
Sample-Efficient Learning of Probabilistic Causes for Reachability in Markov Decision Processes with Probabilistic Guarantees

Ryohei Oura^{1,2}

Georgios Fainekos¹

Hideki Okamoto¹

Bardh Hoxha¹

¹Toyota Research Institute of North America, Toyota Motor North America

²Frontier Research Center, Toyota Motor Corporation

Abstract

Probabilistic model checking for Markov decision processes (MDPs) provides quantitative guarantees, but often offers limited insight into why undesired outcomes occur. Probability-raising (PR) causality addresses this by identifying states whose visitation increases the probability of reaching designated states. Existing PR-cause identification methods, however, use MDP modifications not well-suited for learning: the gap between conditional and unconditional reachability probabilities can be hard to detect from transition samples, and construction requires reachability probabilities of the MDP, which are unavailable when transition probabilities are unknown. We study unknown MDPs and propose a learning approach with probabilistic guarantees for PR-cause identification. Our key ingredient is a *restart-based MDP modification* that reduces PR-cause checking to two conditional reachability queries without using reachability values of the original MDP. We prove correctness, establish sample-complexity bounds, and develop an anytime learning-and-checking algorithm based on two-sided value iteration that progressively classifies states as causal, non-causal, or undecided. Experiments on two benchmarks demonstrate reliable and fast identification of PR causes.

1 INTRODUCTION

Probabilistic model checking (PMC) [Baier and Katoen, 2008] is widely used to analyze Markov decision processes (MDPs) that capture probabilistic and nondeterministic behavior, such as that arising from permissive policies, randomized decisions, and human-related uncertainties in various applications [Drager et al., 2015, Junges et al., 2018, Chen et al., 2022]. While PMC can quantify the probabili-

ties of undesired outcomes, it typically does not explain why such a behavior happens [Baier et al., 2021a]. In practice, one often wants to know not just the probability but which states are responsible for it: for instance, states whose visitation raises the violation probability no matter which action a permissive controller selects, or states whose traversal raises the satisfaction probability no matter how a human or other agent acts. Identifying such states helps localize the source of undesired/desired behavior. Cause-effect reasoning aims to close this gap by connecting outcomes for given specifications to concrete system elements such as states or actions [Baier et al., 2024, Oura and Ito, 2025, Kazemi et al., 2025, Coenen et al., 2022, Triantafyllou et al., 2022] (see Section 5.1 for details).

Probability-raising (PR) causality provides a verification-friendly explanation for probabilistic reachability in MDPs [Baier et al., 2022, 2024]. In this framework, a cause is a set of states whose visitation increases the probability of reaching designated states, such as rejecting states for logical specifications. The authors have also introduced algorithms based on model checking to compute PR causes under some optimality criteria, but existing approaches have presumed that the MDP is fully known. We focus on reachability since it is a fundamental property to which linear temporal logic specifications are reduced via a standard product construction [Baier and Katoen, 2008].

Learning-based verification provides probably approximately correct (PAC) guarantees for unknown stochastic systems [Ashok et al., 2019, Agarwal et al., 2025, Fu and Topcu, 2014, Perez et al., 2024] (see Section 5.3 for details), but it does not directly yield cause identifications. In particular, they are not designed to decide which states (or state sets) are responsible for undesired behavior. Our work builds on a two-sided value iteration scheme [Ashok et al., 2019], but targets certified cause identification rather than property verification alone.

Bringing PR-cause computation to unknown MDPs raises challenges not typically encountered when the model is fully

known. The causation decision depends on state-specific probability gaps that are unknown in advance, making a single fixed-accuracy learning phase potentially insufficient or wasteful. Moreover, standard PR-cause pipelines rely on MDP transformations [Baier et al., 2022] that are ill-suited for learning. They compare a conditional reachability probability with an unconditional one in a way that can be statistically unfavorable. The unconditional quantity averages over runs that do and do not visit the candidate state, which can dilute the gap, especially because the policy optimization used for PR-cause checking tends to steer toward truly causal candidates. Finally, these transformations for PR-cause identification, as well as existing MDP transformations for computing conditional probabilities more generally, require reachability probabilities of the original MDP during construction [Baier et al., 2014, 2022], which are unavailable in unknown MDPs (see Section 5.2 for details).

As a closely related and complementary work, Oura and Ito [2025] has studied PAC-guaranteed identification of PR causes using the existing MDP transformation for an uncertain MDP by instantiating transition probabilities. However, the authors do not consider finding a true PR cause in an unknown MDP.

To address these challenges, we propose a learning approach for PR-cause identification with probabilistic guarantees. First, we introduce a restart-based MDP modification that reduces PR-cause checking to comparing reachability probabilities conditioned on visiting a candidate state and on bypassing it, and we prove its correctness (Proposition 1). Second, we analyze and compare the sample-complexity bounds for learning PR causes with high confidence in unknown MDPs under the proposed modification and the existing one (Theorem 1). Third, built upon two-sided value iteration, we present an anytime learning algorithm to progressively classify states into three classes (causal/non-causal/undecided). We prove its correctness (Theorems 2 and 3). Finally, we evaluate the proposed method on two benchmarks against two baselines: one combining [Baier et al., 2022] with [Ashok et al., 2019], and the other combining learning of a PAC-guaranteed interval MDP [Suilen et al., 2022] with [Ourra and Ito, 2025].

2 PRELIMINARIES

We introduce MDPs and probabilistic causation in an MDP as a set of states, aiming to capture and explain cause-effect relationships for reachability properties on the MDP.

2.1 MARKOV DECISION PROCESSES AND PROBABILITY-RAISING CAUSALITY

A Markov decision process (MDP) is a tuple $\mathcal{M} = (S, A, s_I, P)$, where S is a finite set of states, A is a fi-

nite set of actions, $s_I \in S$ is an initial state, and $P : S \times A \times S \rightarrow [0, 1]$ is a transition probability function with $\sum_{s' \in S} P(s, a, s') \in \{0, 1\}$ for any $s \in S$ and any $a \in A$.

For any $s \in S$, we denote the set of actions enabled at s as $\mathcal{A}(s) = \{a \in A \mid \exists s' \text{ s.t. } P(s, a, s') > 0\}$. We say that $s \in S$ is terminal if $\mathcal{A}(s) = \emptyset$. We denote the set of terminal states by E , assuming $E \neq \emptyset$ and $s_I \notin E$. A path is a finite or infinite sequence $s_0 a_0 s_1 \dots$, where $P(s_i, a_i, s_{i+1}) > 0$ for any $s_i, s_{i+1} \in S$ and $a_i \in A$ with $i = 0, 1, \dots$. Intuitively, a path that reaches E is considered undesirable. A policy $\pi : \text{FinPath} \times A \rightarrow [0, 1]$ is a function that assigns a probability to each action for a finite path, where FinPath is the set of all finite paths. A policy π is memoryless if $\pi(\rho, a) = \pi(\text{last}(\rho), a)$ for any $\rho \in \text{FinPath}$ and any $a \in A$, where $\text{last}(\rho)$ denotes the last state of ρ . For any MDP $\mathcal{M}$, any state $s \in S$, and any policy π , a unique probability measure $\Pr_{\mathcal{M},s}^\pi$ on measurable sets of paths starting from s under π in a standard way [Baier and Katoen, 2008]. For brevity, we write $\Pr_{\mathcal{M}}^\pi$ to denote $\Pr_{\mathcal{M},s_I}^\pi$. We also denote the minimal and maximal probabilities over all policies by $\Pr_{\mathcal{M},s}^{\min}$ and $\Pr_{\mathcal{M},s}^{\max}$, respectively. We denote by $\text{Post}_{s,a}$ the set of successor states of s by executing a . We refer to an MDP as unknown if the successors $\text{Post}_{s,a}$ are known for any $s \in S$ and $a \in A$, but its transition probability function is unknown.

For convenience, we use linear temporal logic (LTL)-like notations such as $\neg$ (negation), $\wedge$ (and), $\diamond$ (eventually), and U (until) to represent certain path properties. For any $X_0, X_1 \subseteq S$, the formula $X_0 U X_1$ is satisfied by a (state) path $\rho = s_0 s_1 \dots$ such that there exists $j \geq 0$ such that, for any $i < j$, $s_i \in X_0$ and $s_j \in X_1$ hold, and we denote $\diamond X_1 = S U X_1$. Furthermore, $\neg \diamond X_1$ is satisfied by a path ρ such that all states in ρ are not contained in X_1 . Throughout, when a set argument is a singleton, we omit the braces. For example, we write $\diamond c$ for $\diamond\{c\}$ and XUc for $XU\{c\}$.

Definition 1. For any MDP $\mathcal{M}$, the set of terminal states $E \subseteq S$, and a nonempty subset $C \subseteq S \setminus E$, we say that C is a strict probability-raising (SPR) cause for E on $\mathcal{M}$ if and only if the following conditions (C1) and (C2) hold:

$$(C1) \quad \forall c \in C, \forall \pi \text{ s.t. } \Pr_{\mathcal{M}}^\pi(\neg CUc) \in (0, 1),$$

$$\Pr_{\mathcal{M}}^\pi(\diamond E \mid \neg CUc) > \Pr_{\mathcal{M}}^\pi(\diamond E), \quad (1)$$

$$(C2) \quad \forall c \in C, \exists \pi \text{ s.t. } \Pr_{\mathcal{M}}^\pi(\neg CUc) \in (0, 1).$$

Intuitively, $\diamond E$ denotes eventually reaching E , while $\neg CUc$ means reaching c before any other state in C . An SPR cause is a set C of waypoint states that increases the probability of reaching E , as required by (C1). We exclude policies under which $\Pr_{\mathcal{M}}^\pi(\neg CUc) = 0$ (the conditional probability is undefined) or 1 (the conditional equals the unconditional). (C2) ensures that (C1) is not satisfied vacuously: for each $c \in C$, there exists at least one policy under which this

comparison in (C1) is non-trivial. For each $c \in S$, we call c a causal state if $\{c\}$ is an SPR cause, and a non-causal state otherwise.

Example 1. We consider an MDP depicted in Fig. 1. The sets of actions and terminal states are $A = \{a, b\}$ and $E = \{s_5\}$, respectively. The only enabled action is a except for s_0 . $C = \{s_2, s_3\}$ forms an SPR cause in this MDP for the following reasons. First, there exist two paths $s_0 a s_2$ and $s_0 a s_1 a s_3$. Thus, C satisfies the condition (C2). Second, we have $\Pr(\Diamond s_5 \mid \Diamond s_2) = 0.6$ and $\Pr(\Diamond s_5 \mid \Diamond s_3) = 1$ while $\Pr_{\mathcal{M}}^{\max}(\Diamond s_5) = 0.48$. Hence, C also satisfies (C1).

Concrete application scenarios for Def. 1, including permissive robotic planning with unsafe states and linear temporal logic planning against an adversary, are discussed in Appendix F.

We consider lower and upper bounds for reachability probabilities. For any MDP $\mathcal{M}$, we say that $\underline{x}^{\text{opt}}, \bar{x}^{\text{opt}} : S \rightarrow [0, 1]$ are lower and upper bounds for the reachability probabilities for the set of terminal states E if and only if

$$\forall s \in S, \underline{x}^{\text{opt}}(s) \leq \Pr_{\mathcal{M},s}^{\text{opt}}(\Diamond E) \leq \bar{x}^{\text{opt}}(s), \quad (2)$$

where $\text{opt} \in \{\min, \max\}$. Moreover, for any multi-sample of transitions $\mathcal{D}$ from $\mathcal{M}$ and any $\delta \in (0, 1)$, we consider a lower bound $\hat{P}_\delta : S \times A \times S \rightarrow [0, 1]$ for the transition probability P as, for any $(s, a, s') \in S \times A \times S$,

$$\hat{P}_\delta(s, a, s') = \begin{cases} \max(0, \frac{N(s, a, s')}{N(s, a)} - \sqrt{\frac{-\ln \delta}{2N(s, a)}}) & \text{if } N(s, a) > 0 \wedge |\text{Post}_{s,a}| > 1, \\ 1 & \text{if } \text{Post}_{s,a} = \{s'\}, \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where $N(s, a, s')$ denotes the number of times the transition (s, a, s') appears in $\mathcal{D}$ and $N(s, a) = \sum_{s' \in S} N(s, a, s')$. We denote $\hat{P}_\delta \leq P$ if and only if $\hat{P}_\delta(s, a, s') \leq P(s, a, s')$ for any $(s, a, s') \in S \times A \times S$. Intuitively, $\hat{P}_\delta$ is derived based on Hoeffding's inequality [Ashok et al., 2019].

