, 1974.

C. F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2): 319–323, 1990.

C. F. Manski and D. S. Nagin. Bounding Disagreements about Treatment Effects: A Case Study of Sentencing and Recidivism. *Sociological Methodology*, 28:99–137, 1998.

C. F. Manski and J. V. Pepper. Deterrence and the death penalty: Partial identification analysis using repeated cross sections. *Journal of Quantitative Criminology*, 29:123–141, 2013.

H. Namkoong, R. Keramati, S. Yadlowsky, and E. Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

P. A. Ortega, M. Kunesch, G. Delétang, T. Genewein, J. Grau-Moya, J. Veness, J. Buchli, J. Degrave, B. Piot, J. Perolat, T. Everitt, C. Tallec, E. Parisotto, T. Erez, Y. Chen, S. Reed, M. Hutter, N. de Freitas, and S. Legg. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*, 2021.

I. Osband and B. Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2701–2710. PMLR, 8 2017.

I. Osband, D. Russo, and B. Van Roy. (More) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

A. Pace, H. Yèche, B. Schölkopf, G. Ratsch, and G. Tennenholtz. Delphic offline reinforcement learning under nonidentifiable hidden confounding. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

J. Pearl. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*, 19(2):3, 2000.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994. ISBN 978-0-47161977-2.

J. M. Robins. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pages 113–159, 1989.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

M. T. J. Spaan. Partially observable markov decision processes. In M. Wiering and M. van Otterlo, editors, *Reinforcement Learning: State-of-the-Art*, pages 387–414. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-27645-3.

A. L. Strehl and M. L. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

G. Tennenholtz, A. Hallak, G. Dalal, S. Mannor, G. Chechik, and U. Shalit. On covariate shift of latent confounders in imitation and reinforcement learning. In *International Conference on Learning Representations*, 2022.

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

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

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

J. Zhang and E. Bareinboim. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1778–1780, 2017.

J. Zhang and E. Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. *Advances in Neural Information Processing Systems*, 32, 2019.

A Proofs

A.1 Proof of Theorem 4

We first prove the following lemma:

Lemma 10. *The causal effect of the action on the transition probabilities and expected reward is respectively bounded by*

$$\begin{aligned} P(O_{t+1} = o', A_t = a | O_t = o) &\leq P(O_{t+1} = o' | O_t = o, \mathrm{do}(A_t = a)) \\ &\leq 1 - P(O_{t+1} \neq o', A_t = a | O_t = o) \end{aligned}$$

and

$$\begin{aligned} &E(R_t | O_t = o, A_t = a) P(A_t = a | O_t = o) + r_{min} \cdot P(A_t \neq a | O_t = o) \\ &\leq E(R_t | O_t = o, \mathrm{do}(A_t = a)) \\ &\leq E(R_t | O_t = o, A_t = a) P(A_t = a | O_t = o) + r_{max} \cdot P(A_t \neq a | O_t = o). \end{aligned}$$

for all $o, o' \in \mathcal{O}$ and $a \in \mathcal{A}$.

Proof. The logging data is generated following a Markov Decision Process, which is a Bayesian network, hence a simple SCM [Bongers et al., 2021]. The following is a rewriting of Theorem 7.5.2 and Corollary 7.5.3 of Forré and Mooij [2023], which are in turn an adaptations of Manski and Nagin [1998] and Manski and Pepper [2013].

