Robust Counterfactual Inference in Markov Decision Processes

Jessica Lally
King's College London
London, United Kingdom
jessica.lally@kcl.ac.uk

Milad Kazemi
King's College London
London, United Kingdom
milad.kazemi@kcl.ac.uk

Nicola Paoletti
King's College London
London, United Kingdom
nicola.paoletti@kcl.ac.uk

ABSTRACT

This paper addresses a key limitation in existing counterfactual inference methods for Markov Decision Processes (MDPs). To make counterfactual distributions identifiable, existing approaches assume a specific causal model of the system; however, there are typically many causal models consistent with the observed and interventional distributions of an MDP, each yielding different counterfactual probabilities. Thus, relying on a single model can limit the validity and usefulness of counterfactual inference. We propose a novel *non-parametric* approach that computes tight bounds on counterfactual transition probabilities across all compatible causal models. Unlike previous methods that require solving prohibitively large optimisation problems, our approach provides closed-form expressions for these bounds, making computation highly efficient even for large-scale MDPs. Using these bounds, we construct an *interval* counterfactual MDP, and identify robust counterfactual policies that optimise the worst-case reward over the uncertain MDP probabilities. We evaluate our method on various case studies, demonstrating improved robustness over existing methods.

KEYWORDS

Counterfactual Inference; Markov Decision Processes

ACM Reference Format:
Jessica Lally, Milad Kazemi, and Nicola Paoletti. 2026. Robust Counterfactual Inference in Markov Decision Processes. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/TXUQ4572

1 INTRODUCTION

Markov Decision Processes (MDPs) are a fundamental mathematical framework for modelling sequential decision-making processes under uncertainty, including reinforcement learning (RL) problems. However, evaluating RL-learnt policies can be challenging, especially in safety-critical domains like healthcare, where testing these policies directly on patients would be both risky and unethical.

Counterfactual inference of MDPs enables offline policy evaluation, i.e., without "deploying" the alternative policy into the environment. Given an observed sequence of actions and outcomes, counterfactual inference estimates what the outcome would have been if different actions had been taken. Counterfactual outcomes yielding higher rewards than the observation can serve as explanations for how the observed policy could be improved.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/TXUQ4572

Counterfactual inference is increasingly being applied to MDPs for various RL applications, e.g., for generating counterfactual explanations [23, 30, 31], facilitating policy transfer [13], and augmenting datasets with counterfactual paths to address data scarcity [17, 28], as well as to other sequential models, such as LLMs [5, 26]. These works compute counterfactual probabilities by assuming a specific causal model of the system (e.g., the Gumbel-max structural causal model (SCM) [23]). However, given an observation and MDP, the system's causal model is generally *non-identifiable*: there can be many causal models compatible with this data, each yielding different counterfactual probabilities [38]. As a result, any counterfactual analysis based on a single assumed causal model may be inaccurate, which is particularly concerning in safety-critical domains.

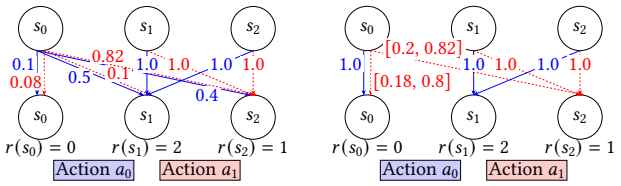
Partial counterfactual inference methods address this problem by computing bounds (as opposed to sharp values) on counterfactual probabilities, over *all* causal models compatible with the given data. One important work in this area is the canonical SCM approach developed by Zhang et al. [38], which formulates partial counterfactual inference as an optimisation problem. However, while this approach can compute counterfactual probability bounds in a wide range of settings (including those with unobserved confounders), their optimisation procedure is highly inefficient, as the number of constraints grows exponentially with the size of the MDP [6, 37, 38].

Contributions. In this paper, we demonstrate how the partial counterfactual inference approach of Zhang et al. [38] can be applied to MDPs, and prove this optimisation problem reduces to exact analytical bounds in the MDP setting (i.e., a Markovian setting with no unobserved confounders), thus successfully addressing the complexity of this optimisation problem. Next, using these bounds, we construct interval counterfactual MDPs, which we solve using pessimistic value iteration [19] to derive counterfactual policies that optimise the worst-case counterfactual rewards across all possible causal models compatible with the given data, ensuring robustness to uncertainty in the true underlying causal model[1]. Finally, we evaluate the average performance and robustness of our approach on a range of MDP benchmarks, demonstrating that our policies are more robust to causal model uncertainty than those derived from the Gumbel-max SCM. Thanks to the analytical bounds, our approach results in a speedup of 4-251x over the Gumbel-max SCM. For the full paper, including appendices and proofs, see [14].

2 BACKGROUND

In this section, we provide background on counterfactual inference and an overview of related work on counterfactual inference in MDPs and partial counterfactual inference.

[1]This notion of robustness aligns with existing work on robust MDPs, which design policies resilient to uncertainty in the transition probabilities [21], but differs from other notions, e.g., robustness to out-of-distribution environments [25].



(a) Interventional probabilities for two actions a_0 (blue, solid) and a_1 (red, dashed). Optimal policy: $\forall s \in \mathcal{S}, \pi^*(s) = a_0$.

(b) Counterfactual probabilities given observed transition, $s_0, a_0 \to s_0$. Optimal counterfactual policy: $\pi^*_{CF}(s_0) = a_1, \pi^*_{CF}(s_1) = a_0, \pi^*_{CF}(s_2) = a_0$.

Figure 1: Counterfactual inference in a toy example MDP.

Markov Decision Processes. MDPs are a class of stochastic models for representing sequential decision-making processes. In an MDP $\mathcal{M}$, at each step t, an agent in state s_t performs some action a_t determined by a policy π. The agent then transitions to a new state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$, and receives a reward $R(s_t, a_t)$. Formally, an MDP is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, P_I, R)$ where $\mathcal{S}$ is the discrete *state space*, $\mathcal{A}$ is the set of *actions*, $P : (\mathcal{S} \times \mathcal{A} \times \mathcal{S}) \to [0, 1]$ is the *transition probability* function, $P_I : \mathcal{S} \to [0, 1]$ is the *initial state distribution*, and $R : (\mathcal{S} \times \mathcal{A}) \to \mathbb{R}$ is the *reward function*. A (deterministic) *policy* π for $\mathcal{M}$ is a function $\pi : \mathcal{S} \to \mathcal{A}$. A path τ of $\mathcal{M}$ under policy π is a sequence $\tau = (s_0, a_0), \ldots, (s_{T-1}, a_{T-1})$ where $T = |\tau|$ is the path length, $P_I(s_0) > 0$, $a_t = \pi(s_t)$ for all $t = 0, \ldots, T-1$, and $P(s_{t+1} \mid s_t, a_t) > 0$ for all $t = 0, \ldots, T-2$.

Counterfactual Inference. Structural causal models (SCMs) [9, 24] provide a mathematical framework for causal inference. Formally, a SCM is a tuple $\mathcal{C} = (\mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}))$, where $\mathbf{V}$ is the set of endogenous (observed) variables, $\mathbf{U}$ is a set of exogenous (unobserved) variables, $P(\mathbf{U})$ is a joint distribution over the possible values of each $U \in \mathbf{U}$, and $\mathcal{F}$ is a set of structural equations where each $f_V \in \mathcal{F}$ determines the value of endogenous variable V, given a fixed realisation of $U_V \in \mathbf{U}$ and set of direct causes (parents) $\mathbf{PA}_V \subseteq \mathbf{V}$. A causal graph can be defined from an SCM by drawing, for every $V \in \mathbf{V}$, directed edges from the nodes in $\mathbf{PA}_V$ and U_V into V. In the following, we denote random variables with capital letters and specific values of the variables in lowercase.