We consider $S_{\mathcal{M}}^{\min} = \{s \in S \mid \Pr_{\mathcal{M},s}^{\min}(\Diamond E) = 0\}$ and $S_{\mathcal{M}}^{\max} = \{s \in S \mid \Pr_{\mathcal{M},s}^{\max}(\Diamond E) = 0\}$ for any MDP $\mathcal{M}$. Note that both $S_{\mathcal{M}}^{\min}$ and $S_{\mathcal{M}}^{\max}$ can be computed in polynomial time in the number of states and actions [Baier and Katoen, 2008]. Based on the lower-bound estimate $\hat{P}_\delta$ of the transition probability function, we consider *modified* Bellman operators that compute lower/upper bounds of the reachability probability as follows [Ashok et al., 2019].

Definition 2. For any MDP $\mathcal{M}$, any multi-sample of transitions $\mathcal{D}$, any $\delta \in (0, 1)$, we define operators $f_{\mathcal{M},\delta}^{\text{opt}^{(1)},\text{opt}^{(2)}} : [0, 1]^S \rightarrow [0, 1]^S$ such that $f_{\mathcal{M},\delta}^{\text{opt}^{(1)},\text{opt}^{(2)}}[x](s) = 1$ for any $s \in E$, $f_{\mathcal{M},\delta}^{\text{opt}^{(1)},\text{opt}^{(2)}}[x](s) = 0$ for any $s \in S_{\mathcal{M}}^{\text{opt}^{(1)}}$, and, for

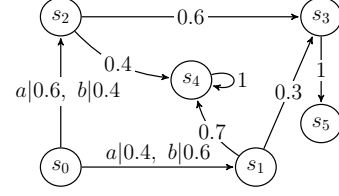


Figure 1: MDP in Example 1 with $E = \{s_5\}$. The action enabled at states except for s_0 is only a , and hence we abbreviate it.

any $s \in S \setminus (E \cup S_{\mathcal{M}}^{\text{opt}^{(1)}})$,

$$f_{\mathcal{M},\delta}^{\text{opt}^{(1)},\text{opt}^{(2)}}[x](s) := \text{opt}_{a \in A(s)}^{(1)} \sum_{s' \in \text{Post}_{s,a}} \hat{P}_\delta(s, a, s') x(s') + \text{opt}_{s' \in \text{Post}_{s,a}}^{(2)} x(s') (1 - \sum_{s' \in \text{Post}_{s,a}} \hat{P}_\delta(s, a, s')), \quad (4)$$

where $\hat{P}_\delta$ is defined by (3) for $\mathcal{M}$ and $\text{opt}^{(1)}, \text{opt}^{(2)} \in \{\min, \max\}$.

Intuitively, repeatedly applying $f_{\mathcal{M},\delta}^{\text{opt},\min}$ and $f_{\mathcal{M},\delta}^{\text{opt},\max}$ to a function $x : S \rightarrow [0, 1]$ provide lower and upper bounds for $\Pr_{\mathcal{M},\cdot}^{\text{opt}}(\Diamond E)$.

2.2 PROBLEM SETTING

Our goal is to identify an SPR cause that maximally covers undesired paths with high probability on a given unknown MDP. To formalize this intuition, we pose the following problem by employing the same path coverage objective as in prior works [Baier et al., 2022, Oura and Ito, 2025].

Problem 1. Given an unknown MDP $\mathcal{M}$, for any $\delta_{\mathcal{M}} > 0$, design an algorithm that finds a subset of states $C^* \subset S$ such that, w. p. at least $1 - \delta_{\mathcal{M}}$, C^* is an SPR cause for E on $\mathcal{M}$ and, for any SPR cause C for E on $\mathcal{M}$,

$$\min_{\pi} \Pr_{\mathcal{M}}^{\pi}(\Diamond C^* \mid \Diamond E) - \Pr_{\mathcal{M}}^{\pi}(\Diamond C \mid \Diamond E) \geq 0. \quad (5)$$

Intuitively, we aim to find an SPR cause whose coverage for undesired path, that is $\Pr_{\mathcal{M}}^{\pi}(\Diamond C^* \mid \Diamond E)$, is maximal under all possible policies among all SPR causes. In the following, we do not mention the set of terminal states E for any MDP when it is clear from the context.

3 CAUSE LEARNING METHODOLOGY

In this section, we propose an anytime learning method for identifying PR causes in unknown MDPs. The method builds on a proposed MDP modification tailored to learning. For each candidate causal state, we redirect its outgoing transitions to the initial state, reducing PR-cause checking to

two conditional reachability queries. This modification does not require reachability probabilities of the original MDP and makes the relevant probability gaps easier to detect from samples. Building on it, we develop a learning-and-checking loop that repeatedly tightens lower and upper bounds on reachability probabilities for both the original and modified MDPs. As these bounds tighten, more candidate states are classified into three classes (causal/non-causal/undecided), leaving fewer states unclassified, yielding an anytime procedure with confidence guarantees.

3.1 MDP MODIFICATIONS FOR SPR CAUSE LEARNING IN UNKNOWN MDPs

We first review an existing MDP modification introduced in [Baier et al., 2022, 2024] for computing SPR causes and point out that it can be sample-inefficient when combined with MDP learning. We then propose a restart-based MDP modification that enables more sample-efficient learning of SPR causes while preserving correctness.

The existing transformation rewrites the transitions outgoing from a specified state c so that the transition probability to a state in E equals $\Pr_{\mathcal{M},c}^{\min}(\Diamond E)$, and the transition probability to an added sink state equals $1 - \Pr_{\mathcal{M},c}^{\min}(\Diamond E)$.

Definition 3 (Minimal-Reachability-based Transformation for MDPs [Baier et al., 2022, 2024]). *For any MDP $\mathcal{M}$ and any $c \in S$, we define a transformed MDP as a tuple $\mathcal{M}_{\text{mr}}^{[c]} = (S \cup \{s_{\text{noeff}}\}, A, s_I, P_{\text{mr}}^c)$ such that $P_{\text{mr}}^c(s_{\text{noeff}}, a, s_{\text{noeff}}) = 1$ for any $a \in A$, $P_{\text{mr}}^c(s, a, s') = P(s, a, s')$ for any $(s, a, s') \in S \setminus \{c\} \times A \times S$, and $P_{\text{mr}}^c(c, a, s_{\text{eff}}) = \Pr_{\mathcal{M},c}^{\min}(\Diamond E)$ and $P_{\text{mr}}^c(c, a, s_{\text{noeff}}) = 1 - \Pr_{\mathcal{M},c}^{\min}(\Diamond E)$ for any $a \in A(c)$, where s_{eff} is a fixed state in E .*

Def. 3 allows us to check (C1) in Def. 1 for each $c \in S$ via $\Pr_{\mathcal{M},c}^{\min}(\Diamond E) - \Pr_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max}(\Diamond E)$. To see this, recall that in $\mathcal{M}_{\text{mr}}^{[c]}$ the outgoing transitions of c are rewritten so that c leads to E exactly with probability $\Pr_{\mathcal{M},c}^{\min}(\Diamond E)$ and otherwise to the sink s_{noeff} , which never reaches E . A policy attaining $\Pr_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max}(\Diamond E)$ therefore reaches E as much as possible before reaching c and, after reaching c , steers to E with the minimal probability. Comparing this maximal reachability against $\Pr_{\mathcal{M},c}^{\min}(\Diamond E)$, the reachability after a visit to c , realizes the test of (C1). However, when c is causal, a policy attaining $\Pr_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max}(\Diamond E)$ typically steers toward c with high probability, making the reachability probability gap between the original and transformed MDPs small and requiring many transition samples to determine whether c is causal. Moreover, Def. 3 requires the (unknown) maximal reachability probabilities at construction time; thus, we must, e.g., replace transition probabilities from c in $\mathcal{M}_{\text{mr}}^{[c]}$ with estimates, which in turn needs additional samples to guarantee correctness. To address these issues, we introduce a restart-based modification that removes the effect of the

probability of reaching c from the initial state and avoids using estimated reachability probabilities during construction.

Definition 4 (Restart-based Modification for MDPs). *For any MDP $\mathcal{M}$ and any $c \in S$, a modified MDP is a tuple $\mathcal{M}^{[c]} = (S, A, s_I, P^c)$ such that $P^c(s, a, s') = P(s, a, s')$ for any $(s, a, s') \in S \setminus \{c\} \times A \times S$, and $P^c(c, a, s_I) = 1$ for any $a \in A(c)$.*

Intuitively, the transition from c is reset to the initial state. Hence, this enables us to evaluate the reachability probability conditioned on bypass of c in $\mathcal{M}$ by the unconditional reachability probability on $\mathcal{M}^{[c]}$ as indicated in Lemma 1.

Lemma 1. *For any MDP $\mathcal{M}$, any $c \in S \setminus E$, and any policy π with $\Pr_{\mathcal{M}}^{\pi}(\neg \Diamond c) > 0$, there exists a policy $\tilde{\pi}$ on $\mathcal{M}^{[c]}$ such that we have*

$$\Pr_{\mathcal{M}^{[c]}}^{\tilde{\pi}}(\Diamond E) = \Pr_{\mathcal{M}}^{\pi}(\Diamond E \mid \neg \Diamond c). \quad (6)$$

Moreover, if π is memoryless, then we can choose $\tilde{\pi} = \pi$.

The identity holds because Def. 4 restarts every path reaching c back to s_I , redistributing the contribution of c -visiting paths to fresh paths that never visit c . Proof of Lemma 1 is in Appendix A. Moreover, we explicitly avoid relying on any reachability probabilities on $\mathcal{M}$ to construct $\mathcal{M}^{[c]}$.

Proposition 1 ensures that states are classified correctly as causal/non-causal even when using Def. 4 in place of Def. 3. Noting that $\Diamond c$ equals $\neg c U c$, to capture the policies for which the probability comparison in (C1) of Def. 1 is non-trivial, for any $c \in S$, we define

$$\Pi_{\mathcal{M},c} = \{\pi \mid \Pr_{\mathcal{M}}^{\pi}(\Diamond c) \in (0, 1)\}. \quad (7)$$

Proposition 1 (Correctness of Restart-based Modification). *For any MDP $\mathcal{M}$ and any $c \in S$, if $\Pi_{\mathcal{M},c} = \emptyset$, then c is not causal, otherwise we have the following properties:*

1. *If $\Pr_{\mathcal{M},c}^{\min}(\Diamond E) > \Pr_{\mathcal{M}^{[c]}}^{\max}(\Diamond E)$, then c is causal.*
2. *If $\Pr_{\mathcal{M},c}^{\min}(\Diamond E) < \Pr_{\mathcal{M}^{[c]}}^{\max}(\Diamond E)$, then c is not causal.*
3. *If $\Pr_{\mathcal{M},c}^{\min}(\Diamond E) = \Pr_{\mathcal{M}^{[c]}}^{\max}(\Diamond E)$, then:*
 - (a) *if c is not reachable from s_I in $\mathcal{M}_{\text{max}}^{[c]}$, c is causal,*
 - (b) *otherwise, c is not causal,*

where $\mathcal{M}^{[c]}$ is defined by Def. 4 and $\mathcal{M}_{\text{max}}^{[c]}$ is the sub-MDP of $\mathcal{M}^{[c]}$ induced by $\mathcal{A}^{\text{max}}(s) = \{a \in A \mid \Pr_{\mathcal{M}^{[c]},s}^{\max}(\Diamond E) = \sum_{s' \in S} P^c(s, a, s') \Pr_{\mathcal{M}^{[c]},s'}^{\max}(\Diamond E)\}$.

At a high level, the cases contrast best-policy reachability after visiting c with worst-policy reachability when bypassing c to decide (C1). In the equality case with c reachable, some policy makes the two conditional probabilities coincide and thus violates (C1); otherwise (C1) holds. Proof for Proposition 1 is in Appendix A

Remark 1. We mention that the soundness of the restart-based modification remains even if we instead use an existential quantifier for policies in (C1) of Def. 1. Please see Proposition 5 in Appendix A.

For any $\delta \in (0, 1)$, we introduce a lower bound $\hat{P}^c : S \times A \times S \rightarrow [0, 1]$ for P^c of $\mathcal{M}^{[c]}$ such that, for any $(s, a, s') \in S \times A \times S$,

$$\hat{P}_\delta^c(s, a, s') = \begin{cases} \hat{P}_\delta(s, a, s') & \text{if } s \neq c, \\ 1 & \text{if } s = c, s' = s_I, \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

where $\hat{P}_\delta$ is defined by (3) for $\mathcal{M}$ with $\mathcal{D}$.

We will analyze the number of transition samples required to correctly classify a state as causal or non-causal under both MDP constructions. Our modification can substantially reduce the required number of samples compared with a straightforward adaptation of Def. 3 to PAC model checking via Def. 2. We first introduce the mixing time of an MDP.