In the SCM framework, the *potential outcome* $O_{t+1}^{\mathrm{do}(A_t=a)}$ is defined as any random variable that is almost surely equal to the random variable O_t in the SCM under the intervention $\mathrm{do}(A_t = a)$ (Bongers et al. [2021], Definition 8.6). Using this notation we have $P(O_{t+1}^{\mathrm{do}(A_t=a)} = o' | O_t = o) = P(O_{t+1} = o' | O_t = o, \mathrm{do}(A_t = a))$. These potential outcomes satisfy the *consistency* property $O_{t+1}^{\mathrm{do}(A_t=A_t)} = O_{t+1}$ (Forré and Mooij [2023], Theorem 7.5.1), using which we get

$$\begin{aligned} P(O_{t+1} = o', A_t = a | O_t = o) &= P(O_{t+1}^{\mathrm{do}(A_t=a)} = o', A_t = a | O_t = o) \\ &\leq P(O_{t+1}^{\mathrm{do}(A_t=a)} = o' | O_t = o) \\ &= P(O_{t+1}^{\mathrm{do}(A_t=a)} = o', A_t = a | O_t = o) \\ &\quad + P(O_{t+1}^{\mathrm{do}(A_t=a)} = o', A_t \neq a | O_t = o') \\ &\leq P(O_{t+1} = o', A_t = a | O_t = o) + P(A_t \neq a | O_t = o) \\ &= 1 - P(O_{t+1} \neq o', A_t = a | O_t = o). \end{aligned}$$

The above can also be proven without potential outcomes, using response variables instead [Balke and Pearl, 1997].

Following Manski and Pepper [2013] we write

$$\begin{aligned} E(R_t^{\mathrm{do}(A_t=a)} | O_t = o) &= E(R_t^{\mathrm{do}(A_t=a)} | O_t = o, A_t = a) P(A_t = a | O_t = o) \\ &\quad + E(R_t^{\mathrm{do}(A_t=a)} | O_t = o, A_t \neq a) P(A_t \neq a | O_t = o) \\ &= E(R_t | O_t = o, A_t = a) P(A_t = a | O_t = o) \\ &\quad + E(R_t^{\mathrm{do}(A_t=a)} | O_t = o, A_t \neq a) P(A_t \neq a | O_t = o) \end{aligned}$$

and since $E(R_t^{\mathrm{do}(A_t=a)} | O_t = o, A_t \neq a) \in [r_{min}, r_{max}]$, we get the desired result. □

By writing Lemma 10, in terms of the parameters of the marginalized MDP $\mathcal{M}_m$ and behavioral policy π_β and the observational quantities and dropping the dependency on time since the transition model, reward function and behavioral policy are time-homogeneous, we get for all $o, o' \in \mathcal{O}$ and $a \in \mathcal{A}$:

$$P(o', a | o) \leq P_m(o' | o, \mathrm{do}(A = a)) \leq 1 - \sum_{o'' \neq o'} P(o'', a | o) \tag{4}$$

and

$$\begin{aligned} &E(R | o, a) \pi_{\beta m}(a | o) + r_{min} \cdot (1 - \pi_{\beta m}(a | o)) \leq \\ &\quad r_m(o, a) \leq E(R | o, a) \pi_{\beta m}(a | o) + r_{max} \cdot (1 - \pi_{\beta m}(a | o)). \end{aligned}$$

By noticing that $\sum_{o'' \neq o'} P(o'', a | o) = \pi_{\beta m}(a | o) - P(o', a | o)$ and that $P(o', a | o) = P(o' | o, a) \pi_{\beta m}(a | o)$ and replacing these terms in Equation (4) we conclude the proof.

A.2 Proof of Proposition 6

Proposition 6. *(Tightness of the natural bounds) The width of the causal bounds satisfy the following:*

$$\beta_p(o', a, o) - \alpha_p(o', a, o) = 1 - \pi_{\beta m}(a|o)$$
$$\beta_r(o, a) - \alpha_r(o, a) = (r_{max} - r_{min})(1 - \pi_{\beta m}(a|o))$$

Furthermore:

$$\sum_a \beta_p(o', a, o) - \alpha_p(o', a, o) = K - 1$$
$$\sum_a \beta_r(o, a) - \alpha_r(o, a) = (r_{max} - r_{min})(K - 1)$$

for all o', a, o.