SCMs enable the evaluation of causal effects, that is, how the distribution of some (outcome) variable Y changes after applying a so-called *intervention* $X \leftarrow x$ on a (treatment) variable X. Such an intervention corresponds to replacing the structural assignment of X with the constant value x. We can also perform *counterfactual inference* to estimate, given an observation $\mathbf{V} = \mathbf{v}$, the hypothetical values of $\mathbf{V}$ had we applied some intervention. Counterfactual inference involves inferring the value of the exogenous variables present in the observation $\mathbf{v}$, i.e., $P(\mathbf{U} \mid \mathbf{v})$, then evaluating the structural equations by replacing $P(\mathbf{U})$ with the inferred $P(\mathbf{U} \mid \mathbf{v})$, and applying the intervention.

Counterfactual Inference in MDPs. MDPs can be represented as SCMs with the causal graph in Figure 2 and the following structural equations:

$$S_{t+1} = f(S_t, A_t, U_t); \quad A_t = \pi(S_t); \quad S_0 = f_I(U_I) \tag{1}$$

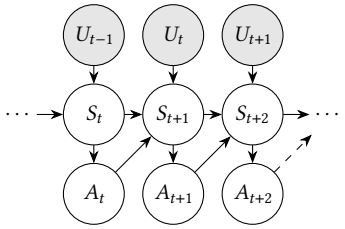


Figure 2: MDP causal graph. White nodes represent endogenous/observable variables; grey nodes represent exogenous/unobserved variables.

In the MDP context, the interventional distribution is the MDP's transition probabilities[2], and an observation (or realisation) is a path of the MDP, i.e., a sequence of transitions (s_t, a_t, s_{t+1}). Counterfactual inference is applied at the level of the individual timesteps: given the observed transition $s_t, a_t \to s_{t+1}$ at time t, we can evaluate the counterfactual probability of reaching outcome $S_{t+1} = \widetilde{s}'$, had we been in state $S_t = \widetilde{s}$ and performed action $A_t = \widetilde{a}$. Formally, we define this counterfactual probability $\widetilde{P}_t(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})$ as:

$$\widetilde{P}_t(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) = P(S_{t+1(S_t=\widetilde{s}, A_t=\widetilde{a})} = \widetilde{s}' \mid S_t = s_t, A_t = a_t, S_{t+1} = s_{t+1}),$$

where $S_{t+1(S_t=\widetilde{s}, A_t=\widetilde{a})}$ denotes variable S_{t+1} in the SCM after the intervention $S_t \leftarrow \widetilde{s}$ and $A_t \leftarrow \widetilde{a}$.

These counterfactual probabilities can be combined across multiple timesteps to construct a non-stationary *counterfactual* MDP [30]. This construction allows us to reason about counterfactual outcomes given an observed path.

Definition 2.1 (Counterfactual MDP (CFMDP)). *Given an observed path τ and MDP $\mathcal{M}$, the corresponding counterfactual MDP is a tuple $\widetilde{\mathcal{M}}_\tau = (\mathcal{S}^+, \mathcal{A}, \widetilde{P}, \widetilde{P}_I, \mathcal{R}^+)$ where $\mathcal{S}^+$ and $\mathcal{R}^+$ are the state space and reward function of $\mathcal{M}$, augmented with time; and:*

(1) $\widetilde{P}_I(s) = 1$ if $s = s_0$, 0 otherwise;

(2) $\widetilde{P}_t(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) = P(S_{t+1(S_t=\widetilde{s}, A_t=\widetilde{a})} = \widetilde{s}' \mid S_t = s_t, A_t = a_t, S_{t+1} = s_{t+1}), \forall t \in \{0, \ldots, |\tau| - 1\}$.

Counterfactual inference in MDPs is challenging because there are typically many SCMs consistent with the same observation and interventional distribution, yet they produce different counterfactual probabilities [38]. Most existing works avoid this problem by assuming a particular causal model to make counterfactual probabilities identifiable. One such model is the Gumbel-max SCM [23], which has been widely used for counterfactual analysis in MDPs [2, 12, 13, 22, 30, 39] due to its desirable property of counterfactual stability (see Definition 3.3). Notably, Tsirtsis et al. [30] applied the Gumbel-max SCM to derive alternative sequences of actions, i.e., counterfactual paths, that would have produced a higher reward than the observed path[3]. In this sense, counterfactual paths can be seen as *counterfactual explanations* for how the observed policy could have been improved. However, since the Gumbel-max SCM is just one of many possible SCMs compatible with a given

[2] in that if we apply an intervention $S_t \leftarrow s$ and $A_t \leftarrow a$, then the interventional distribution of S_{t+1} corresponds to the MDP transition probabilities $P(\cdot \mid s, a)$.

[3] See Appendix A for more details on the construction of counterfactual MDPs using the Gumbel-max SCM. There also exist other works which apply bijective causal models [20, 31], or genetic algorithms [7] to generate counterfactual paths.

	Binary outcomes	Categorical outcomes
Exact bounds	[1, 4, 11]	[15, 38], **ours**
Approximate bounds	[18, 27]	[6, 33–38][4]

Table 1: Related works on partial counterfactual inference.

MDP, the derived counterfactual MDP may be inaccurate, leading to unreliable counterfactual explanations.

Partial Counterfactual Inference. Partial counterfactual inference methods bound counterfactual outcomes by considering all SCMs compatible with observational and interventional data. Key works are summarised in Table 1. Notably, the *canonical SCM* approach, developed by Zhang et al. [38], constructs a linear or polynomial optimisation problem (depending on the causal graph structure) that identifies exact counterfactual probability bounds. However, this optimisation is very inefficient, as the number of linear constraints grows exponentially with the number and cardinality of the endogenous variables. Thus, research has shifted from exact methods to developing approximation methods. One notable exception is recent work by Li and Pearl [15], who identified counterfactual probability bounds for categorical systems, given observational and interventional distributions. While their bounds are exact (that is, capturing all causal models compatible with the observation and interventional distributions), their bounds are often uninformative, i.e., very wide or trivial ([0, 1]), because they make no additional assumptions about the underlying causal model. In a multi-step setting such as MDPs, having many transitions with loose or trivial bounds can quickly compound over several time steps, severely limiting the usefulness of counterfactual inference. In this paper, we propose adding two common and reasonable assumptions to tighten the bounds and improve the usefulness of counterfactual inference. We discuss the similarities and differences between their work and ours in more detail in Section 4.

3 PARTIAL COUNTERFACTUAL INFERENCE VIA CANONICAL SCMS

Zhang et al. [38] introduced a family of canonical SCMs capable of capturing all possible counterfactual distributions for any causal graph. Their partial counterfactual inference approach optimises over these canonical SCMs to identify exact counterfactual probability bounds. In this section, we demonstrate how this approach can be applied to MDPs, and how additional assumptions can be incorporated into the optimisation problem to ensure counterfactuals are plausible and useful, addressing issues in existing work.