Definition 5. For any MDP $\mathcal{M}$, any policy π , any $s \in S$, and any $\epsilon > 0$, ϵ -value mixing time from s is $T_{\mathcal{M}, \epsilon, s}^\pi = \min\{T \mid |\Pr_{\mathcal{M}, s}^\pi(\diamond^{\leq T} E) - \Pr_{\mathcal{M}, s}^\pi(\diamond E)| \leq \epsilon\}$, where $\diamond^{\leq T} E$ denotes the event of reaching E within T steps.

Intuitively, $T_{\mathcal{M}, \epsilon, s}^\pi$ denotes the minimal time T such that, under π , the difference between the probability of eventually reaching E and the probability of reaching E within T steps is less than ϵ . For simplicity, let $T_{\mathcal{M}, \epsilon, s}^{\text{opt}} = T_{\mathcal{M}, \epsilon, s}^\pi$ for any π with $\Pr_{\mathcal{M}}^\pi(\diamond E) = \Pr_{\mathcal{M}}^{\text{opt}}(\diamond E)$, where $\text{opt} \in \{\min, \max\}$.

Let $\underline{x}_{\mathcal{M}}^{\text{opt},*}$ and $\bar{x}_{\mathcal{M}}^{\text{opt},*}$ be fixed points of $f_{\mathcal{M}}^{\text{opt}, \min}$ and $f_{\mathcal{M}}^{\text{opt}, \max}$, respectively. We denote the maximum number of successors by $d_{\max} = \max_{(s,a) \in (S \setminus E) \times A} |\text{Post}_{s,a}|$ and the set of probabilistic transitions by $\text{Tr}(\mathcal{M}) = \{(s, a, s') \in S \times A \times S \mid P(s, a, s') \in (0, 1)\}$, respectively. Similar to (8), for any $\delta \in (0, 1)$, we define a plug-in lower-bound estimate $\hat{P}_{\text{mr}, \delta}^c : S \times A \times S \rightarrow [0, 1]$ of P_{mr}^c in $\mathcal{M}_{\text{mr}}^{[c]}$ by setting $\hat{P}_{\text{mr}, \delta}^c(s, a, s') = \hat{P}_\delta(s, a, s')$ for any $s \neq c$, and $\hat{P}_{\text{mr}, \delta}^c(c, a, s_{\text{eff}}) = \underline{x}_{\mathcal{M}}^{\min,*}(c)$ and $\hat{P}_{\text{mr}, \delta}^c(c, a, s_{\text{noeff}}) = 1 - \bar{x}_{\mathcal{M}}^{\min,*}(c)$ for any $a \in \mathcal{A}(c)$.

Theorem 1 (Sample Complexity). For any MDP $\mathcal{M}$, any random N transition samples $\mathcal{D}$, any $\delta_{\mathcal{M}}$, and any $c \in S$, if $\Delta_c := \Pr_{\mathcal{M}, c}^{\min}(\diamond E) - \Pr_{\mathcal{M}^{[c]}}^{\max}(\diamond E) \neq 0$ and, for any $(s, a) \in (S \setminus E) \times A$,

$$N(s, a) \geq K \frac{d_{\max}^2 T^2}{\Delta_c^2} \ln \frac{2|\text{Tr}(\mathcal{M})|}{\delta_{\mathcal{M}}}, \quad (9)$$

then $\underline{x}_{\mathcal{M}}^{\min,*}(c) > \bar{x}_{\mathcal{M}^{[c]}}^{\max,*}(s_I)$ or $\bar{x}_{\mathcal{M}}^{\min,*}(c) < \underline{x}_{\mathcal{M}^{[c]}}^{\max,*}(s_I)$ w. p. at least $1 - \delta_{\mathcal{M}}$, moreover, if $\Delta_{\text{mr}, c} := \Pr_{\mathcal{M}, c}^{\min}(\diamond E) - \Pr_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max}(\diamond E) \neq 0$ and, for any $(s, a) \in (S \setminus E) \times A$,

$$N(s, a) \geq K \frac{d_{\max}^4 T_{\text{mr}}^4}{\Delta_{\text{mr}, c}^2} \ln \frac{2|\text{Tr}(\mathcal{M})|}{\delta_{\mathcal{M}}}, \quad (10)$$

then $\underline{x}_{\mathcal{M}}^{\min,*}(c) > \bar{x}_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max,*}(s_I)$ or $\bar{x}_{\mathcal{M}}^{\min,*}(c) < \underline{x}_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max,*}(s_I)$ w. p. at least $1 - \delta_{\mathcal{M}}$, where $T = \max\{T_{\mathcal{M}, \frac{|\Delta_c|}{K_1}, c}^{\min}, T_{\mathcal{M}^{[c]}, \frac{|\Delta_c|}{K_1}, s_I}^{\max}\}$, $T_{\text{mr}} = \min\{T \mid T \geq \max\{T_{\mathcal{M}, \epsilon, c}^{\min}, T_{\mathcal{M}_{\text{mr}}^{[c]}, \epsilon, s_I}^{\max}\}\}$ with $\epsilon = |\Delta_{\text{mr}, c}|/(K_1(Td_{\max} + 1))$, and $K, K_1 > 0$ are constants.

Proof for Theorem 1 is in Appendix A. Combined with Proposition 1, the respective lower bound in (9) and (10) acts as a number of samples required for correct classification of states as causal or non-causal.

Remark 2 (Comparison of Sample Complexities). The restart-based transformation $\mathcal{M}^{[c]}$ can yield more favorable sample-complexity bound than the existing modification $\mathcal{M}_{\text{mr}}^{[c]}$. First, $\mathcal{M}_{\text{mr}}^{[c]}$ embeds the unknown reachability probability $\Pr_{\mathcal{M}, c}^{\min}(\diamond E)$ into the transition probabilities from c . So, estimation errors enter twice through learning the original transitions and through plugging an estimated reachability probability back into the transformed MDP. This explains the stronger $d_{\max}^4 T_{\text{mr}}^4$ scaling, compared with $d_{\max}^2 T^2$ for $\mathcal{M}^{[c]}$. Second, this plug-in construction imposes a stricter finite-horizon accuracy requirement, so that T_{mr} is evaluated at a smaller tolerance $((T_{\text{mr}}d_{\max} + 1)\epsilon = |\Delta_{\text{mr}, c}|/K_1$ instead of $\epsilon = |\Delta_c|/K_1$), which can further increase T_{mr} . Finally, when c is causal, a policy that achieves $\Pr_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max}(\diamond E)$ tends to steer towards c , which can severely reduce $\Delta_{\text{mr}, c}$, whereas $\mathcal{M}^{[c]}$ emphasizes the gap between reaching E when passing through c and when bypassing c .

3.2 PROBABLY APPROXIMATELY CORRECT SPR CAUSE LEARNING

We provide a solution to Problem 1 in Algorithm 1. We first precompute a set of predetermined non-causal states (Line 2), then iteratively collect k transition samples and update the lower/upper bounds for $\Pr_{\mathcal{M}, s}^{\min}(\diamond E)$ and $\Pr_{\mathcal{M}^{[c]}}^{\max}(\diamond E)$ (Lines 4-6). Using Proposition 1, each remaining candidate state c is progressively assigned to $C_{\top}$ or $C_{\perp}$ when the corresponding bound intervals certify the ordering (Lines 7-12). With introducing a margin $\tau \geq 0$, the procedure supports a three-way outcome (causal/non-causal/undecided): for $\tau > 0$, a still-unclassified state c is assigned to $C_{?}$ once the current confidence interval of its reachability probability gap is entirely included by $[-\tau, \tau]$, capturing near-tie cases. In practice, we optionally stop bound updates for c and decide its class as soon as the compared intervals become disjoint, since the ordering is then fixed by monotonicity of (4) (see Lemma 7 in Appendix A). We describe a practical initialization and an early stopping scheme in Appendix B. After all candidates have been assigned, we re-check the classifications using Lines 6-12 for all classified states; if $(C_{\top}, C_{\perp}, C_{?})$ is unchanged, we return C^* as the “front” of $C_{\top}$ to satisfy Def. 1, and otherwise we resume learning.

Algorithm 1: PAC Cause Identification**Input:** Unknown MDP $\mathcal{M}$, $\delta_{\mathcal{M}} > 0$, $k > 0$, $\tau \geq 0$.**Output:** $C^* \subset S$

```

1  $\mathcal{D}, C_{\top}, C_{\perp}, C_{?} \leftarrow \emptyset$ , and  $\delta \leftarrow \frac{\delta_{\mathcal{M}}}{\text{Tr}(\mathcal{M})}$ .
2  $C_{\perp}^{\text{pre}} \leftarrow \{c \in S \mid \Pi_{\mathcal{M},c} = \emptyset\} \cup E$ .
3 repeat
4   Sample  $k$  transitions in  $\mathcal{M}$  and add them to  $\mathcal{D}$ .
5   Initialize lower and upper bounds:  $[\underline{x}_{\mathcal{M}}^{\min}, \bar{x}_{\mathcal{M}}^{\min}]$ , and
     for all  $c \in S \setminus C_{\perp}^{\text{pre}}$ ,  $[\underline{x}_{\mathcal{M}[c]}^{\max}, \bar{x}_{\mathcal{M}[c]}^{\max}]$ .
6   Compute the fixed points for (4).
7   for  $c \in S \setminus (C_{\top} \cup C_{\perp} \cup C_{?} \cup C_{\perp}^{\text{pre}})$  do
8     Add  $c$  to  $C_{\top}$  if  $\underline{x}_{\mathcal{M}}^{\min,*}(c) > \bar{x}_{\mathcal{M}[c]}^{\max,*}(s_I)$ .
9     Add  $c$  to  $C_{\perp}$  if  $\bar{x}_{\mathcal{M}}^{\min,*}(c) < \underline{x}_{\mathcal{M}[c]}^{\max,*}(s_I)$ .
10    if  $[\underline{x}_{\mathcal{M}}^{\min,*}(c) - \bar{x}_{\mathcal{M}[c]}^{\max,*}(s_I), \bar{x}_{\mathcal{M}}^{\min,*}(c) -$ 
11       $\underline{x}_{\mathcal{M}[c]}^{\max,*}(s_I)] \subseteq [-\tau, \tau]$  and  $c \notin C_{\top} \cup C_{\perp}$  then
12       $\tau > 0$ : Add  $c$  to  $C_{?}$ .
       $\tau = 0$ : Add  $c$  to  $C_{\top}$  if 3-(a) in Proposition 1
      is satisfied, otherwise add  $c$  to  $C_{\perp}$ .
13 until  $S \setminus (C_{\top} \cup C_{\perp} \cup C_{?} \cup C_{\perp}^{\text{pre}}) = \emptyset$ ;
14 if  $(C_{\top}, C_{\perp}, C_{?})$  remains unchanged under  $\mathcal{D}$  then
15   Return  $C^* = \{c \in C_{\top} \mid \exists \rho \text{ s.t. } \rho \models \neg C_{\top} \cup c\}$ .
16 Go back to Line 4 with  $C_{\top}, C_{\perp}, C_{?} \leftarrow \emptyset$ .

```

In the following Example 2, we show how well Algorithm 1 finds an SPR cause that satisfies (5).

Example 2. We applied Algorithm 1 to the MDP in Example 1. Let $\delta_{\mathcal{M}} = 0.05$ so that the found SPR cause is guaranteed w. p. at least 0.95, $\tau = 0$ to classify all states, and sample $k = 100$ transitions at each iteration by selecting k state-action pairs uniformly at random. We conducted 30 experimental runs. In every run, we consistently found s_3 as a causal state first and then obtained $C^* = \{s_2, s_3\}$. This is because s_3 has a higher difference in probabilities in terms of (C1) in Def. 1. Algorithm 1 converged within 130 iterations per run.

Theorem 2 shows that Algorithm 1 is a solution to Problem 1. Theorem 3 guarantees an anytime correctness, that is, $C_{\top}$ and $C_{\perp}$ in Algorithm 1 are always a correct set of causal and non-causal states, respectively, and $C_{?}$ contains only states with near-tie cases, with a high probability.

Theorem 2. For any MDP $\mathcal{M}$, any $\delta_{\mathcal{M}} > 0$, and any $\tau \geq 0$, C^* obtained from Algorithm 1 satisfies (5) for all SPR causes included by $S \setminus C_{?}$ w. p. at least $1 - \delta_{\mathcal{M}}$. Moreover, if $\tau = 0$, Algorithm 1 is a solution to Problem 1.