Proof. The proof is straightforward, let us first recall the definition of the causal bounds for the transition model: $\alpha_p(o', a, o) = P(o', a|o)$ and $\beta_p(o', a, o) = 1 - (\pi_{\beta m}(a|o) - P(o', a|o))$.

Then we have that, for all $o, o' \in \mathcal{O}$, $a \in \mathcal{A}$:

$$\begin{aligned}\beta_p(o', a, o) - \alpha_r(o', a, o) &= 1 - (\pi_\beta(a|o) - P(o', a|o)) - P(o', a|o) \\ &= 1 - \pi_{\beta m}(a|o)\end{aligned}$$

Therefore, by summing over actions:

$$\sum_{a \in \mathcal{A}} \beta_p(o', a, o) - \alpha_r(o', a, o) = \sum_{a \in \mathcal{A}} 1 - \pi_{\beta m}(a|o) = K - 1$$

Similarly, for the reward, $\alpha_r(o, a) = \mathbb{E}(R|o, a)\pi_{\beta m}(a|o) + r_{min}(1 - \pi_{\beta m}(a|o))$ and $\beta_r(o, a) = \mathbb{E}(R|o, a)\pi_{\beta m}(a|o) + r_{max}(1 - \pi_{\beta m}(a|o'))$, we have that, for all $o \in \mathcal{O}$, $a \in \mathcal{A}$:

$$\begin{aligned}\beta_r(o, a) - \alpha_r(o, a) &= \mathbb{E}(R|o, a)\pi_{\beta m}(a|o) + r_{max}(1 - \pi_{\beta m}(a|o')) - \\ &\qquad \mathbb{E}(R|o, a)\pi_{\beta m}(a|o) - r_{min}(1 - \pi_{\beta m}(a|o)) \\ &= r_{max}(1 - \pi_{\beta m}(a|o')) - r_{min}(1 - \pi_{\beta m}(a|o)) \\ &= (r_{max} - r_{min})(1 - \pi_{\beta m}(a|o'))\end{aligned}$$

and by summing over actions:

$$\sum_{a \in \mathcal{A}} \beta_r(o, a) - \alpha_r(o, a) = \sum_{a \in \mathcal{A}} 1 - \pi_{\beta m}(a|o) = (r_{max} - r_{min})(K - 1).$$

□

A.3 Proof of Proposition 7

Proof. The proof is straightforward. From theorem 4, the transition and reward satisfy for $o, o' \in \mathcal{O}$, $a \in \mathcal{A}$, and for all $j = 1, .., L$:

$$\begin{aligned}&\alpha_p^j(o, a, o') \le P_m(o'|o, \mathrm{do}(a)) \le \beta_p^j(o, a, o'), \\ &\alpha_r^j(o, a) \le r_m(o, a) \le \beta_r^j(o, a),\end{aligned}$$

Hence the lower and upper transition bounds hold for $\max_j \alpha_p^j(o, a, o')$ and $\min_j \beta_p^j(o, a, o')$, and the lower and upper reward bounds hold for $\max_j \alpha_r^j(o, a)$ and $\min_j \beta_r^j(o, a))$ respectively.

□

A.4 Proof of Proposition 8

Proposition 8. *Consider a Multi-Armed Bandits problem, with 2 actions $\{a_0, a_1\}$ and two confounders $\mathcal{U} = \{u_0, u_1\}$, with $p(u_0) = p(u_1) = \frac{1}{2}$ and no observations. With binary rewards $R|A = a_j, U = u_i \sim Bern(\mu_{i,j})$. $\mu_{0,0} = \mu_{1,1} = \frac{1}{2} + \delta$ and $\mu_{0,1} = \mu_{1,0} = \frac{1}{2} - \delta$, for some $0 < \delta < \frac{1}{2}$.*