3.1 Optimisation Procedure

To perform partial counterfactual inference in MDPs, we must convert the MDP SCM (1) to its equivalent canonical SCM. To this end, we need to identify the c-component of the exogenous variable U_t in the MDP causal graph (Figure 2), which is defined as follows:

Definition 3.1 (c-component [29]). *Given a causal graph $\mathcal{G}$, a subset of its endogenous variables $\mathbf{C} \subseteq \mathbf{V}$ is a **c-component** if any two variables $V_i, V_j \in \mathbf{C}$ are connected by a sequence of bi-directed edges $V_i \leftarrow U_k$ and $V_j \leftarrow U_k$, where each U_k is an exogenous parent shared by V_i and V_j.*

Definition 3.2 (Canonical SCM [38]). *A **canonical SCM** is a tuple $C = \langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, P \rangle$, where:*

1. *For every endogenous variable $V \in \boldsymbol{V}$, its values v are determined by a structural equation $v \leftarrow f_V(pa_V, u_V)$ where, for any pa_V and u_V, $f_V(pa_V, u_V)$ is contained within a finite domain Ω_V.*
2. *For every exogenous $U \in \boldsymbol{U}$, its values u are drawn from a finite domain Ω_U, where the cardinality of U is equal to the total number of functions that map all possible inputs $pa_V \in \Omega_{PA_V}$ to values $v \in V$ for every endogenous V in the c-component covering U[5], i.e., $|U| = \prod_{V \in \mathbf{C}(U)} |\Omega_{PA_V} \mapsto \Omega_V|$*

The c-component covering the single exogenous variable U_t in an MDP is $\mathbf{C}(U_t) = \{S_{t+1}\}$. Therefore, for an MDP, $|U|$ is equal to the total number of functions that map all possible values of S_t and A_t (i.e., all possible combinations of states $s \in \mathcal{S}$ and actions $a \in \mathcal{A}$) to all possible values of S_{t+1}. Thus, the cardinality of U_t in the canonical MDP SCM is $|U_t| = |\mathcal{S}|^{|\mathcal{S}|\times|\mathcal{A}|}$. Each value $u_t \in U_t$ indexes a unique structural equation, which deterministically maps all possible state-action pairs (s, a) to a next state s'.

Given the canonical SCM representation of an MDP and an observed transition $s_t, a_t \rightarrow s_{t+1}$, we can define an optimisation procedure to find the minimum and maximum counterfactual probabilities for every transition [38]. The first step is to define a mapping between exogenous values $u_t \in U_t$ and the structural equation they index. We define an indicator variable $\mu \in \{0, 1\}^{|\mathcal{S}|\times|\mathcal{A}|\times|U_t|\times|\mathcal{S}|}$ such that, for any s_t, a_t, u_t and s_{t+1}:

$$\mu_{s_t,a_t,u_t,s_{t+1}} = \begin{cases} 1 & \text{if } f(s_t, a_t, u_t) = s_{t+1} \\ 0 & \text{otherwise} \end{cases}$$

Next, we define a vector $\theta \in \mathbb{R}^{|U_t|}$, where each θ_{u_t} represents the probability $P(U_t = u_t)$. Each instantiation of θ defines a unique SCM. Given an observed transition $s_t, a_t \rightarrow s_{t+1}$, we can write the counterfactual probability $\widetilde{P}_t(\widetilde{s'} \mid \widetilde{s}, \widetilde{a})$ of any transition $\widetilde{s}, \widetilde{a} \rightarrow \widetilde{s'}$ in terms of μ and θ:

$$\widetilde{P}_t(\widetilde{s'} \mid \widetilde{s}, \widetilde{a}) = \frac{\sum_{u_t=1}^{|U_t|} \mu_{\widetilde{s},\widetilde{a},u_t,\widetilde{s'}} \cdot \mu_{s_t,a_t,u_t,s_{t+1}} \cdot \theta_{u_t}}{P(s_{t+1} \mid s_t, a_t)} \tag{2}$$

To identify the counterfactual probability bounds, we search over all values of θ consistent with the observed transition and interventional data (i.e., the MDP's transition probabilities). This optimisation problem is defined as:

$$\begin{aligned} &\min/\max_\theta \sum_{u_t=1}^{|U_t|} \mu_{\widetilde{s},\widetilde{a},u_t,\widetilde{s'}} \cdot \mu_{s_t,a_t,u_t,s_{t+1}} \cdot \theta_{u_t} \\ &\text{s.t.} \sum_{u_t=1}^{|U_t|} \mu_{s,a,u_t,s'} \cdot \theta_{u_t} = P(s' \mid s, a), \forall s, a, s' \\ &\qquad 0 \le \theta_{u_t} \le 1, \forall u_t, \qquad \sum_{u_t=1}^{|U_t|} \theta_{u_t} = 1 \end{aligned} \tag{3}$$

[4] Zhang et al. [38] provide both an exact and approximation method in their paper.

[5] $\Omega_{PA_V} \mapsto \Omega_V$ denotes the set of all possible mappings from values in Ω_{PA_V} to values in Ω_V.

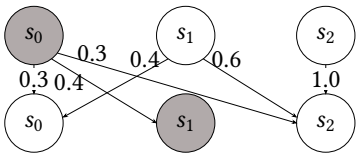


Figure 3: Example MDP where Gumbel-max produces unintuitive CF probabilities. The observed path is $s_0 \to s_1$.

3.2 Incorporating Additional Assumptions

While this optimisation correctly considers all SCMs consistent with the interventional and observational data, it can result in wide or trivial $[0, 1]$ bounds. As previously discussed, having many MDP transitions with loose or trivial bounds can severely limit the usefulness of counterfactual inference in MDP settings. However, incorporating reasonable assumptions about the causal model can help to tighten counterfactual probability bounds enough to derive useful (but still robust) counterfactual policies. In this work, we adopt two assumptions (similar to those in [10, 16]) when deriving counterfactual MDPs. The first assumption is *counterfactual stability* [23], which, in the context of an MDP, is defined as:

Definition 3.3 (Counterfactual stability). *An MDP SCM* (1) *satisfies counterfactual stability if, given we observed the transition $s_t, a_t \to s_{t+1}$ at time t, the counterfactual outcome under a different state-action pair $(\widetilde{s}, \widetilde{a})$ will not change to some $S_{t+1} = \widetilde{s}' \neq s_{t+1}$ unless $\frac{P(s_{t+1} \mid \widetilde{s}, \widetilde{a})}{P(s_{t+1} \mid s_t, a_t)} < \frac{P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})}{P(\widetilde{s}' \mid s_t, a_t)}$.*

However, even causal models that assume counterfactual stability can yield unreasonable counterfactual probabilities. Consider the toy MDP in Figure 3. Table 2 compares the counterfactual transition probabilities obtained from the optimisation approach in Eq. (3) (with no assumptions) and the Gumbel-max SCM (which satisfies counterfactual stability). Under the Gumbel-max SCM, the counterfactual probability of transition $s = 1, a = 0 \to s' = 2$ (highlighted in Table 2) is greater than its nominal probability, even though state $s' = 2$ was reachable from the observed state $s_t = 0$, but not observed (see Appendix B for further explanation). Arguably, a state that was reachable from the observed state-action pair, but was not observed, should not become more likely in the counterfactual world (i.e., under the counterfactual state-action pair). To formalise this intuition, we introduce *counterfactual monotonicity*[6]:

Definition 3.4 (Counterfactual monotonicity). *An MDP SCM* (1) *satisfies counterfactual monotonicity if, upon observing the transition $s_t, a_t \to s_{t+1}$, then, $\forall \widetilde{s} \in \mathcal{S}, \widetilde{a} \in \mathcal{A}$:*

(Mon1) $\widetilde{P}_t(s_{t+1} \mid \widetilde{s}, \widetilde{a}) \geq P(s_{t+1} \mid \widetilde{s}, \widetilde{a})$ (i.e., observing an outcome cannot make it less likely in the counterfactual world), and

(Mon2) $\forall \widetilde{s}' \neq s_{t+1}$ with $P(\widetilde{s}' \mid s_t, a_t) > 0$, $\widetilde{P}_t(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) \leq P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})$ (i.e., not observing a possible outcome cannot make it more likely in the counterfactual world).