Theorem 3 (Anytime Correctness). For any MDP $\mathcal{M}$, any $\delta_{\mathcal{M}} > 0$, any $\tau \geq 0$, and any multi-sample $\mathcal{D}$ of transitions, $C_{\top}$, $C_{\perp}$, and $C_{?}$ computed in Algorithm 1, respectively, satisfy the following properties:

1. $\forall c \in C_{\top}, \{c\} \models (\text{C1})$ w. p. at least $1 - \delta_{\mathcal{M}}$,

2. $\forall c \in C_{\perp}, \{c\} \not\models (\text{C1})$ w. p. at least $1 - \delta_{\mathcal{M}}$,
3. $\forall c \in C_{?}, |\Pr_{\mathcal{M}}^{\min}(\Diamond E \mid \Diamond c) - \Pr_{\mathcal{M}}^{\max}(\Diamond E \mid \neg \Diamond c)| \leq \tau$ w. p. at least $1 - \delta_{\mathcal{M}}$.

Proofs for Theorems 2 and 3 are in Appendix A. Almost sure convergence to a true SPR cause that satisfies (5) when the amount of data goes to infinity is also guaranteed under some conditions (see Propositions 3 and 4 in Appendix A).

Remark 3 (On Weaker Assumptions for MDPs). The above guarantees rely on knowing the successor supports of $\mathcal{M}$, but this assumption can be relaxed. Def. 4 uses none of c 's successor support: it resets every action at c to s_I regardless of c 's original outgoing transitions, and it uses no reachability value of $\mathcal{M}$. Hence, the construction is unaffected if we weaken “successor supports are known” to, e.g., “only a lower bound on the minimum nonzero transition probability is known” [Ashok et al., 2019]. Algorithm 1 is still applicable through a black-box two-sided value iteration [Ashok et al., 2019]. In contrast, Def. 3 requires reachability information of the original MDP, so weakening the assumption loosens its estimated bounds and further hampers the separation check for (C1).

4 EXAMPLES

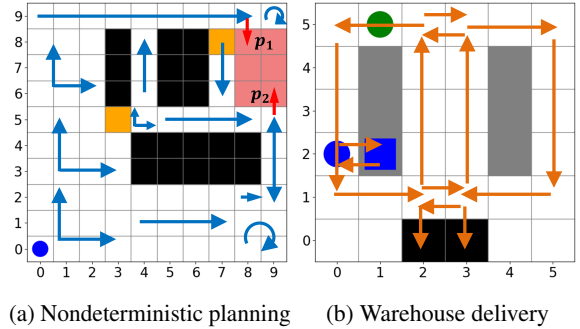


Figure 2: (a) Nondeterministic planning and (b) warehouse delivery. Blue/orange arrows denote enabled actions; black cells are obstacles (a) or goals (b); red cells are the undesired set E , entered with the probabilities p_1 and p_2 . Orange cells in (a) mark the SPR cause satisfying (5).

We demonstrate how correctly and quickly the proposed method identifies causal states in two environments: (i) a path planning scenario with nondeterministic actions and (ii) a warehouse delivery example under a reinforcement learning controller, as depicted in Figs. 2a and 2b. We implemented Algorithm 1 in Python 3.11 with the initialization and the early stopping described in Appendix B, and all experiments ran on a computer equipped with Intel Core i9-14900HX and 62GB of RAM. We compared our method against two baselines. Baseline1 runs Algorithm 2 in [Baier et al., 2022], but compares lower/upper bounds on reachability probabilities computed using [Ashok et al., 2019] from k

sampled transitions, as in Algorithm 1. It replaces the transition probabilities in Def. 3 with the estimated lower bounds and repeats the procedure until all states are classified. An algorithmic summary of Baseline1 is provided in Appendix C. Baseline2 first learns intervals for transition probabilities from k transition samples such that P is contained in the intervals w. p. at least $1 - \delta_{\mathcal{M}}$ [Suilen et al., 2022]. It then runs Algorithm 1 in [Oura and Ito, 2025] by instantiating 10 MDPs from these intervals. We treat an experiment as timed out if the iteration count exceeds 2×10^4 . For all methods, we sample k state-action pairs uniformly at random and draw transitions. This sampling scheme is an experimental choice, and we could instead use an exploration policy.

4.1 SETTING FOR PATH PLANNING

In the path-planning environment (Fig. 2a), the available actions are Up, Right, Down, Left, and Stay; entering the red cells is undesired. The robot moves in the intended direction with probability 0.9 and moves in a perpendicular direction with probability 0.05 each. It remains in the same cell if it chooses Stay or if the intended move would hit an obstacle (black cells) or leave the grid. When adjacent to walls or obstacles, actions that would move into them are disabled. When the robot is at $(9, 5)$ (resp., $(8, 9)$), it enters the red cells with probability p_1 (resp., p_2) and moves to the intended cell with the remaining probability. In the cells adjacent to the red cells, except for $(9, 5)$ and $(8, 9)$, the robot moves as intended with probability 1. We conducted 20 runs for each baseline and our method. We set $k = 5 \times 10^4$, $\delta_{\mathcal{M}} = 0.05$, and $(p_1, p_2) \in (0.2, 0.7), (0.4, 0.5)$. Experiments with other parameter settings are in Appendix D.

In this example, when $p_1 < p_2$, the true SPR cause is the orange cells shown in Fig. 2a. Intuitively, traversing these cells makes the robot more likely to reach the red cells from below via the middle corridor than via alternative routes. Our method gradually identifies these cells by comparing upper and lower bounds on reachability probabilities.

4.2 SETTING FOR WAREHOUSE DELIVERY

In the warehouse environment (Fig. 2b), two robots jointly learn policies via an automata-guided approach [Oura et al., 2020] to deliver the item (blue square) to a goal (black cells) while avoiding collisions, which constitute the undesired behavior. The enabled actions Up, Right, Down, Left, and Pickup for each region are indicated by orange arrows. The Pickup action is available only when a robot is at the item location. To encourage picking up the item, the blue robot is forced to choose Right with probability 0.4 when adjacent to the item location. Detailed parameter settings are provided in Appendix E. The undesired behavior is a collision. When not carrying an item, each robot moves in the intended

direction with probability 1. When carrying the item, the robot moves as intended with probability 0.6 and stays in place with probability 0.4, reflecting the slowdown due to the item’s weight. The state space consists of 2808 states. We set $k = 5 \times 10^5$, $\delta_{\mathcal{M}} = 0.1$, and conducted 20 runs for both our method and the baselines.

Intuitively, collisions are more likely when the green robot is on the left side and behind the blue robot while it is carrying the item, since the green robot can catch up with positive probability. In contrast, if the green robot stays on the right side of the warehouse while the blue robot is carrying the item, the collision probability is 0.

4.3 RESULTS

Performance comparisons. Our method successfully and quickly identified causal states during the learning process and produced SPR causes that satisfy (5) for both examples, whereas the baselines failed to do so. Table 1 reports the accuracy that states in $C_{\top}/C_{\perp}$ are truly causal/non-causal, and that C^* is an SPR cause satisfying (5); results are from our method with $\tau = 0$ and each baseline. Across all experiments, every intermediate state labeled causal/non-causal by our method was correct, and the output C^* was also correct. Baseline1 also produced correct labels, but required far more samples; in nondeterministic planning, all runs timed out and left $(3, 5)$ and $(4, 5)$ unclassified. This occurred because, in the transformed MDP in Def. 3, a policy maximizing reachability probability to E tends to take paths reaching $(3, 5)$ (and similarly $(4, 5)$), making the reachability gap between the original and transformed MDPs extremely small; our method mitigates this issue. Baseline2 converged quickly, but when $(p_1, p_2) = (0.4, 0.5)$ its success frequency fell below $1 - \delta_{\mathcal{M}}$. In some generated MDPs, transitions from $(9, 5)$ to the red cell had smaller/larger probabilities under Up/Down than those from $(7, 9)$, leading to the incorrect SPR cause $\{(3, 5), (7, 5)\}$. In contrast, the warehouse delivery example does not have nondeterministic actions, and the gap of reachability probabilities under a visit to a causal state and the bypass of it is not tiny. So, causal states are easier to detect. Moreover, for Baseline2, enforcing that both the lower bound of probability over the learned intervals that a returned set of states is an SPR cause and its confidence are near 1, e.g., $1 - 10^{-6}$, requires about 1.4×10^7 MDP instances by Theorem 1 in [Oura and Ito, 2025], which is infeasible since checking all states for one instantiated MDP took at least 10 seconds in both environments. Further experimental results are in Appendix D.

Convergence comparisons. Figs. 3a and 4c show, for the nondeterministic planning scenario, the numbers of causal and non-causal states correctly identified at each iteration by our method and Baseline1. For readability, we plot the

Table 1: Means and standard deviations (stds.) of the accuracies for the identified causal states, non-causal states, and final SPR causes, as well as of the number of iterations each method required to obtain the final outputs. The left and middle columns report results for the first nondeterministic planning example, and the right column reports results for the warehouse example. In all nondeterministic planning experiments, **Baseline1 timed out**.

	Path planning with $(p_1, p_2) = (0.4, 0.5)$				Path planning with $(p_1, p_2) = (0.2, 0.7)$				Warehouse			
	$C_\top$	$C_\perp$	C^*	Iter	$C_\top$	$C_\perp$	C^*	Iter	$C_\top$	$C_\perp$	C^*	Iter
Ours	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1449 ± 188	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	168 ± 5	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	57 ± 11
Baseline1	1.0 ± 0.0	1.0 ± 0.0	–	TO	1.0 ± 0.0	1.0 ± 0.0	–	TO	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	86 ± 12
Baseline2	–	–	0.25 ± 0.09	55 ± 69	–	–	0.95 ± 0.05	18 ± 2	–	–	1.0 ± 0.0	5 ± 4

results only up to 2000 and 300 iterations in Figs. 3a and 4c, respectively. Overall, our method identifies states earlier than Baseline1, and the advantage is more pronounced for causal states; particularly, when $(p_1, p_2) = (0.4, 0.5)$, the causal states are identified substantially earlier. Moreover, even when $(p_1, p_2) = (0.4, 0.5)$, our method classified all but two causal states into $C_\top$ within 300 iterations; only (3, 5) and (4, 5) remain unclassified.

Effects of τ and early stopping. Figs. 3c and 3d show that, as τ increases, $|C_\tau|$ increases while the number of iterations decreases, enabling faster convergence at the expense of more undecided states. In particular, when $\tau \geq 0.1$ and $(p_1, p_2) = (0.2, 0.7)$, our method converged in fewer than 30 iterations, with $|C_\tau|$ at most five at termination. For all τ , the causal/non-causal state classifications were correct. In Table 2, we compare the wall-clock time over the first 100 iterations with and without early stopping for bound updates. Early stopping was about $1.2\times$ faster on average and preserved the correctness of causal/non-causal state classification in all experiments.

Scalability. Our method is amenable to scaling. PR-cause checking for each state can be parallelized, since each candidate c is classified independently, and the interval value iteration can be GPU-accelerated [Farrington et al., 2025]. Both reduce wall-clock time without altering the algorithm. Realizing such parallel and GPU-accelerated implementations is left as future work.

Table 2: Means and standard deviations of the measured wall-clock times over the first 100 iterations, with and without the early stopping (ES) described in Sec. 3.2.

	$(p_1, p_2) = (0.4, 0.5)$	$(p_1, p_2) = (0.2, 0.7)$
Ours with ES	2140 ± 264 [sec.]	1971 ± 701 [sec.]
Ours w/o ES	2609 ± 747 [sec.]	2450 ± 1042 [sec.]

5 RELATED WORKS

5.1 CAUSALITY AND EXPLANATIONS FOR DECISION MAKING

A line of work has studied actual causality, often defined over known system models [Coenen et al., 2022, Finkbeiner

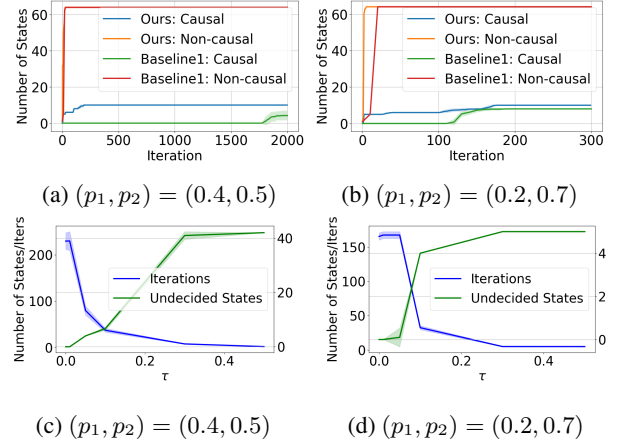


Figure 3: Nondeterministic planning. (a) and (b): mean numbers of causal/non-causal states identified by our method and Baseline1. (c) and (d): mean numbers of $|C_\tau|$ and iteration count vs. τ . Shaded areas show standard deviation; see Fig 4a-4i for more details.

et al., 2024a, Dimitrova et al., 2020, Clarkson et al., 2014, Finkbeiner and Siber, 2023, Finkbeiner et al., 2024b, Carelli et al., 2025, Oberst and Sontag, 2019, Tsirtsis et al., 2021]. These approaches aim to explain why a particular outcome occurred. Rafieioskouei et al. [2025] has considered causal states on tree-structured known Markov chains. In contrast, we focus on estimating certified causal states for probabilistic behavior over unknown MDPs. Closest to our setting are PR causes as sets of responsible states for Markovian models and computational procedures for them [Baier et al., 2021b, 2022, 2024]. However, existing approaches assume a known MDP, and applying their methods to an unknown MDP leads to sample inefficiency. As a context of explanations for verification, Wimmer et al. [2014], Feng et al. [2018], Jantsch et al. [2020], Jansen et al. [2011] have provided a subsystem and trajectories as evidence of a specification violation. However, they did not consider either unknown systems or causal principles. Recent works on causal explanation for reinforcement learning are also related [Madumal et al., 2020b,a, Yu et al., 2023], but they typically aim to explain agents' actions on fixed-length trajectories using learned or known SCM-like models. Moreover, they do not consider correctness and nondeterministic behavior.