Then, for all $0 < \epsilon < 1$, there exist behavioral policies π_0 and π_1 with reward bounds $(\alpha_r^0(a), \beta_r^0(a))$ and $(\alpha_r^1(a), \beta_r^1(a))$ respectively, such that

$$\sum_a \underline{\delta}_r(a) \leq 2\epsilon.$$

Proof. Consider two behavioral policies π_0 and π_1 such that:

$$\pi_j(a_i|u_0) = \begin{cases} 1 - \frac{\epsilon}{2} & \text{if } i = j \\ \frac{\epsilon}{2} & \text{if } i \neq j \end{cases} \qquad \text{and,} \qquad \pi_j(a_i|u_1) = \begin{cases} 1 - \frac{3\epsilon}{2} & \text{if } i = j \\ \frac{3\epsilon}{2} & \text{if } i \neq j \end{cases}$$

For ease of notation, we denote by $\pi_j(a_i)$ the marginal behavioral policies $\pi_j(a_i) = p(u_0)\pi_j(a_i|u_0) + p(u_1)\pi_j(a_i|u_1)$ such that:

$$\pi_j(a_i) = \begin{cases} 1 - \epsilon & \text{if } i = j \\ \epsilon & \text{if } i \neq j \end{cases}$$

and,

$$\begin{aligned} P(R = 1|a_i; \pi_j) &= P(R = 1, u_0|a_i; \pi_j) + P(R = 1, u_1|a_i; \pi_j) \\ &= P(R = 1|a_i, u_0; \pi_j)\frac{\pi_j(a_i|u_0)p(u_0)}{\pi_j(a_i)} + P(R = 1|a_i, u_1; \pi_j)\frac{\pi_j(a_i|u_1)p(u_1)}{\pi_j(a_i)} \end{aligned}$$

since $p(u_0) = p(u_1) = \frac{1}{2}$, we have:

$$P(R = 1|a_i; \pi_j) = \frac{\mu_{0,i}\pi_j(a_i|u_0) + \mu_{1,i}\pi_j(a_i|u_1)}{2\pi_j(a_i)}$$

For short we let $\tilde{\mu}_i^j = P(R = 1|A = a_i; \pi_j)$ We evaluate now $\tilde{\mu}_i^j$ for all (i, j) pairs. We have:

$$\begin{aligned} \tilde{\mu}_0^0 &= \frac{(\frac{1}{2} + \delta)(1 - \frac{\epsilon}{2}) + (\frac{1}{2} - \delta)(1 - \frac{3\epsilon}{2})}{2(1 - \epsilon)} = \frac{1}{2} + \frac{\delta\epsilon}{2(1 - \epsilon)} \\ \tilde{\mu}_1^0 &= \frac{(\frac{1}{2} - \delta)\frac{\epsilon}{2} + (\frac{1}{2} + \delta)\frac{3\epsilon}{2}}{2\epsilon} = \frac{1 + \delta}{2} \\ \tilde{\mu}_0^1 &= \frac{(\frac{1}{2} + \delta)\frac{\epsilon}{2} + (\frac{1}{2} - \delta)\frac{3\epsilon}{2}}{2\epsilon} = \frac{1 - \delta}{2} \\ \tilde{\mu}_1^1 &= \frac{(\frac{1}{2} - \delta)(1 - \frac{\epsilon}{2}) + (\frac{1}{2} + \delta)(1 - \frac{3\epsilon}{2})}{2(1 - \epsilon)} = \frac{1}{2} - \frac{\delta\epsilon}{2(1 - \epsilon)} \end{aligned}$$