[6]This is different from other definitions of monotonicity, which assume an ordering on interventions and outcomes, e.g., [32].

s	a	s'	$P(s' \mid s, a)$	Optimisation (3)		Gumbel-Max [23]	Optimisation (3)-(6)	
				LB	UB		LB	UB
0	0	0	0.3	0.0	0.0	0.0	0.0	0.0
0	0	1	0.4	1.0	1.0	1.0	1.0	1.0
0	0	2	0.3	0.0	0.0	0.0	0.0	0.0
1	0	0	0.4	0.0	1.0	0.35	0.4	0.4
1	0	1	0.0	0.0	0.0	0.0	0.0	0.0
1	**0**	**2**	**0.6**	**0.0**	**1.0**	**0.65**	**0.6**	**0.6**
2	0	0	0.0	0.0	0.0	0.0	0.0	0.0
2	0	1	0.0	0.0	0.0	0.0	0.0	0.0
2	0	2	1.0	1.0	1.0	1.0	1.0	1.0

Table 2: Counterfactual transition probabilities produced by various methods.

The counterfactual stability (4) and monotonicity assumptions (5,6) can be added as constraints to the optimisation problem (3) as:

$$\widetilde{P}_t(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) = 0 \quad \text{if } \frac{P(s_{t+1} \mid \widetilde{s}, \widetilde{a})}{P(s_{t+1} \mid s_t, a_t)} > \frac{P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})}{P(\widetilde{s}' \mid s_t, a_t)} \text{ and } P(\widetilde{s}' \mid s_t, a_t) > 0 \tag{4}$$

$$\widetilde{P}_t(s_{t+1} \mid \widetilde{s}, \widetilde{a}) \geq P(s_{t+1} \mid \widetilde{s}, \widetilde{a}) \quad \text{if } P(s_{t+1} \mid \widetilde{s}, \widetilde{a}) > 0 \tag{5}$$

$$\widetilde{P}_t(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) \leq P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) \quad \text{if } P(\widetilde{s}' \mid s_t, a_t) > 0, \forall \widetilde{s}' \neq s_{t+1} \tag{6}$$

4 DERIVING ANALYTICAL BOUNDS

While the optimisation in Eq. (3)-(6) yields exact bounds, it can be very inefficient as the number of constraints grows exponentially with the sizes of the state and action spaces. However, in the MDP setting (Markovian, no unobserved confounders), we have proven that this linear optimisation problem always reduces to exact closed-form solutions, illustrated below.

4.1 Analytical Bounds

The support of a state-action pair $(\widetilde{s}, \widetilde{a})$ is defined as the set of next states with nonzero probability when action $\widetilde{a}$ is taken in state $\widetilde{s}$. Given the observed transition $s_t, a_t \to s_{t+1}$, the counterfactual probability of any transition $\widetilde{s}, \widetilde{a} \to \widetilde{s}'$ depends only on whether the state-action pair $(\widetilde{s}, \widetilde{a})$ is the observed state-action pair (s_t, a_t), or whether its support is disjoint from or overlaps with the support of the observed state-action pair. The lower and upper bounds, denoted by $\widetilde{P}_t^{LB}$ and $\widetilde{P}_t^{UB}$, are given in Theorems 4.1-4.3. For clarity, we provide a proof sketch below and defer to the full proof in Appendix D.1.

Theorem 4.1. *For the observed state-action pair (s_t, a_t), the linear program will produce the following bounds:*

$$\widetilde{P}_t^{LB}(s_{t+1} \mid s_t, a_t) = \widetilde{P}_t^{UB}(s_{t+1} \mid s_t, a_t) = 1$$
$$\forall \widetilde{s}' \in \mathcal{S} \setminus \{s_{t+1}\}, \widetilde{P}_t^{LB}(\widetilde{s}' \mid s_t, a_t) = \widetilde{P}_t^{UB}(\widetilde{s}' \mid s_t, a_t) = 0$$

Theorem 4.2. *For state-action pairs $(\widetilde{s}, \widetilde{a}) \neq (s_t, a_t)$ which have disjoint support from the observed (s_t, a_t), the linear program will produce, $\forall \widetilde{s}' \in \mathcal{S}$, the following bounds:*

$$\widetilde{P}_t^{UB}(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) = \begin{cases} \frac{P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})}{P(s_{t+1} \mid s_t, a_t)} & \text{if } P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) < P(s_{t+1} \mid s_t, a_t) \\ 1 & \text{otherwise} \end{cases}$$

$$\widetilde{P}_t^{LB}(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) = \begin{cases} \dfrac{P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) - \big(1 - P(s_{t+1} \mid s_t, a_t)\big)}{P(s_{t+1} \mid s_t, a_t)} & \text{if } P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) > 1 - P(s_{t+1} \mid s_t, a_t) \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 4.3. *For state-action pairs $(\widetilde{s}, \widetilde{a}) \neq (s_t, a_t)$ which have overlapping support with the observed (s_t, a_t), the linear program will produce, $\forall \widetilde{s}' \in \mathcal{S}$, the following upper bounds for $\widetilde{P}_t^{UB}(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})$:*

$$\begin{cases} \dfrac{\min\Big(P(s_{t+1} \mid s_t, a_t), P(s_{t+1} \mid \widetilde{s}, \widetilde{a})\Big)}{P(s_{t+1} \mid s_t, a_t)} & \text{if } \widetilde{s}' = s_{t+1} \\ 0 & \text{if CS conditions} \\ \min\Big(P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}), 1 - P(s_{t+1} \mid \widetilde{s}, \widetilde{a})\Big) & \text{if } P(\widetilde{s}' \mid s_t, a_t) > 0 \\ \min\left(1 - P(s_{t+1} \mid \widetilde{s}, \widetilde{a}), \dfrac{P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})}{P(s_{t+1} \mid s_t, a_t)}\right) & \text{otherwise} \end{cases}$$

and the following lower bounds for $\widetilde{P}_t^{LB}(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})$:

$$\begin{cases} \max\Big(P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}), 1 - \sum_{s' \in \mathcal{S} \setminus \{\widetilde{s}'\}} \widetilde{P}_t^{UB}(s' \mid \widetilde{s}, \widetilde{a})\Big) & \text{if } \widetilde{s}' = s_{t+1} \\ 0 & \text{if CS conditions} \\ \max\Big(0, 1 - \sum_{s' \in \mathcal{S} \setminus \{\widetilde{s}'\}} \widetilde{P}_t^{UB}(s' \mid \widetilde{s}, \widetilde{a})\Big) & \text{otherwise} \end{cases}$$

where the counterfactual stability (CS) conditions are $P(\widetilde{s}' \mid s_t, a_t) > 0$ and $\dfrac{P(s_{t+1} \mid \widetilde{s}, \widetilde{a})}{P(s_{t+1} \mid s_t, a_t)} > \dfrac{P(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})}{P(\widetilde{s}' \mid s_t, a_t)}$.