5.2 TRANSFORMATIONS FOR CONDITIONAL QUERIES IN MARKOVIAN MODELS

Several works have studied the reduction of conditional reachability to unconditional reachability via model transformations [Baier et al., 2014, 2017, Märcker et al., 2017]. These methods and tools are primarily developed for known models; in the MDP setting, these pipelines require reachability probabilities in the original model during transformations, which are unavailable in unknown MDPs. Our proposed modification follows the same high-level idea but avoids querying reachability probabilities.

5.3 LEARNING-BASED VERIFICATION AND PR CAUSES UNDER UNCERTAINTY

PAC learning for unknown MDPs provides sampling-based verification and policy synthesis under specifications and stopping criteria [Ashok et al., 2019, Agarwal et al., 2025, Fu and Topcu, 2014, Wen and Topcu, 2016]. While these methods underpin our bound computation, they have mainly aimed at property verification rather than causal identification. PR causes have also been studied under explicit uncertain MDPs, where causes are inferred by reasoning over a specified uncertain set for transition probabilities [Oura and Ito, 2025]. Our focus instead is a single unknown MDP, and we target (i) sample-efficient learning of PR causes and (ii) anytime certification.

6 CONCLUSION

We studied PR-cause identification for reachability properties in unknown MDPs from transition samples. Our key ingredient is a restart-based MDP modification that addresses learning-specific challenges; we prove its correctness by reducing PR-cause checking to conditional reachability queries in the modified MDP, without requiring reachability values of the original MDP. We also derive and compare sample-complexity bounds for the existing and proposed MDP modifications. Building on two-sided value iteration, we develop an anytime learning-and-checking algorithm with probabilistic guarantees that progressively classifies candidate states as causal, non-causal, or undecided, and outputs a PR cause. Experiments on two benchmarks demonstrate that the approach reliably and quickly identifies the desired PR cause. Future work includes extending the approach to temporal logic formulas as causes, investigating adaptive sampling strategies, partial observability, and multi-agent settings, and considering other policy classes.

References

Chaitanya Agarwal, Shibashis Guha, Jan Křetínský, and M Pazhamalai. Pac statistical model checking of mean

payoff in discrete-and continuous-time mdp. *Formal Methods in System Design*, 66(2):195–237, 2025.

Pranav Ashok, Jan Křetínský, and Maximilian Weininger. Pac statistical model checking for markov decision processes and stochastic games. In *Proceedings of the 31st International Conference on Computer Aided Verification*, pages 497–519. Springer, 2019.

Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. MIT press, 2008.

Christel Baier, Joachim Klein, Sascha Klüppelholz, and Steffen Märcker. Computing conditional probabilities in markovian models efficiently. In *Proceedings of the 20th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 515–530. Springer, 2014.

Christel Baier, Joachim Klein, Sascha Klüppelholz, and Sascha Wunderlich. Maximizing the conditional expected reward for reaching the goal. In *Proceedings of the 23rd International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 269–285. Springer, 2017.

Christel Baier, Clemens Dubslaff, Florian Funke, Simon Jantsch, Rupak Majumdar, Jakob Piribauer, and Robin Ziemek. From verification to causality-based explications. In *Proceedings of the 48th International Colloquium on Automata, Languages, and Programming*, pages 1–1, 2021a.

Christel Baier, Florian Funke, Simon Jantsch, Jakob Piribauer, and Robin Ziemek. Probabilistic causes in Markov chains. In *Proceedings of the 19th International Symposium on Automated Technology for Verification and Analysis*, pages 205–221, 2021b.

Christel Baier, Florian Funke, Jakob Piribauer, and Robin Ziemek. On probability-raising causality in Markov decision processes. *arXiv preprint arXiv:2201.08768*, 2022.

Christel Baier, Jakob Piribauer, and Robin Ziemek. Foundations of probability-raising causality in Markov decision processes. *Logical Methods in Computer Science*, 20, 2024.

Mishel Carelli, Bernd Finkbeiner, and Julian Siber. Closure and complexity of temporal causality. *arXiv preprint arXiv:2505.10186*, 2025.

Shenghui Chen, Kayla Boggess, David Parker, and Lu Feng. Multi-objective controller synthesis with uncertain human preferences. In *Proceedings of the ACM/IEEE 13th International Conference on Cyber-Physical Systems*, pages 170–180, 2022.

- Michael R Clarkson, Bernd Finkbeiner, Masoud Koleini, Kristopher K Micinski, Markus N Rabe, and César Sánchez. Temporal logics for hyperproperties. In *Proceedings of the 3rd International Conference on Principles of Security and Trust*, pages 265–284. Springer Verlag, 2014.
- Norine Coenen, Bernd Finkbeiner, Hadar Frenkel, Christopher Hahn, Niklas Metzger, and Julian Siber. Temporal causality in reactive systems. In *Proceedings of the 20th International Symposium on Automated Technology for Verification and Analysis*, pages 208–224. Springer, 2022.
- Luca De Alfaro. *Formal verification of probabilistic systems*. stanford university, 1998.
- Rayna Dimitrova, Bernd Finkbeiner, and Hazem Torfah. Probabilistic hyperproperties of Markov decision processes. In *Proceedings of the 18th International Symposium on Automated Technology for Verification and Analysis*, pages 484–500. Springer, 2020.
- Klaus Drager, Vojtech Forejt, Marta Kwiatkowska, David Parker, and Mateusz Ujma. Permissive controller synthesis for probabilistic systems. *Logical Methods in Computer Science*, 11, 2015.
- Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.
- Joseph Farrington, Wai Keong Wong, Kezhi Li, and Martin Utley. Going faster to see further: graphics processing unit-accelerated value iteration and simulation for perishable inventory control using jax: J. farrington et al. *Annals of Operations Research*, 349(3):1609–1638, 2025.
- Lu Feng, Mahsa Ghasemi, Kai-Wei Chang, and Ufuk Topcu. Counterexamples for robotic planning explained in structured language. In *Proceedings of the International Conference on Robotics and Automation*, pages 7292–7297. IEEE, 2018.
- Bernd Finkbeiner and Julian Siber. Counterfactuals modulo temporal logics. In *Proceedings of the 24th International Conference on Logic*, volume 94, pages 181–204, 2023.
- Bernd Finkbeiner, Hadar Frenkel, Niklas Metzger, and Julian Siber. Synthesis of temporal causality. In *Proceedings of the 36th International Conference on Computer Aided Verification*, pages 87–111. Springer, 2024a.
- Bernd Finkbeiner, Felix Jahn, and Julian Siber. Counterfactual explanations for mitl violations. In *Proceedings of the 44th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 22–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2024b.
- Jie Fu and Ufuk Topcu. Probably approximately correct mdp learning and control with temporal logic constraints. In *Proceedings of Robotics: Science and Systems*, 2014.
- Serge Haddad and Benjamin Monmege. Interval iteration algorithm for mdps and imdps. *Theoretical Computer Science*, 735:111–131, 2018.
- Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- Nils Jansen, Erika Ábrahám, Jens Katelaan, Ralf Wimmer, Joost-Pieter Katoen, and Bernd Becker. Hierarchical counterexamples for discrete-time markov chains. In *Proceedings of the 9th International Symposium on Automated Technology for Verification and Analysis*, pages 443–452. Springer, 2011.
- Simon Jantsch, Florian Funke, and Christel Baier. Minimal witnesses for probabilistic timed automata. In *Proceedings of the 18th International Symposium on Automated Technology for Verification and Analysis*, pages 501–517. Springer, 2020.
- Sebastian Junges, Nils Jansen, Joost-Pieter Katoen, Ufuk Topcu, Ruohan Zhang, and Mary Hayhoe. Model checking for safe navigation among humans. In *Proceedings of the 15th International Conference on Quantitative Evaluation of Systems*, pages 207–222. Springer, 2018.
- Milad Kazemi, Jessica Lally, and Nicola Paoletti. Causal temporal reasoning for markov decision processes. *Research Directions: Cyber-Physical Systems*, 3:1–23, 2025.
- Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Distal explanations for explainable reinforcement learning agents. *arXiv preprint arXiv:2001.10284*, 2, 2020a.
- Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the 34th Annual AAAI conference on artificial intelligence*, volume 34, pages 2493–2500, 2020b.
- Steffen Märcker, Christel Baier, Joachim Klein, and Sascha Klüppelholz. Computing conditional probabilities: implementation and evaluation. In *Proceedings of the 15th International Conference on Software Engineering and Formal Methods*, pages 349–366. Springer, 2017.
- Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *Proceedings of the 15th International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

- Ryohei Oura and Yuji Ito. Probability-raising causality for uncertain parametric markov decision processes with pac guarantees. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence*, 2025.
- Ryohei Oura, Ami Sakakibara, and Toshimitsu Ushio. Reinforcement learning of control policy for linear temporal logic specifications using limit-deterministic generalized büchi automata. *Control Systems Letters*, 4(3):761–766, 2020.
- Mateo Perez, Fabio Somenzi, and Ashutosh Trivedi. A pac learning algorithm for ltl and omega-regular objectives in mdps. In *Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence*, volume 38, pages 21510–21517, 2024.
- Arshia Rafieioskouei, Kenneth Rogale, and Borzoo Bonakdarpour. Efficient discovery of actual causality with uncertainty. *arXiv preprint arXiv:2507.09000*, 2025.
- Marnix Suilen, Thiago D Simão, David Parker, and Nils Jansen. Robust anytime learning of markov decision processes. *Advances in Neural Information Processing Systems*, 35:28790–28802, 2022.
- Stelios Triantafyllou, Adish Singla, and Goran Radanovic. Actual causality and responsibility attribution in decentralized partially observable Markov decision processes. In *Proceedings of the 5th AAAI/ACM Conference on AI, Ethics, and Society*, pages 739–752, 2022.
- Stratis Tsirtsis, Abir De, and Manuel Rodriguez. Counterfactual explanations in sequential decision making under uncertainty. *Advances in Neural Information Processing Systems*, 34:30127–30139, 2021.
- Min Wen and Ufuk Topcu. Probably approximately correct learning in stochastic games with temporal logic specifications. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3630–3636, 2016.
- Ralf Wimmer, Nils Jansen, Erika Ábrahám, Joost-Pieter Katoen, and Bernd Becker. Minimal counterexamples for linear-time probabilistic verification. *Theoretical Computer Science*, 549:61–100, 2014.
- Zhongwei Yu, Jingqing Ruan, and Dengpeng Xing. Explainable reinforcement learning via a causal world model. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pages 4540–4548, 2023.

Supplementary Material

Ryohei Oura^{1,2}

Georgios Fainekos¹

Hideki Okamoto¹

Bardh Hoxha¹

¹Toyota Research Institute of North America, Toyota Motor North America

²Frontier Research Center, Toyota Motor Corporation

A PROOFS OF THEOREMS AND PROPOSITIONS

We start with proving that, for any policy $\pi : \text{FinPath} \times A \rightarrow [0, 1]$, there exists a policy $\tilde{\pi} : \text{FinPath} \times A \rightarrow [0, 1]$ such that the reachability probability conditioned on bypass of a state c on the original MDP under π is equivalent to the reachability probability on our modified MDP under $\tilde{\pi}$, that is, $\Pr_{\mathcal{M}}^{\pi}(\Diamond E \mid \neg \Diamond c) = \Pr_{\mathcal{M}^{[c]}}^{\tilde{\pi}}(\Diamond E)$. We then show that each state c can be classified as causal or non-causal using the modified MDP $\mathcal{M}^{[c]}$. Let InfPath be the set of infinite paths on $\mathcal{M}$. Recall that we denote $\Pi_{\mathcal{M},c} = \{\pi : \text{FinPath} \times A \rightarrow [0, 1] \mid \Pr_{\mathcal{M}}^{\pi}(\Diamond c) \in (0, 1)\}$ for any MDP $\mathcal{M}$ and any state $c \in S$. For any $\rho \in \text{FinPath}$, we denote $C_{\rho} = \{\rho' \mid \rho' \in \text{InfPath}, \rho \in \text{pref}(\rho')\}$, where $\text{pref}(\rho)$ denotes the set of prefixes of ρ . We define the concatenation $\rho\rho'$ of any $\rho = s_0a_0 \dots s_n$ and any $\rho' = s'_0a'_0 \dots s'_m$ such that $s_n = s'_0$ by $\rho\rho' = s_0a_0 \dots s_na'_0 \dots s'_m$. Moreover, for any $\rho = s_0a_0 \dots s_n \in \text{FinPath}$, we denote $\rho_t = s_t$, $\rho_{t:} = s_t a_t \dots s_n$, $\rho_{:,t} = s_0 a_0 \dots s_t$, and its length by $|\rho| = n + 1$. In what follows, without loss of generality, each terminal state $s \in E$ is treated as absorbing by adding a dummy self-loop action.