From Corollary 5, we have:

$$\alpha_r^j(a_i) = \pi_j(a_i)\tilde{\mu}_i^j$$

and,

$$\beta_r^j(a_i) = 1 - (1 - \pi_j(a_i))\tilde{\mu}_i^j$$

Now we evaluate the causal bounds for each policy-action pair and their maximums and minimums. For the lower causal bound of action a_0:

$$\begin{cases} \alpha_r^0(a_0) = \frac{1-(1-\delta)\epsilon}{2} \\ \alpha_r^1(a_0) = \frac{(1-\delta)\epsilon}{2} \end{cases} \qquad \text{Hence,} \qquad \overline{\alpha_r}(a_0) = \max_j \alpha_r^j(a_0) = \begin{cases} \frac{(1-\delta)\epsilon}{2} \text{ if } \epsilon \geq \frac{1}{2(1-\delta)} \\ \frac{1-(1-\delta)\epsilon}{2} \text{ otherwise} \end{cases}$$

For the lower causal bound of action a_1:

$$\begin{cases} \alpha_r^0(a_1) = \frac{(1+\delta)\epsilon}{2} \\ \alpha_r^1(a_1) = \frac{1-(1+\delta)\epsilon}{2} \end{cases} \qquad \text{Hence,} \qquad \overline{\alpha_r}(a_1) = \max_j \alpha_r^j(a_1) = \begin{cases} \frac{(1+\delta)\epsilon}{2} \text{ if } \epsilon \geq \frac{1}{2(1+\delta)} \\ \frac{1-(1+\delta)\epsilon}{2} \text{ otherwise} \end{cases}$$

For the upper causal bound of action a_0:

$$\begin{cases} \beta_r^0(a_0) = \frac{1+(1+\delta)\epsilon}{2} \\ \beta_r^1(a_0) = \frac{2-(1+\delta)\epsilon}{2} \end{cases} \qquad \text{Hence,} \qquad \underline{\beta_r}(a_0) = \min_j \beta_r^j(a_0) = \begin{cases} \frac{2-(1+\delta)\epsilon}{2} \text{ if } \epsilon \geq \frac{1}{2(1+\delta)} \\ \frac{1+(1+\delta)\epsilon}{2} \text{ otherwise} \end{cases}$$

For the upper causal bound of action a_1:

$$\begin{cases} \beta_r^0(a_1) = \frac{2-(1-\delta)\epsilon}{2} \\ \beta_r^1(a_1) = \frac{1+(1-\delta)\epsilon}{2} \end{cases} \qquad \text{Hence,} \qquad \underline{\beta_r}(a_1) = \min_j \beta_r^j(a_1) = \begin{cases} \frac{2-(1-\delta)\epsilon}{2} \text{ if } \epsilon \geq \frac{1}{2(1-\delta)} \\ \frac{1+(1-\delta)\epsilon}{2} \text{ otherwise} \end{cases}$$

Therefore, we distinguish three cases depending on the value of ϵ as a function of the gap $\Delta = 2\delta$:

Case 1: $\epsilon < \frac{1}{2(1+\frac{\Delta}{2})}$:

In this case, we have:

$$\begin{cases} \overline{\alpha_r}(a_0) = \frac{1-(1-\delta)\epsilon}{2} \\ \underline{\beta_r}(a_0) = \frac{1+(1+\delta)\epsilon}{2} \\ \overline{\alpha_r}(a_1) = \frac{1-(1+\delta)\epsilon}{2} \\ \underline{\beta_r}(a_1) = \frac{1+(1-\delta)\epsilon}{2} \end{cases}$$

Hence,

$$\underline{\delta_r}(a_0) + \underline{\delta_r}(a_1) = \underline{\beta_r}(a_0) - \overline{\alpha_r}(a_0) + \underline{\beta_r}(a_1) - \overline{\alpha_r}(a_1) = 2\epsilon$$

Case 2: $\frac{1}{2(1+\frac{\Delta}{2})} \leq \epsilon < \frac{1}{2(1-\frac{\Delta}{2})}$:

In this case, we have:

$$\begin{cases} \overline{\alpha_r}(a_0) = \frac{1-(1-\delta)\epsilon}{2} \\ \underline{\beta_r}(a_0) = \frac{2-(1+\delta)\epsilon}{2} \\ \overline{\alpha_r}(a_1) = \frac{(1+\delta)\epsilon}{2} \\ \underline{\beta_r}(a_1) = \frac{1+(1-\delta)\epsilon}{2} \end{cases}$$

Hence,

$$\underline{\delta_r}(a_0) + \underline{\delta_r}(a_1) = 1 - 2\delta\epsilon = 1 - \Delta\epsilon$$

Since $\frac{1}{2(1+\frac{\Delta}{2})} \leq \epsilon$, we have:

$$\underline{\delta_r}(a_0) + \underline{\delta_r}(a_1) \leq 2\epsilon$$

.