Proof Sketch. Consider an observed transition $s_t, a_t \to s_{t+1}$ and counterfactual transition $\widetilde{s}, \widetilde{a} \to \widetilde{s}'$. For a fixed θ, we can compute the counterfactual probability $\widetilde{P}_t(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})$ as in Eq. (2). Since $P(s_{t+1} \mid s_t, a_t)$ is fixed, $\widetilde{P}_t(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})$ depends only on:

$$\sum_{u_t=1}^{|U_t|} \mu_{\widetilde{s},\widetilde{a},u_t,\widetilde{s}'} \cdot \mu_{s_t,a_t,u_t,s_{t+1}} \cdot \theta_{u_t} = \sum_{\substack{u_t \in U_t \\ f(s_t,a_t,u_t)=s_{t+1} \\ f(\widetilde{s},\widetilde{a},u_t)=\widetilde{s}'}} \theta_{u_t} \tag{7}$$

Therefore, to minimise/maximise the counterfactual probability of a particular transition, we choose θ to minimise/maximise the value of this sum. The main challenge is that θ must also satisfy the optimisation constraints for *all* state-action pairs in the MDP. To address this, the proof proceeds in two steps:

(1) **Induction over state-action pairs (D.1.2)**. We show that, for any MDP $\mathcal{M}$, if we can identify assignments of θ that satisfy the constraints individually for all $|\mathcal{S}| \times |\mathcal{A}|$ state-action pairs in $\mathcal{M}$, then these assignments can be combined to form a single θ that satisfies the constraints of all $|\mathcal{S}| \times |\mathcal{A}|$ state-action pairs simultaneously.
(2) **Closed-form Probability Bounds (D.1.3)** We then analyse how to minimise/maximise a particular counterfactual transition probability, $\widetilde{P}_t(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})$. When the supports of the observed and counterfactual state-action pairs overlap, the counterfactual stability and counterfactual monotonicity assumptions influence the counterfactual probability bounds, leading to different closed-form expressions. In D.1.3 we systematically cover all possible cases in an MDP and prove, subject to the constraints of the optimisation problem, the minimum and maximum values of Eq. (7) to identify the counterfactual probability bounds.

Equivalence with Existing Work. As discussed in Section 2, Li and Pearl [15] have also derived counterfactual probability bounds for categorical models, given observational and interventional distributions. In this work, we have independently derived equivalent counterfactual probability bounds in the MDP setting, using the canonical SCM framework [38]. However, unlike Li and Pearl [15], we additionally incorporate reasonable assumptions about the causal model to tighten the counterfactual probability bounds and enhance the practical utility of counterfactual inference for MDPs. In Appendix E, we prove our closed-form bounds are equivalent to the bounds of Li and Pearl [15] when the counterfactual stability and monotonicity assumptions are removed (or equivalently, for the case where the state-action pair has *disjoint support* with the observed state-action pair[7]). This equivalence reaffirms the correctness of our bounds and suggests we could reformulate them for an observational distribution (as opposed to a single observed path), which would be an interesting future direction.

Flexibility of Assumptions. While the bounds given in Theorems 4.1-4.3 incorporate the assumptions of counterfactual stability and counterfactual monotonicity, our approach is flexible: these assumptions can be removed if they do not hold in a particular environment (for example, if a domain expert determines they are not applicable). This modularity ensures that our approach can adapt to a variety of settings without requiring major changes to the underlying procedure. Appendix C provides further discussion on the plausibility of our assumptions, as well as closed-form solutions for the case where only the standard counterfactual stability assumption is adopted and for the case where both assumptions are removed.

5 ROBUST COUNTERFACTUAL POLICIES

With our analytical bounds, we can compute a non-stationary interval counterfactual MDP for any MDP $\mathcal{M}$ and observed path τ. Formally, an interval Markov decision process (IMDP) [8] is a tuple $(\mathcal{S}, \mathcal{A}, P_\updownarrow, \mathcal{R})$, extending a standard MDP with an uncertain transition probability function $P_\updownarrow$ that maps each transition to a probability within its bounds $P_\updownarrow(s' \mid s, a) = [P^{LB}(s' \mid s, a), P^{UB}(s' \mid s, a)]$. To construct an interval counterfactual MDP, we compute the counterfactual probability bounds for every transition in the MDP, given each observed transition in τ:

Definition 5.1 (Interval Counterfactual MDP (ICFMDP)). *Given an observed path τ and MDP $\mathcal{M}$, the counterfactual probability for each transition in the interval counterfactual MDP (ICFMDP) $\widetilde{\mathcal{M}}_\tau^\updownarrow$ is defined, for $t = 0, \ldots, T-1$, as $\widetilde{P}_t^\updownarrow(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}) = [\widetilde{P}_t^{LB}(\widetilde{s}' \mid \widetilde{s}, \widetilde{a}), \widetilde{P}_t^{UB}(\widetilde{s}' \mid \widetilde{s}, \widetilde{a})]$.*

To derive optimal counterfactual policies for this ICFMDP, we apply robust value iteration [19], which pessimistically optimises the expected reward over the worst-case CFMDP within the ICFMDP:

$$V^*(s) = \max_\pi \min_{\widetilde{P}_t \in \widetilde{P}_t^\updownarrow} \mathbb{E}_{s' \sim \widetilde{P}_t(\cdot \mid s, \pi(s))} \Big[R(s, \pi(s)) + V^*(s')\Big] \tag{8}$$

This yields a *robust* counterfactual policy: its performance on the true (unknown) causal model is guaranteed to be at least as good as the worst-case performance over the ICFMDP. Moreover, as we

[7] When the state-action pair has disjoint support, counterfactual stability and monotonicity hold vacuously.

prove in Appendix D.2, any CFMDP entailed by the ICFMDP is a valid (i.e., the ICFMDP formulation does not introduce spurious CFMDPs), meaning the value function bounds are tight[8].

6 EVALUATION

We apply our methodology to derive counterfactual policies for various MDPs, addressing four main research questions: (1) how does our CF inference approach perform in off-policy evaluation problems; (2) how does our policy's performance and robustness compare to the Gumbel-max SCM approach; (3) how do the counterfactual stability and monotonicity assumptions impact the probability bounds; and (4) how fast is our approach compared with the Gumbel-max SCM method? We conduct experiments in four environments, spanning simple navigation (GridWorld and Frozen Lake), clinical decision-making (Sepsis), and safety-critical control (Aircraft), with varying levels of stochasticity and complexity (see Appendix G for details). Experiments were run on a 128-core Intel Xeon CPU (see Appendix F for more information), and the code is available at: https://github.com/ddv-lab/robust-cf-inference-in-MDPs.

6.1 Robust CF Inference and Off-Policy Evaluation (OPE)

We assess the validity of our robust CF inference method in an OPE setting [3, 23], where we want to predict the expected return (cumulative reward) of a target policy π^* given only paths observed under a behavioural policy π. In this experiment, π is a suboptimal policy and π^* an improved one. Procedurally, we sample a number of paths (up to 100) under π; for each sampled path, we use our method to derive upper and lower bounds on the counterfactual return of that path under π^*. If our method is unbiased (and enough paths are sampled), then the average of these bounds should contain the true expected return under π^*. Using the same set of paths, we perform a similar evaluation for the Gumbel-max approach, except that this method yields a crisp return value rather than bounds.