Proof of Lemma 1. Fix a state $c \in S \setminus E$ and a policy $\pi : \text{FinPath} \times A \rightarrow [0, 1]$. For any finite path ρ on $\mathcal{M}^{[c]}$, write $\rho = s_0a_0s_1 \dots a_{n-1}s_n$. We define *episode suffix* $\text{suf}_c(\rho)$ as the suffix of ρ that starts at the most recent occurrence of s_I that is the initial or is jumped from c : let

$$i^*(\rho) = \max\{i \mid s_i = s_I \text{ and } (i = 0 \text{ or } s_{i-1} = c)\}, \quad (11)$$

and set

$$\text{suf}_c(\rho) = \begin{cases} s_{i^*(\rho)}a_{i^*(\rho)} \dots a_{n-1}s_n & \text{if } i^*(\rho) < n, \\ s_n & \text{otherwise.} \end{cases} \quad (12)$$

Now we define the policy $\tilde{\pi} : \text{FinPath} \times A \rightarrow [0, 1]$ as follows: for any $\rho \in \text{FinPath}$ and any $a \in \mathcal{A}(\text{last}(\rho))$,

$$\tilde{\pi}(\rho, a) = \pi(\text{suf}_c(\rho), a). \quad (13)$$

For any $t > 0$, we consider $A_t = \{\rho = s_0a_0 \dots s_t a_t s_{t+1} \in \text{FinPath} \mid c \notin \rho_{:,t-1}, s_t = c, s_{t+1} = s_I\}$. Then, consider the event H_t of the first hitting time of c is $t < \infty$, that is, $H_t = \bigcup_{\rho \in A_t} C_{\rho}$. For any A_t , any $\rho = s_0a_0 \dots s_{t+1} \in A_t$, and any $\rho' = s_{t+1}a_{t+1} \dots s_n \in \text{FinPath}$ such that $s_{t+1} = s_I$, by (13), we have

$$\begin{aligned} \Pr_{\mathcal{M}^{[c]}}^{\tilde{\pi}}(C_{\rho\rho'}) &= \Pi_{i=0}^{n-1} P^c(s_i, a_i, s_{i+1}) \tilde{\pi}(s_0a_0 \dots s_i, a_i) \\ &= \Pi_{i=0}^t P^c(s_i, a_i, s_{i+1}) \tilde{\pi}(s_0a_0 \dots s_i, a_i) \Pi_{j=t+1}^{n-1} P^c(s_j, a_j, s_{j+1}) \tilde{\pi}(s_{t+1}a_{t+1} \dots s_j, a_j) \\ &= \Pr_{\mathcal{M}^{[c]}}^{\tilde{\pi}}(C_{\rho}) \Pr_{\mathcal{M}^{[c]}}^{\tilde{\pi}}(C_{\rho'}). \end{aligned} \quad (14)$$

Let $D_t(\rho') = \bigcup_{\rho \in A_t} C_{\rho\rho'}$. Since $C_{\rho_1} \cap C_{\rho_2} = \emptyset$ for any $\rho_1, \rho_2 \in A_t$ such that $\rho_1 \neq \rho_2$, we have, when $\Pr_{\tilde{\mathcal{M}}^{[c]}}(H_t) > 0$,

$$\begin{aligned} \Pr_{\tilde{\mathcal{M}}^{[c]}}(D_t(\rho') \mid H_t) &= \frac{\Pr_{\tilde{\mathcal{M}}^{[c]}}(D_t(\rho') \cap H_t)}{\Pr_{\tilde{\mathcal{M}}^{[c]}}(H_t)} \\ &= \frac{\sum_{\rho \in A_t} \Pr_{\tilde{\mathcal{M}}^{[c]}}(D_t(\rho') \cap C_\rho)}{\Pr_{\tilde{\mathcal{M}}^{[c]}}(H_t)} \\ &= \sum_{\rho \in A_t} \Pr_{\tilde{\mathcal{M}}^{[c]}}(D_t(\rho') \mid C_\rho) \Pr_{\tilde{\mathcal{M}}^{[c]}}(C_\rho \mid H_t). \end{aligned}$$

By (14), we have $\Pr_{\tilde{\mathcal{M}}^{[c]}}(C_{\rho\rho'} \mid C_\rho) = \Pr_{\tilde{\mathcal{M}}^{[c]}}(C_{\rho'})$ for each $\rho \in A_t$. Hence,

$$\begin{aligned} &= \Pr_{\tilde{\mathcal{M}}^{[c]}}(C_{\rho'}) \sum_{\rho \in A_t} \Pr_{\tilde{\mathcal{M}}^{[c]}}(C_\rho \mid H_t) \\ &= \Pr_{\tilde{\mathcal{M}}^{[c]}}(C_{\rho'}) \end{aligned} \quad (15)$$

Consider a set of paths of length $n+1$ starting from s_I that reach E : $\mathcal{P}_n = \{\rho' \in \text{FinPath} \mid \rho'_0 = s_I, |\rho'| = n+1, \exists i > 0 \text{ s.t. } \forall j \geq i, \rho_j \in E\}$, and let $B_{t,n} = \bigcup_{\rho' \in \mathcal{P}_n} D_t(\rho')$. By (15), we have

$$\begin{aligned} \Pr_{\tilde{\mathcal{M}}^{[c]}}(B_{t,n} \mid H_t) &= \sum_{\rho' \in \mathcal{P}_n} \Pr_{\tilde{\mathcal{M}}^{[c]}}(C_{\rho'}) \\ &= \Pr_{\tilde{\mathcal{M}}^{[c]}}\left(\bigcup_{\rho' \in \mathcal{P}_n} C_{\rho'}\right) \\ &= \Pr_{\tilde{\mathcal{M}}^{[c]}}(\diamond^{\leq n} E), \end{aligned} \quad (16)$$

where $\diamond^{\leq T} E$ denotes the event of reaching E within T steps. Note that, since E is the set of terminal states, if $m < n$, then, for any $\rho \in \mathcal{P}_m$, there exists $\rho' \in \mathcal{P}_n$ such that $\rho'_m = \rho$. Thus, $(B_{t,n})_{n>0}$ is increasing and $\bigcup_{n>0} B_{t,n} = \{\rho \in \text{InfPath} \mid \rho \models \diamond E\} \cap H_t$. Hence,

$$\Pr_{\tilde{\mathcal{M}}^{[c]}}(\diamond E \mid H_t) = \Pr_{\tilde{\mathcal{M}}^{[c]}}(\diamond E). \quad (17)$$

Thus, since $\diamond c$ represents $\bigcup_{t>0} H_t$ and $H_t \cap H_{t'} = \emptyset$ if $t \neq t'$, we have, when $\Pr_{\tilde{\mathcal{M}}^{[c]}}(\diamond c) > 0$,

$$\begin{aligned} \Pr_{\tilde{\mathcal{M}}^{[c]}}(\diamond E \mid \diamond c) &= \sum_{t>0} \Pr_{\tilde{\mathcal{M}}^{[c]}}(\diamond E \mid H_t) \Pr_{\tilde{\mathcal{M}}^{[c]}}(H_t \mid \diamond c) \\ &= \Pr_{\tilde{\mathcal{M}}^{[c]}}(\diamond E) \sum_{t>0} \Pr_{\tilde{\mathcal{M}}^{[c]}}(H_t \mid \diamond c) \\ &= \Pr_{\tilde{\mathcal{M}}^{[c]}}(\diamond E). \end{aligned} \quad (18)$$

By Def. 4 and (13), for any $\rho = s_0 a_0 \dots s_{t+1} \in \text{FinPath}$ such that $c \notin \rho_{:t}$, we have

$$\begin{aligned} \Pr_{\tilde{\mathcal{M}}}^\pi(C_\rho) &= \prod_{i=0}^t P(s_i, a_i, s_{i+1}) \pi(s_0 a_0 \dots s_i, a_i) \\ &= \prod_{i=0}^t P^c(s_i, a_i, s_{i+1}) \tilde{\pi}(s_0 a_0 \dots s_i, a_i) \\ &= \Pr_{\tilde{\mathcal{M}}^{[c]}}^{\tilde{\pi}}(C_\rho). \end{aligned} \quad (19)$$

Hence, combined with (18), we have

$$\Pr_{\tilde{\mathcal{M}}^{[c]}}^{\tilde{\pi}}(\diamond E) = \Pr_{\tilde{\mathcal{M}}^{[c]}}^{\tilde{\pi}}(\diamond E \wedge \neg \diamond c) + \Pr_{\tilde{\mathcal{M}}^{[c]}}^{\tilde{\pi}}(\diamond c) \Pr_{\tilde{\mathcal{M}}^{[c]}}^{\tilde{\pi}}(\diamond E) \quad (20)$$

$$= \Pr_{\tilde{\mathcal{M}}}^\pi(\diamond E \wedge \neg \diamond c) + \Pr_{\tilde{\mathcal{M}}}^\pi(\diamond c) \Pr_{\tilde{\mathcal{M}}^{[c]}}^{\tilde{\pi}}(\diamond E) \quad (21)$$

$$= \Pr_{\tilde{\mathcal{M}}}^\pi(\diamond E \mid \neg \diamond c). \quad (22)$$

Moreover, if π is memoryless, then $\pi(\rho, a) = \pi(\text{last}(\rho), a)$, and hence $\pi(\text{suf}_c(\rho), a) = \pi(\rho, a)$ for any $\rho \in \text{FinPath}$ and any $a \in \mathcal{A}(\text{last}(\rho))$. Thus, we have $\tilde{\pi} = \pi$. $\square$

Lemma 2. For any MDP $\mathcal{M}$, any policy π_1, π_2 , and any $\lambda \in [0, 1]$, there exists a policy π_λ such that, for any $\rho \in \text{FinPath}$,

$$\Pr_{\mathcal{M}}^{\pi_\lambda}(C_\rho) = \lambda \Pr_{\mathcal{M}}^{\pi_1}(C_\rho) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi_2}(C_\rho), \quad (23)$$

where $C_\rho = \{\rho\rho' \mid \rho' \in \text{InfPath}\}$ for any $\rho \in \text{FinPath}$.

Proof. For any π_1 and π_2 , and any $\lambda \in [0, 1]$, we define $\pi_\lambda : \text{FinPath} \times A \rightarrow [0, 1]$ such that, for any $\rho \in \text{FinPath}$,

$$\pi_\lambda(\rho, a) = \alpha(\rho)\pi_1(\rho, a) + (1 - \alpha(\rho))\pi_2(\rho, a), \quad (24)$$

where

$$\alpha(\rho) = \begin{cases} \frac{\lambda \Pr_{\mathcal{M}}^{\pi_1}(C_\rho)}{\lambda \Pr_{\mathcal{M}}^{\pi_1}(C_\rho) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi_2}(C_\rho)} & \text{if } \Pr_{\mathcal{M}}^{\pi_1}(C_\rho) > 0 \vee \Pr_{\mathcal{M}}^{\pi_2}(C_\rho) > 0, \\ 1 & \text{otherwise.} \end{cases} \quad (25)$$

We show the claim by induction.