Case 3**:** $\epsilon \geq \frac{1}{2(1+\frac{\Delta}{2})}$**:**

$$\begin{cases} \overline{\alpha_r}(a_0) = \frac{(1-\delta)\epsilon}{2} \\ \underline{\beta_r}(a_0) = \frac{2-(1+\delta)\epsilon}{2} \\ \overline{\alpha_r}(a_1) = \frac{(1+\delta)\epsilon}{2} \\ \underline{\beta_r}(a_1) = \frac{2-(1-\delta)\epsilon}{2} \end{cases}$$

$$\underline{\delta_r}(a_0) + \underline{\delta_r}(a_1) = 2(1-\epsilon)$$

Since $\epsilon \geq \frac{1}{2(1+\frac{\Delta}{2})} \geq \frac{1}{2}$, we have:

$$\underline{\delta_r}(a_0) + \underline{\delta_r}(a_1) \leq 2\epsilon$$

In summary $\forall 0 \leq \epsilon \leq 1$:

$$\underline{\delta_r}(a_0) + \underline{\delta_r}(a_1) = \underline{\beta_r}(a_0) - \overline{\alpha_r}(a_0) + \underline{\beta_r}(a_1) - \overline{\alpha_r}(a_1) \leq 2\epsilon$$

Which concludes the proof. □

B Environments details

B.1 Switching RiverSwim

B.1.1 Environment description

The Switching RiverSwim (Figure 6a) is a chain similar to RiverSwim introduced by Strehl and Littman [2008]. The environment consists of N observations { o_1, o_1,.., o_N}, two confounders { *switch*, $\neg$*switch*}, two actions {*left*, *right*}. The agent starts at o_1 and has to reach o_N. At each timestep the confounder is sampled at random independently from the observation and action. When the confounder is "$\neg$*switch*" the agent transitions to the left with probability $1-\epsilon$ and to the right with probability ϵ when taking action "*left*" and transitions to the right with probability 1 when taking action "*right*" However, when the confounder is "*switch*", the actions are switched and the agent transitions to the right with probability one if it takes action "*left*" and it transitions to the left with probability $1-\epsilon$ and to the right with probability ϵ when taking action "*right*". Whenever a transition to the right occurs, the agent receives a small positive reward $r \in (0, 1)$, otherwise it receives a reward of 0. Finally, once o_N is reached, the episode is terminated and the agent receives a reward of 1.

B.1.2 The behavioral policies

The behavioral policy observes the confounders value u, and at each timestep picks with probability $1-\epsilon_\pi$ $a = u$ and $a = \neg u$ with probability ϵ_π. For CT-PSRL-1, we use $\epsilon_\pi^0 = 0.1$. For CT-PSRL-2-UC, we use additional a behavioral policy with parameter $\epsilon_\pi^1 = 0.9$ and for CT-PSRL-2-NO, the additional behavioral policy has parameter $\epsilon_\pi^2 = 0.2$. For our simulation, we choose $r = 0.02$, $N = 3$, $H = 5$. For each behavioral policy, we collect a dataset of 50 000 transitions to compute the causal bounds.