Figure 4 illustrates the results of this experiment for the GridWorld ($p = 0.4$) MDP. As expected, the averaged pessimistic and optimistic returns obtained from our approach correctly bound the true return under the target policy. The Gumbel-max approach also closely approximates the target return, indicating that it is also a correct and unbiased approach. However, as we demonstrate in the following sections, the Gumbel-max approach is less robust to causal model uncertainty compared to our method.

6.2 Performance of Counterfactual Policies

To compare policy performance, we measure average rewards of counterfactual paths induced by our policy and the Gumbel-max policy by uniformly sampling 200 counterfactual MDPs from the ICFMDP and generating 10,000 counterfactual paths over each sampled CFMDP. Since the interval CFMDP depends on the observed path, we select three paths of varying optimality: a slightly suboptimal path needing minimal changes to reach the goal, a catastrophic path ending in a terminal low-reward state, and an almost catastrophic path that narrowly avoided a catastrophic state.

[8]For some applications, one might be interested in an *optimistic* variant of the problem. This is supported by our method and boils down to maximising (instead of minimising) the expected reward under the best-case CFMDP within the ICFMDP.

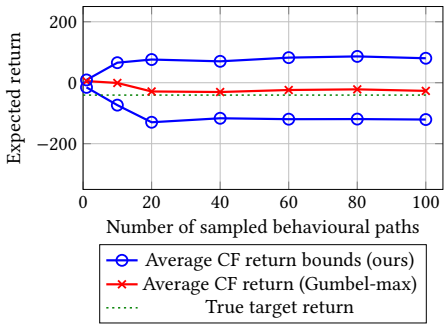


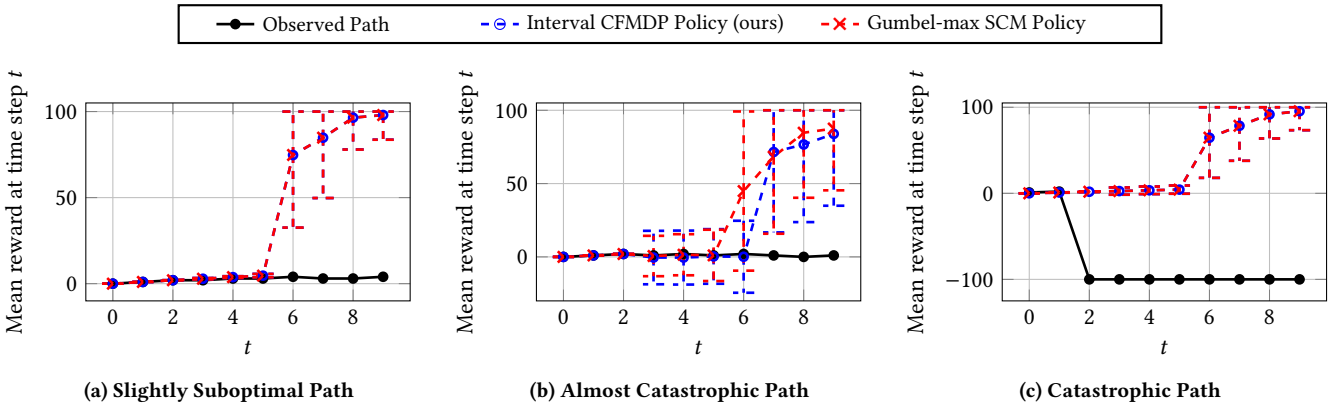
Figure 4: CF inference approaches for off-policy evaluation (GridWorld ($p = 0.4$))

Figures 5-7 show the average counterfactual rewards induced by our policy vs. the Gumbel-max policy in the GridWorld and Sepsis MDPs (similar results are observed in Frozen Lake and Aircraft, see Appendix H). Table 3 reports the worst-case cumulative reward across all induced counterfactual paths, demonstrating that the worst-case cumulative reward obtained by our robust policy is never below that obtained by the Gumbel-max policy.

GridWorld ($p = 0.9$). When $p = 0.9$, the counterfactual probability bounds are typically narrow (see Table 5 for average measurements). Consequently, as shown in Figure 5, both policies are nearly identical and perform similarly well across the slightly suboptimal and catastrophic paths. However, for the almost catastrophic path, our interval CFMDP path is more conservative and follows the observed path more closely (as this is where the probability bounds are narrowest), which typically requires one additional step to reach the goal state than the Gumbel-max SCM policy.

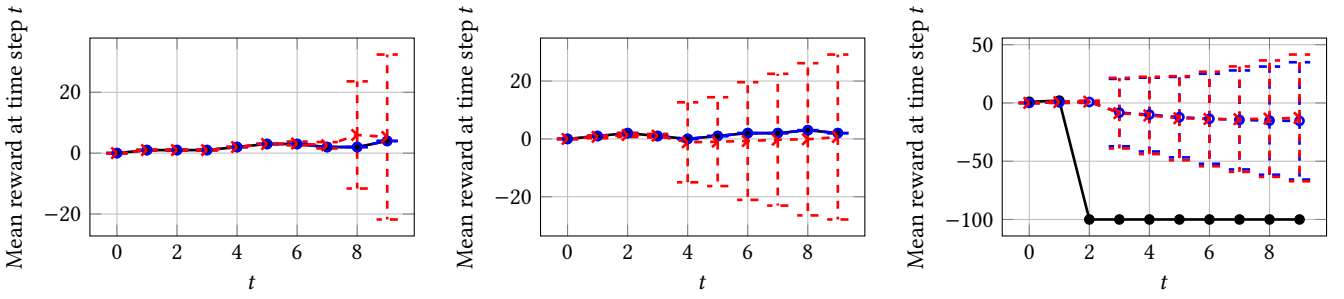
GridWorld ($p = 0.4$). When $p = 0.4$, the GridWorld environment is more uncertain, increasing the risk of entering the dangerous state even if correct actions are chosen. Thus, as shown in Figure 6, the interval CFMDP policy adopts a more conservative approach, and does not deviate from the observed path if it cannot guarantee higher counterfactual rewards (see the slightly suboptimal and almost catastrophic paths). The Gumbel-max policy is more inconsistent: it can yield higher rewards, but also much lower rewards, as reflected by the wide error bars. For the catastrophic path, both policies must significantly deviate from the observed path to achieve a higher reward and perform similarly.