1. If $|\rho| = 1$, that is the length of ρ is 0, then we have

$$\Pr_{\mathcal{M}}^{\pi_\lambda}(C_\rho) = \begin{cases} 1 & \text{if } \rho = s_I, \\ 0 & \text{otherwise.} \end{cases} \quad (26)$$

Likewise,

$$\lambda \Pr_{\mathcal{M}}^{\pi_1}(C_\rho) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi_2}(C_\rho) = \begin{cases} 1 & \text{if } \rho = s_I, \\ 0 & \text{otherwise.} \end{cases} \quad (27)$$

2. Suppose that we have (23) for any $\rho \in \text{FinPath}$ with $|\rho| = n$, then, for any $\rho' = \rho a s'$, we have

$$\Pr_{\mathcal{M}}^{\pi_\lambda}(C_{\rho'}) = \Pr_{\mathcal{M}}^{\pi_\lambda}(C_\rho) \pi_\lambda(\rho, a) P(s, a, s'),$$

where s is the last state of ρ , by the assumption and the definition of π_λ , we have

$$\begin{aligned} &= (\lambda \Pr_{\mathcal{M}}^{\pi_1}(C_\rho) \pi_1(\rho, a) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi_2}(C_\rho) \pi_2(\rho, a)) P(s, a, s') \\ &= \lambda \Pr_{\mathcal{M}}^{\pi_1}(C_{\rho'}) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi_2}(C_{\rho'}). \end{aligned} \quad (28)$$

□

Lemma 3. For any MDP $\mathcal{M}$, any policy π_1, π_2 , and any $\lambda \in [0, 1]$, there exists a policy π_λ such that, for any measurable sets $\mathcal{T} \subseteq \text{InfPath}$,

$$\Pr_{\mathcal{M}}^{\pi_\lambda}(\mathcal{T}) = \lambda \Pr_{\mathcal{M}}^{\pi_1}(\mathcal{T}) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi_2}(\mathcal{T}). \quad (29)$$

Proof. This immediately follows from Lemma 2 and Theorem 2.1.6 in [Durrett, 2019].

□

Based on Lemmas 1 and 3, we prove Proposition 1.

Proof of Proposition 1. Suppose $\Pi_{\mathcal{M},c} = \emptyset$. Then, $\{c\}$ violates the condition **(C2)** in Def. 1. Next, we consider $\Pi_{\mathcal{M},c} \neq \emptyset$.

1. Suppose $\Pr_{\mathcal{M},c}^{\min}(\diamond E) > \Pr_{\mathcal{M},c}^{\max}(\diamond E)$. Then, by Lemma 1, for any $\pi \in \Pi_{\mathcal{M},c}$, there exists a policy $\tilde{\pi}$ such that $\Pr_{\mathcal{M},c}^{\tilde{\pi}}(\diamond E) = \Pr_{\mathcal{M}}^{\pi}(\diamond E \mid \neg \diamond c)$. Thus,

$$\Pr_{\mathcal{M}}^{\pi}(\diamond E \mid \diamond c) - \Pr_{\mathcal{M}}^{\pi}(\diamond E) = (1 - \Pr_{\mathcal{M}}^{\pi}(\diamond c))(\Pr_{\mathcal{M}}^{\pi}(\diamond E \mid \diamond c) - \Pr_{\mathcal{M},c}^{\tilde{\pi}}(\diamond E)),$$

hence,

$$\begin{aligned} &\geq (1 - \Pr_{\mathcal{M}}^{\pi}(\diamond c))(\Pr_{\mathcal{M},c}^{\min}(\diamond E) - \Pr_{\mathcal{M},c}^{\max}(\diamond E)) \\ &> 0. \end{aligned} \quad (30)$$

2. Suppose $\Pr_{\mathcal{M},c}^{\min}(\diamond E) < \Pr_{\mathcal{M}[c]}^{\max}(\diamond E)$. Then, there exists $\pi_{2\text{mem}} : \text{FinPath} \rightarrow A$ such that $\pi_{2\text{mem}}$ is identical to a memoryless policy $\pi_{\max}$ with $\Pr_{\mathcal{M}[c]}^{\pi_{\max}}(\diamond E) = \Pr_{\mathcal{M}[c]}^{\max}(\diamond E)$ before hitting c , and identical to a memoryless policy $\pi_{\min}$ with $\Pr_{\mathcal{M},c}^{\pi_{\min}}(\diamond E) = \Pr_{\mathcal{M},c}^{\min}(\diamond E)$ after hitting c . Moreover, there exists a policy π such that π is identical to a memoryless policy π' with $\Pr_{\mathcal{M}}^{\pi'}(\diamond c) \in (0, 1)$ before hitting c , and identical $\pi_{\min}$ after hitting c . By Lemma 3, for any $\lambda \in (0, 1)$, there exists a policy π_{λ} such that, for any measurable set $\mathcal{T} \subseteq \text{InfPath}$,

$$\Pr_{\mathcal{M}}^{\pi_{\lambda}}(\mathcal{T}) = \lambda \Pr_{\mathcal{M}}^{\pi_{2\text{mem}}}(\mathcal{T}) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\mathcal{T}). \quad (31)$$

Hence,

$$\Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond c) \geq (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\diamond c) > 0, \quad (32)$$

$$\Pr_{\mathcal{M}}^{\pi_{\lambda}}(\neg \diamond c) \geq (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\neg \diamond c) > 0. \quad (33)$$

Moreover,

$$\begin{aligned} \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \diamond c) &= \frac{\lambda \Pr_{\mathcal{M}}^{\pi_{2\text{mem}}}(\diamond E \wedge \diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\diamond E \wedge \diamond c)}{\lambda \Pr_{\mathcal{M}}^{\pi_{2\text{mem}}}(\diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\diamond c)} \\ &= \Pr_{\mathcal{M},c}^{\min}(\diamond E). \end{aligned} \quad (34)$$

So, we have

$$\begin{aligned} \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \diamond c) - \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E) &= (1 - \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond c))(\Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \diamond c) - \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \neg \diamond c)) \\ &= (1 - \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond c))(\Pr_{\mathcal{M},c}^{\min}(\diamond E) - \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \neg \diamond c)). \end{aligned} \quad (35)$$

We consider the two cases.

- (a) $\Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \neg \diamond c) = \Pr_{\mathcal{M}}^{\max}(\diamond E \mid \neg \diamond c)$: then, we have

$$\begin{aligned} \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \neg \diamond c) &= \frac{\lambda \Pr_{\mathcal{M}}^{\pi_{\max}}(\diamond E \wedge \neg \diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\diamond E \wedge \neg \diamond c)}{\lambda \Pr_{\mathcal{M}}^{\pi_{\max}}(\neg \diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\neg \diamond c)} \\ &= \Pr_{\mathcal{M}}^{\pi_{\max}}(\diamond E \mid \neg \diamond c), \end{aligned} \quad (36)$$

by Lemma 1,

$$= \Pr_{\mathcal{M}[c]}^{\max}(\diamond E). \quad (37)$$

Thus, by the assumption, $\Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \diamond c) < \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E)$.

- (b) $\Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \neg \diamond c) < \Pr_{\mathcal{M}}^{\max}(\diamond E \mid \neg \diamond c)$: then,

$$\begin{aligned} \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \neg \diamond c) &= \frac{\lambda \Pr_{\mathcal{M}}^{\pi_{2\text{mem}}}(\diamond E \wedge \neg \diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\diamond E \wedge \neg \diamond c)}{\lambda \Pr_{\mathcal{M}}^{\pi_{2\text{mem}}}(\neg \diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\neg \diamond c)} \\ &= \frac{\lambda \Pr_{\mathcal{M}}^{\pi_{\max}}(\diamond E \wedge \neg \diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\diamond E \wedge \neg \diamond c)}{\lambda \Pr_{\mathcal{M}}^{\pi_{\max}}(\neg \diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\neg \diamond c)} \end{aligned} \quad (38)$$

is continuous with respect to λ . Moreover, as $\lambda \rightarrow 1$, we have $\Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \neg \diamond c) = \Pr_{\mathcal{M}}^{\pi_{\max}}(\diamond E \mid \neg \diamond c)$. So, by Lemma 1, $\Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \neg \diamond c) = \Pr_{\mathcal{M}[c]}^{\max}(\diamond E)$ as $\lambda \rightarrow 1$. Hence, combined with the assumption, there exists $\lambda < 1$ such that $\Pr_{\mathcal{M},c}^{\min}(\diamond E) < \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \neg \diamond c)$. Thus, $\Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E \mid \diamond c) < \Pr_{\mathcal{M}}^{\pi_{\lambda}}(\diamond E)$.

3. Suppose $\Pr_{\mathcal{M},c}^{\min}(\diamond E) = \Pr_{\mathcal{M}[c]}^{\max}(\diamond E)$. Then, we consider the following two cases.

- (a) We suppose c is not reachable from s_I in $\mathcal{M}_{\max}^{[c]}$. For any π such that $\Pr_{\mathcal{M}[c]}^{\pi}(\diamond E) = \Pr_{\mathcal{M}[c]}^{\max}(\diamond E)$, we have $\Pr_{\mathcal{M}[c]}^{\pi}(\diamond c) = 0$. Thus, for any $\tilde{\pi}$ such that $\Pr_{\mathcal{M}[c]}^{\tilde{\pi}}(\diamond c) \in (0, 1)$, we have $\Pr_{\mathcal{M}[c]}^{\tilde{\pi}}(\diamond E) < \Pr_{\mathcal{M}[c]}^{\max}(\diamond E)$. Thus, by Lemma 1, for any $\pi \in \Pi_{\mathcal{M},c}$, there exists $\tilde{\pi}$ such that

$$\Pr_{\mathcal{M}}^{\pi}(\diamond E \mid \diamond c) - \Pr_{\mathcal{M}}^{\pi}(\diamond E) = (1 - \Pr_{\mathcal{M}}^{\pi}(\diamond c))(\Pr_{\mathcal{M}}^{\pi}(\diamond E \mid \diamond c) - \Pr_{\mathcal{M}[c]}^{\tilde{\pi}}(\diamond E)) \quad (39)$$

$$\geq (1 - \Pr_{\mathcal{M}}^{\pi}(\diamond c))(\Pr_{\mathcal{M},c}^{\min}(\diamond E) - \Pr_{\mathcal{M}[c]}^{\tilde{\pi}}(\diamond E)) \quad (40)$$

by the assumption

$$= (1 - \Pr_{\mathcal{M}}^{\pi}(\diamond c))(\Pr_{\mathcal{M}^{[c]}}^{\max}(\diamond E) - \Pr_{\mathcal{M}^{[c]}}^{\bar{\pi}}(\diamond E)) \quad (41)$$

$$> 0. \quad (42)$$

Hence, the conditions **(C1)** and **(C2)** hold for c .

(b) We suppose c is reachable in $\mathcal{M}_{\max}^{[c]}$. Then, there exists a memoryless policy π such that $\Pr_{\mathcal{M}^{[c]}}^{\pi}(\diamond E) = \Pr_{\mathcal{M}^{[c]}}^{\max}(\diamond E)$ and $\Pr_{\mathcal{M}}^{\pi}(\diamond c) > 0$. We consider the two cases:

i. $\Pr_{\mathcal{M}}^{\pi}(\diamond c) = 1$: by Def. 4, we have $\Pr_{\mathcal{M}^{[c]}}^{\max}(\diamond E) = 0$. Hence, $\Pr_{\mathcal{M},c}^{\min}(\diamond E) = 0$. Thus, by Lemma 1, there exists a memoryless policy π' such that $\Pr_{\mathcal{M},c}^{\pi'}(\diamond E) = \Pr_{\mathcal{M},c}^{\min}(\diamond E)$,

$$\Pr_{\mathcal{M}}^{\pi'}(\diamond E \mid \diamond c) - \Pr_{\mathcal{M}}^{\pi'}(\diamond E) = (1 - \Pr_{\mathcal{M}}^{\pi'}(\diamond c))(\Pr_{\mathcal{M},c}^{\pi'}(\diamond E) - \Pr_{\mathcal{M}^{[c]}}^{\pi'}(\diamond E)) \quad (43)$$

$$= 0. \quad (44)$$

ii. $\Pr_{\mathcal{M}}^{\pi}(\diamond c) \in (0, 1)$: by Lemma 1, we have $\Pr_{\mathcal{M}}^{\pi}(\diamond E \mid \neg \diamond c) = \Pr_{\mathcal{M}^{[c]}}^{\max}(\diamond E)$. Thus, there exists a policy $\pi' \in \Pi_{\mathcal{M},c}$ such that $\Pr_{\mathcal{M}}^{\pi'}(\diamond E \mid \neg \diamond c) = \Pr_{\mathcal{M}^{[c]}}^{\max}(\diamond E)$ and $\Pr_{\mathcal{M},c}^{\pi'}(\diamond E) = \Pr_{\mathcal{M},c}^{\min}(\diamond E)$, hence we have

$$\Pr_{\mathcal{M}}^{\pi'}(\diamond E \mid \diamond c) - \Pr_{\mathcal{M}}^{\pi'}(\diamond E) = (1 - \Pr_{\mathcal{M}}^{\pi'}(\diamond c))(\Pr_{\mathcal{M},c}^{\pi'}(\diamond E) - \Pr_{\mathcal{M}}^{\pi'}(\diamond E \mid \neg \diamond c)) \quad (45)$$

$$= (1 - \Pr_{\mathcal{M}}^{\pi'}(\diamond c))(\Pr_{\mathcal{M},c}^{\min}(\diamond E) - \Pr_{\mathcal{M}^{[c]}}^{\max}(\diamond E)) \quad (46)$$

$$= 0. \quad (47)$$

□

To prove Theorems 1, 2, and 3, we additionally show the following Lemmas. Let $\text{Tr}(\mathcal{M})$ be the number of transitions of an MDP $\mathcal{M}$ whose probabilities are less than 1 and more than 0, that is, $\text{Tr}(\mathcal{M}) = |\{(s, a, s') \in S \times A \times S \mid P(s, a, s') \in (0, 1)\}|$.