B.2 C-ModelWin

B.2.1 Environment description

C-ModelWin [Bennett et al., 2021] depicted in Figure 6b is a confounded variant of the standard ModelWin introduced by Thomas and Brunskill [2016]. The environment consists of 3 observations, 2 actions and 2 confounders taking two values, 0.1 or 0.2 with probabilities 0.3 and 0.7 respectively. The agent starts in state o_0. When taking action a_0 at o_0 at time t, the agent transitions to o_1 with probability $\epsilon_p + U_t$ and to o_2 with probability $1 - \epsilon_p - U_t$. Otherwise, if it takes action a_1, it transitions to o_1 with probability $1 - \epsilon_p + U_t$ and to o_2 with probability $\epsilon_p - U_t$ and in both cases receives a reward of 0. While in o_1 or o_2, the agent transitions to o_0 with probability 1 independently from the action taken, and receives a reward of $r + 20U_t$ if it transitions from o_1 and a reward of $-r - 20U_t$ if it transitions from o_2.

B.2.2 The behavioral policies

The behavioral policy for takes action a_0 and a_1 with probabilities $1 - \epsilon_\pi - U_t$ and $\epsilon_\pi + U_t$ respectively, independently from the observation for some parameter ϵ_π. For CT-PSRL-1 we choose $\epsilon_\pi^0 = 0.1$ and for CT-PSRL-2-UC and CT-PSRL-2-NO consider two additional behavioral policies with parameters $\epsilon_\pi^1 = 0.7$ and $\epsilon_\pi^2 = 0.2$ respectively. Finally, we choose $\epsilon_p = 0.7$ and $r = 10$ similar to Bennett et al. [2021]. Similar to the Switching Riverswim environment, each offline dataset is composed of 50 000 transitions.

B.3 C-Gridworld

B.3.1 Environment description

C-GridWorld (Figure 6c) [Bennett et al., 2021] is a 2-dimensional grid where the goal is to reach the bottom-right corner starting from the top-left. The observations consist of the position of the agent and four actions: either move *up* a_0, *right* a_1, *down* a_2, or *left* a_3. The transitions are deterministic, i.e. the agent moves in the direction of the action taken. Similar to C-WinModel, at each timestep t the confounder U_t can take two possible values 0.1 or 0.2 with probabilities 0.3 and 0.7 respectively. The rewards are either $100 + 100U_t$ if the agent reaches the goal. Otherwise, the reward is $1 + 20U_t$ when taking action *down*, $1 + 30U_t$ for *right*, $-1 - 30U_t$ for *up*, and $-1 - 40U_t$ for *left*. We choose an episode length $H = 10$ and we run experiments for $K = 20$ episodes and we base our adaptation on the GridWorld implementation of Chevalier-Boisvert et al. [2018].

B.3.2 The behavioral policies

The behavioral policy chooses actions hierarchically, where it chooses to move either *down-right* with probability $1 - \epsilon_\pi - U_t$ or *top-left* with probability $\epsilon_\pi + U_t$. When choosing to move *top-left*, the behavioral policy moves *up* with probability $0.5\epsilon_\pi + U_t$ and *right* with probability $1 - 0.5\epsilon_\pi - U_t$. If the behavioral policy chooses to move *down-right*, the action is chosen depending on the position of the agent with respect to the diagonal defined by the *top-left* and *bottom-right* corners. If the agent is above the diagonal, it moves *down* with probability $1 - \epsilon_\pi - U_t$ and *right* with probability $\epsilon_\pi + U_t$. Conversely, if the agent is below the diagonal, it moves *down* with probability $\epsilon_\pi + U_t$ and *right* with probability $1 - \epsilon_\pi - U_t$. Finally when the agent is in the diagonal, the probability of moving *down* and *right* are $0.5\epsilon_\pi + 0.5U_t$ and $1 - 0.5\epsilon_\pi - 0.5U_t$ respectively. We choose a grid size of 5×5, and three behavioral policies with parameters $\epsilon_\pi \in \{0.3, 0.4, 0.7\}$, and each behavioral policy is used to collect a dataset of 100 000 transitions.