Sepsis. Like in the above experiment, the Sepsis MDP is highly stochastic, with many states equally likely to lead to optimal and poor outcomes. As shown in Figure 7, both policies follow the observed almost-catastrophic path to ensure rewards are no worse than the observation. However, to improve the catastrophic path, both policies must deviate from the observation. Here, on average, the Gumbel-max SCM policy outperforms the interval CFMDP policy, but since both have lower error bars clipped at −1000, neither reliably improves upon the observation. In contrast, for the slightly suboptimal path, the interval CFMDP policy performs significantly better, as shown by the higher lower end of its error bars. Moreover, in these two cases, the worst-case counterfactual path generated by



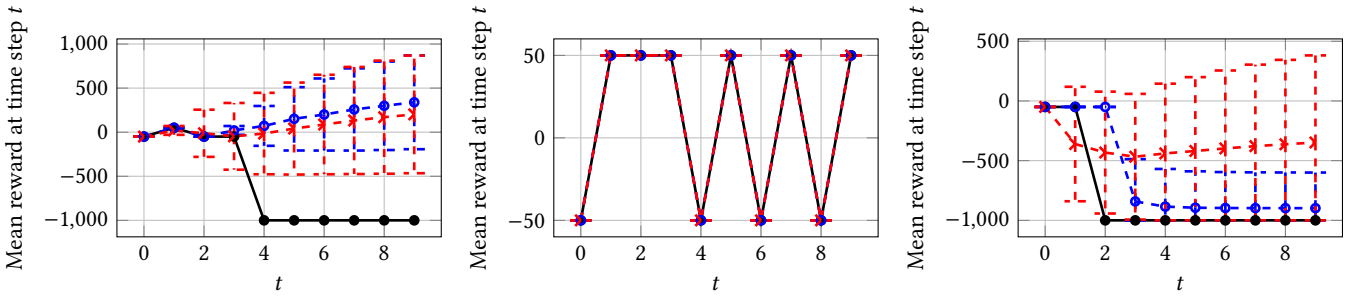
(a) Slightly Suboptimal Path (b) Almost Catastrophic Path (c) Catastrophic Path

Figure 5: Average instant reward of CF paths induced by policies on GridWorld $p = 0.9$. Error bars denote the standard deviation in reward at each time step.



(a) Slightly Suboptimal Path (b) Almost Catastrophic Path (c) Catastrophic Path

Figure 6: Average instant reward of CF paths induced by policies on GridWorld $p = 0.4$.



(a) Slightly Suboptimal Path (b) Almost Catastrophic Path (c) Catastrophic Path

Figure 7: Average instant reward of CF paths induced by policies on Sepsis.

the interval CFMDP policy is better than that of the Gumbel-max SCM policy (see Table 3), indicating its greater robustness.

6.3 Robustness of Counterfactual Policies

The primary advantage of our approach is that it produces policies that are robust with respect to the unknown (true) causal model. To experimentally assess robustness, we examine how our policy and the Gumbel-max policy perform when deployed in the worst-case environment, i.e., when selecting the counterfactual MDP (among the admissible ones) with the lowest reward. Procedurally, given an MDP, we sample from it 100 observed paths following a random policy. For each path, we derive the corresponding Gumbel-max policy and use our approach to construct the corresponding ICFMDP. Recall that the ICFMDP entails all and only the counterfactual MDPs that are compatible with the model and data. Thus, we use the ICFMDP to compute the worst-case performance of the two policies across all CFMDPs within the ICFMDP. Note that, for our policy,

Environment	Slightly Suboptimal Path		Almost Catastrophic		Catastrophic Path	
	ICFMDP (ours)	Gumbel-max	ICFMDP (ours)	Gumbel-max	ICFMDP (ours)	Gumbel-max
GridWorld ($p = 0.9$)	-495	-495	-697	-698	-698	-698
GridWorld ($p = 0.4$)	19	-88	14	-598	-698	-698
Sepsis	-5980	8000	100	100	-7150	-9050
Frozen Lake	41	-97	-68	-87	-87	-87
Aircraft	231	231	-776	-776	-776	-776

Table 3: Lowest cumulative rewards achieved by Interval CFMDP and Gumbel-max SCM policies across sampled counterfactual paths and CFMDPs, for observed paths of varying optimality.

Environment	Pessimistic $V(s_0)$	
	ICFMDP (ours)	Gumbel-max
1. GridWorld ($p = 0.9$)	346 ± 104	304 ± 211
2. GridWorld ($p = 0.4$)	−81.4 ± 197	−230 ± 213
3. Sepsis	1660 ± 1010	−85.4 ± 2860
4. Frozen Lake	37.3 ± 17.9	2.56 ± 51.3
5. Aircraft	−99.0 ± 380	−221 ± 421

Table 4: Average worst-case counterfactual $V(s_0)$ for the ICFMDP and Gumbel-max policies over 100 randomly sampled observed trajectories. Our policy significantly outperforms the Gumbel-max one (Welch T-test resulted in $p < 10^{-4}$ for rows 2,3,4; $p = 0.0381$ for row 1; $p = 0.0163$ for row 5).

the worst-case performance is readily available as the solution of the pessimistic value iteration problem (8). Results are reported in Table 4. Across all environments, the worst-case reward obtained by our approach is consistently much higher than the Gumbel-max approach, demonstrating that our counterfactual inference method is substantially more robust to causal model uncertainty.

6.4 Interval CFMDP Bounds and Runtimes

Table 5 reports average counterfactual probability bound widths (excluding transitions where the upper bound is 0) for each MDP, averaged over 20 observed paths. We compare the bounds under counterfactual stability (CS) and monotonicity (M) assumptions, CS alone, and no assumptions. These results indicate that the assumptions are not overly restrictive (i.e., they do not exclude too many causal models), as they do not significantly narrow the bounds, yet still exclude implausible counterfactuals such as those in the example in Figure 3. In Appendix H we further examine the impact of these assumptions on policy robustness, showing that relaxing the assumptions leads to a slight reduction in performance in most environments, but the policies still outperform the worst-case performance of the Gumbel-max approach. Table 6 compares the average time needed to generate the interval CFMDP vs. the Gumbel-max SCM CFMDP for 20 observations. The GridWorld and Frozen Lake experiments were run single-threaded, while Sepsis and Aircraft were run in parallel. Generating the interval CFMDP is significantly faster as it uses exact analytical bounds, whereas the Gumbel-max CFMDP requires sampling from the Gumbel distribution to estimate counterfactual transition probabilities. In particular, across the five case studies, our approach is 4 to 251 times faster than the Gumbel-max method. Since constructing counterfactual MDPs is the main bottleneck in both approaches, ours is more efficient overall and suitable for larger MDPs.

Environment	Mean Bound Widths		
	CS + M	CS	None
GridWorld ($p = 0.9$)	0.0817	0.0977	0.100
GridWorld ($p = 0.4$)	0.552	0.638	0.646
Sepsis	0.138	0.140	0.140
Frozen Lake	0.307	0.353	0.359
Aircraft	0.180	0.190	0.190

Table 5: Mean width of counterfactual probability bounds in generated CFMDPs.

Environment	ICFMDP (ours)	Gumbel-max
GridWorld ($p = 0.9$)	0.261	56.1
GridWorld ($p = 0.4$)	0.336	54.5
Sepsis	688	2940
Frozen Lake	0.398	100
Aircraft	6.99	74.3

Table 6: Mean execution time (s) for generating CFMDPs.

7 CONCLUSION

We introduced a non-parametric partial counterfactual inference approach for MDPs, leveraging tight analytical bounds on counterfactual probabilities to accelerate the construction of counterfactual models, enabling scaling to larger MDPs. Our interval CFMDP policies are more robust to uncertainty about the true (unknown) causal model, particularly in highly stochastic environments. This robustness yields more reliable counterfactual explanations for improving the agent's policy, which is crucial in safety-critical domains.

Future Work. Like existing work on counterfactual inference in MDPs [3, 23, 30], our approach assumes access to the MDP's transition probabilities. With estimated probabilities, the counterfactual policy may be sensitive to misspecification. A natural extension is to generalise our method to work to uncertain MDPs learned from data, where the probabilities are bounded by confidence intervals learned from observed trajectories. This extension will build directly on the theoretical framework established in this paper. Future work will also explore extending our approach to partially observable and continuous-state settings.

ACKNOWLEDGMENTS

We thank Frederik Mathiesen and Luca Laurenti for their support with the IntervalMDP.jl package. This work was supported by UK Research and Innovation [grant EP/S023356/1] in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org) and by the EPSRC grant EP/W014785/2.

REFERENCES

[1] Alexander Balke and Judea Pearl. 1994. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty in artificial intelligence*. Elsevier, 46–54.

[2] Nina L Corvelo Benz and Manuel Gomez Rodriguez. 2022. Counterfactual inference of second opinions. In *Uncertainty in Artificial Intelligence*. PMLR, 453–463.

[3] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. 2018. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272* (2018).

[4] Zhihong Cai, Manabu Kuroki, Judea Pearl, and Jin Tian. 2008. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics* 64, 3 (2008), 695–701.

[5] Ivi Chatzi, Nina Corvelo Benz, Eleni Straitouri, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. 2024. Counterfactual token generation in large language models. *arXiv preprint arXiv:2409.17027* (2024).

[6] Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. 2023. An automated approach to causal inference in discrete settings. *J. Amer. Statist. Assoc.* (2023), 1–16.

[7] Jasmina Gajcin and Ivana Dusparic. 2024. ACTER: Diverse and Actionable Counterfactual Sequences for Explaining and Diagnosing RL Policies. *arXiv preprint arXiv:2402.06503* (2024).

[8] Robert Givan, Sonia Leach, and Thomas Dean. 2000. Bounded-parameter Markov decision processes. *Artificial Intelligence* 122, 1 (2000), 71–109. https://doi.org/10.1016/S0004-3702(00)00047-3

[9] Joseph Y Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. Part II: Explanations. *The British journal for the philosophy of science* (2005).

[10] Martin B Haugh and Raghav Singal. 2023. Bounding Counterfactuals in Hidden Markov Models and Beyond. *Available at SSRN 4529724* (2023).

[11] Changsung Kang and Jin Tian. 2006. Inequality constraints in causal models with hidden variables. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. 233–240.

[12] Milad Kazemi, Jessica Lally, Ekaterina Tishchenko, Hana Chockler, and Nicola Paoletti. 2025. Counterfactual Influence in Markov Decision Processes. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning (Proceedings of Machine Learning Research, Vol. 275)*, Biwei Huang and Mathias Drton (Eds.). PMLR, 792–817. https://proceedings.mlr.press/v275/kazemi25a.html

[13] Taylor W Killian, Marzyeh Ghassemi, and Shalmali Joshi. 2022. Counterfactually guided policy transfer in clinical settings. In *Conference on Health, Inference, and Learning*. PMLR, 5–31.

[14] Jessica Lally, Milad Kazemi, and Nicola Paoletti. 2025. Robust counterfactual inference in markov decision processes. *arXiv preprint arXiv:2502.13731* (2025).

[15] Ang Li and Judea Pearl. 2024. Probabilities of causation with nonbinary treatment and effect. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 20465–20472.

[16] Guy Lorberbom, Daniel D Johnson, Chris J Maddison, Daniel Tarlow, and Tamir Hazan. 2021. Learning generalized Gumbel-max causal mechanisms. *Advances in Neural Information Processing Systems* 34 (2021), 26792–26803.

[17] Chaochao Lu, Biwei Huang, Ke Wang, José Miguel Hernández-Lobato, Kun Zhang, and Bernhard Schölkopf. 2020. Sample-efficient reinforcement learning via counterfactual-based data augmentation. *arXiv preprint arXiv:2012.09092* (2020).

[18] Charles F Manski. 1990. Nonparametric bounds on treatment effects. *The American Economic Review* 80, 2 (1990), 319–323.

[19] Frederik Baymler Mathiesen, Morteza Lahijanian, and Luca Laurenti. 2024. IntervalMDP.jl: Accelerated Value Iteration for Interval Markov Decision Processes. *IFAC-PapersOnLine* 58, 11, 1–6. https://doi.org/10.1016/j.ifacol.2024.07.416 8th IFAC Conference on Analysis and Design of Hybrid Systems ADHS 2024.

[20] Arash Nasr-Esfahany and Emre Kiciman. 2023. Counterfactual (non-) identifiability of learned structural causal models. *arXiv preprint arXiv:2301.09031* (2023).

[21] Arnab Nilim and Laurent Ghaoui. 2003. Robustness in Markov decision problems with uncertain transition matrices. *Advances in neural information processing systems* 16 (2003).

[22] Kimia Noorbakhsh and Manuel Rodriguez. 2022. Counterfactual temporal point processes. *Advances in Neural Information Processing Systems* 35 (2022), 24810–24823.

[23] Michael Oberst and David Sontag. 2019. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In *ICML*.

[24] Judea Pearl. 2009. *Causality* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

[25] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. 2017. Robust adversarial reinforcement learning. In *International conference on machine learning*. PMLR, 2817–2826.

[26] Edoardo Pona, Milad Kazemi, Yali Du, David Watson, and Nicola Paoletti. 2025. Abstract Counterfactuals for Language Model Agents. *arXiv preprint arXiv:2506.02946* (2025).

[27] James M Robins. 1989. 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* (1989), 113–159.

[28] Yuewen Sun, Erli Wang, Biwei Huang, Chaochao Lu, Lu Feng, Changyin Sun, and Kun Zhang. 2024. ACAMDA: improving data efficiency in reinforcement learning through guided counterfactual data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 15193–15201.

[29] Jin Tian and Judea Pearl. 2002. A general identification condition for causal effects. In *AAAI/IAAI*. 567–573.

[30] Stratis Tsirtsis, Abir De, and Manuel Rodriguez. 2021. Counterfactual explanations in sequential decision making under uncertainty. *Advances in Neural Information Processing Systems* 34 (2021), 30127–30139.

[31] Stratis Tsirtsis and Manuel Rodriguez. 2024. Finding counterfactually optimal action sequences in continuous state spaces. *Advances in Neural Information Processing Systems* 36 (2024).

[32] Athanasios Vlontzos, Bernhard Kainz, and Ciarán M Gilligan-Lee. 2023. Estimating categorical counterfactuals via deep twin networks. *Nature Machine Intelligence* 5, 2 (2023), 159–168.

[33] Marco Zaffalon, Alessandro Antonucci, and Rafael Cabañas. 2020. Structural causal models are (solvable by) credal networks. In *International Conference on Probabilistic Graphical Models*. PMLR, 581–592.

[34] Marco Zaffalon, Alessandro Antonucci, and Rafael Cabañas. 2021. Causal Expectation-Maximisation. In *WHY-21 Workshop*.

[35] Marco Zaffalon, Alessandro Antonucci, Rafael Cabañas, and David Huber. 2023. Approximating counterfactual bounds while fusing observational, biased and randomised data sources. *International Journal of Approximate Reasoning* 162 (2023), 109023.

[36] Marco Zaffalon, Alessandro Antonucci, Rafael Cabañas, David Huber, and Dario Azzimonti. 2022. Bounding counterfactuals under selection bias. In *International Conference on Probabilistic Graphical Models*. PMLR, 289–300.

[37] Marco Zaffalon, Alessandro Antonucci, Rafael Cabañas, David Huber, and Dario Azzimonti. 2024. Efficient computation of counterfactual bounds. *International Journal of Approximate Reasoning* (2024), 109111.

[38] Junzhe Zhang, Jin Tian, and Elias Bareinboim. 2022. Partial Counterfactual Identification from Observational and Experimental Data. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 26548–26558.

[39] Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. 2020. Counterfactual off-policy training for neural dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3438–3448.