Lemma 4. For any MDP, any random samples $\mathcal{D} = (s_i, a_i, s_{i+1})_i$, and any $\delta \in (0, 1)$, we have, for any $(s, a, s') \in S \times A \times S$,

$$\mathbb{P}_{\mathcal{D}} \left(\hat{P}_{\delta}(s, a, s') \leq P(s, a, s') \right) \geq 1 - \delta, \quad (48)$$

and

$$\mathbb{P}_{\mathcal{D}} \left(\hat{P}_{\delta} \leq P \right) \geq 1 - \delta \text{Tr}(\mathcal{M}), \quad (49)$$

where $\mathbb{P}_{\mathcal{D}}$ is the probability measure for $\mathcal{D}$.

Proof. For any $\mathcal{D}$, any $\delta \in (0, 1)$, and any transition (s, a, s') , we have

$$\mathbb{P}_{\mathcal{D}} \left(\hat{P}_{\delta}(s, a, s') \geq P(s, a, s') \right) = \mathbb{P}_{\mathcal{D}} \left(\frac{N(s, a, s')}{N(s, a)} - P(s, a, s') \geq \sqrt{\frac{-\ln \delta}{2N(s, a)}} \right)$$

by Hoeffding's inequality [Hoeffding, 1963],

$$\leq \delta. \quad (50)$$

Moreover, by Bool's inequality, we have

$$\mathbb{P}_{\mathcal{D}} \left(\exists (s, a, s') \in S \times A \times S \text{ s.t. } \hat{P}_{\delta}(s, a, s') \geq P(s, a, s') \right) \quad (51)$$

$$= \mathbb{P}_{\mathcal{D}} \left(\exists (s, a, s') \in \text{Tr}(\mathcal{M}) \text{ s.t. } \hat{P}_{\delta}(s, a, s') \geq P(s, a, s') \right) \quad (52)$$

$$\begin{aligned} &\leq \sum_{(s,a,s') \in \text{Tr}(\mathcal{M})} \mathbb{P}_{\mathcal{D}} \left(\hat{P}_{\delta}(s, a, s') \geq P(s, a, s') \right) \\ &\leq \delta \text{Tr}(\mathcal{M}). \end{aligned} \quad (53)$$

□

For any function $x : S \rightarrow \mathbb{R}$, we consider the infinite norm $\|\cdot\|_\infty$ such that $\|x\|_\infty = \max_{s \in S} |x(s)|$. The distance of functions x and x' under $\|\cdot\|_\infty$ is hence $\|x - x'\|_\infty = \max_{s \in S} |x(s) - x'(s)|$.

Lemma 5. *For any MDP $\mathcal{M}$ and any $\delta, \varepsilon \in (0, 1)$, if $\|P - \hat{P}_\delta\|_\infty \leq \varepsilon$, then, for any $x : S \rightarrow [0, 1]$ and any $\text{opt}^{(1)}, \text{opt}^{(2)} \in \{\min, \max\}$, we have $\|f_{\mathcal{M}, \delta}^{\text{opt}^{(1)}, \text{opt}^{(2)}}[x] - f_{\mathcal{M}}^{\text{opt}^{(1)}}[x]\|_\infty \leq d_{\max} \varepsilon$.*

Proof. Let $\Delta P(s, a, s') = P(s, a, s') - \hat{P}_\delta(s, a, s')$ for any $(s, a, s') \in S \times A \times S$. For any $x : S \rightarrow [0, 1]$ and any $s \in S$, we have

$$f_{\mathcal{M}}^{\text{opt}^{(1)}}[x](s) = \text{opt}_{a \in \mathcal{A}(s)}^{(1)} \left\{ \sum_{s' \in \text{Post}(s, a)} \hat{P}_\delta(s, a, s') x(s') + \sum_{s' \in \text{Post}(s, a)} \Delta P(s, a, s') x(s') \right\}, \quad (54)$$

$$\begin{aligned} f_{\mathcal{M}, \delta}^{\text{opt}^{(1)}, \text{opt}^{(2)}}[x](s) &= \text{opt}_{a \in \mathcal{A}(s)}^{(1)} \left\{ \sum_{s' \in \text{Post}(s, a)} \hat{P}_\delta(s, a, s') x(s') \right. \\ &\quad \left. + \text{opt}_{s' \in \text{Post}(s, a)}^{(2)} x(s') \Delta P(s, a, s') \right\}. \end{aligned} \quad (55)$$

Hence, noting that $\text{opt}_i^{(1)} x_i - \text{opt}_i^{(1)} y_i \leq \max_i |x_i - y_i|$ for any sequences of real numbers $(x_i)_i$ and $(y_i)_i$, we have

$$\begin{aligned} &|f_{\mathcal{M}}^{\text{opt}^{(1)}}[x](s) - f_{\mathcal{M}, \delta}^{\text{opt}^{(1)}, \text{opt}^{(2)}}[x](s)| \\ &\leq \max_a \left\{ \sum_{s'} \Delta P(s, a, s') x(s') - \text{opt}_{s'}^{(2)} x(s') \sum \Delta P(s, a, s') \right\} \\ &\leq \max_a \sum_{s'} \Delta P(s, a, s') \\ &\leq d_{\max} \varepsilon. \end{aligned} \quad (56)$$

□

Lemma 6. *For any MDP $\mathcal{M}$ and any $\delta, \varepsilon \in (0, 1)$, if $\|P - \hat{P}_\delta\|_\infty \leq \varepsilon$, then, for any $x : S \rightarrow [0, 1]$, any $T > 0$, and any $\text{opt}^{(1)}, \text{opt}^{(2)} \in \{\min, \max\}$, we have $\|(f_{\mathcal{M}, \delta}^{\text{opt}^{(1)}, \text{opt}^{(2)}})^T[x] - (f_{\mathcal{M}}^{\text{opt}^{(1)}})^T[x]\|_\infty \leq T d_{\max} \varepsilon$.*

Proof. For any $x : S \rightarrow [0, 1]$, we define $x_t : S \rightarrow [0, 1]$ and $z_t : S \rightarrow [0, 1]$ as follows:

$$x_0 = z_0 = x, \quad (57)$$

$$x_{t+1} = f_{\mathcal{M}, \delta}^{\text{opt}^{(1)}, \text{opt}^{(2)}}[x_t], \quad (58)$$

$$z_{t+1} = f_{\mathcal{M}}^{\text{opt}^{(1)}}[z_t]. \quad (59)$$

Then, we have

$$\begin{aligned} x_{t+1} - z_{t+1} &= f_{\mathcal{M}, \delta}^{\text{opt}^{(1)}, \text{opt}^{(2)}}[x_t] - f_{\mathcal{M}}^{\text{opt}^{(1)}}[z_t], \\ &= f_{\mathcal{M}, \delta}^{\text{opt}^{(1)}, \text{opt}^{(2)}}[x_t] - f_{\mathcal{M}}^{\text{opt}^{(1)}}[x_t] + f_{\mathcal{M}}^{\text{opt}^{(1)}}[x_t] - f_{\mathcal{M}}^{\text{opt}^{(1)}}[z_t]. \end{aligned} \quad (60)$$

Note that $f_{\mathcal{M}}^{\text{opt}^{(1)}}[x_t] - f_{\mathcal{M}}^{\text{opt}^{(1)}}[z_t] \leq \|x_t - z_t\|_\infty$. Hence, by Lemma 5,

$$\begin{aligned} \|x_T - z_T\|_\infty &\leq d_{\max} \varepsilon + \|x_{T-1} - z_{T-1}\|_\infty \\ &\leq T d_{\max} \varepsilon. \end{aligned} \quad (61)$$

□

Lemma 7. *For any MDP $\mathcal{M}$, any multi-sample $\mathcal{D}$ of transitions, and any $\delta \in (0, 1)$, if $\hat{P}_\delta \leq P$ holds, the operators $f_{\mathcal{M}, \delta}^{\text{opt}^{(1)}, \text{opt}^{(2)}}$ defined by (4) with $\text{opt}^{(1)}, \text{opt}^{(2)} \in \{\min, \max\}$ are monotone.*

Proof. For any $x, x' : S \rightarrow [0, 1]$ and any $s \in S$, we have

$$f_{\mathcal{M},\delta}^{\text{opt}^{(1)},\text{opt}^{(2)}}[x'](s) - f_{\mathcal{M},\delta}^{\text{opt}^{(1)},\text{opt}^{(2)}}[x](s) \quad (62)$$

$$= \text{opt}_{a \in \mathcal{A}(s)}^{(1)} \sum_{s' \in \text{Post}(s,a)} \hat{P}_\delta(s, a, s')(x'(s') - x(s')) \\ + (\text{opt}_{s'' \in \text{Post}(s,a)}^{(2)} x'(s'') - \text{opt}_{s'' \in \text{Post}(s,a)}^{(2)} x(s'')) \left(1 - \sum_{s' \in \text{Post}(s,a)} \hat{P}_\delta(s, a, s') \right). \quad (63)$$

Clearly $1 \geq \sum_{s' \in \text{Post}(s,a)} \hat{P}_\delta(s, a, s')$. Thus, if $x \leq x'$, then we have $f_{\mathcal{M},\delta}^{\text{opt}^{(1)},\text{opt}^{(2)}}[x'](s) - f_{\mathcal{M},\delta}^{\text{opt}^{(1)},\text{opt}^{(2)}}[x](s) \geq 0$, otherwise we have $f_{\mathcal{M},\delta}^{\text{opt}^{(1)},\text{opt}^{(2)}}[x'](s) - f_{\mathcal{M},\delta}^{\text{opt}^{(1)},\text{opt}^{(2)}}[x](s) \leq 0$. $\square$

Lemma 8. For any MDP $\mathcal{M}$ and any $\delta, \varepsilon \in (0, 1)$, if $\|P - \hat{P}_\delta\|_\infty \leq \varepsilon$, then, for any $x : S \rightarrow [0, 1]$, any $T > 0$, and any $\text{opt} \in \{\min, \max\}$, we have $\|\underline{x}_{\mathcal{M},\delta}^{\text{opt},*} - \text{Pr}_{\mathcal{M},\cdot}^{\text{opt}}(\diamond E)\|_\infty \leq \|(f_{\mathcal{M},\delta}^{\text{opt},\min})^T[x] - \text{Pr}_{\mathcal{M},\cdot}^{\text{opt}}(\diamond E)\|_\infty$ if $x \leq f_{\mathcal{M},\delta}^{\text{opt},\min}[x]$, and $\|\bar{x}_{\mathcal{M},\delta}^{\text{opt},*} - \text{Pr}_{\mathcal{M},\cdot}^{\text{opt}}(\diamond E)\|_\infty \leq \|(f_{\mathcal{M},\delta}^{\text{opt},\max})^T[x] - \text{Pr}_{\mathcal{M},\cdot}^{\text{opt}}(\diamond E)\|_\infty$ if $x \geq f_{\mathcal{M},\delta}^{\text{opt},\max}[x]$, where $\underline{x}_{\mathcal{M},\delta}^{\text{opt},*}$ and $\bar{x}_{\mathcal{M},\delta}^{\text{opt},*}$ are fixed points of $f_{\mathcal{M},\delta}^{\text{opt},\min}$ and $f_{\mathcal{M},\delta}^{\text{opt},\max}$ defined in (4), respectively.

Proof. Both inequalities immediately follow from Lemma 7. $\square$

Lemma 9. For MDP $\mathcal{M}$, any $c \in S$, and any $\delta, \varepsilon, \epsilon > 0$, if $\|P - \hat{P}_\delta\|_\infty \leq \varepsilon$, then we have

$$\|P^c - \hat{P}_\delta^c\|_\infty \leq \varepsilon, \quad (64)$$

$$\|P_{\text{mr}}^c - \hat{P}_{\text{mr},\delta}^c\|_\infty \leq T_{\mathcal{M},\epsilon,c}^{\min} d_{\max} \varepsilon + \epsilon, \quad (65)$$

where $T_{\mathcal{M},\epsilon,c}^{\min}$ is the minimal mixing time for ϵ defined by (5).

Proof. For any $\mathcal{M}$ and any $c \in S$, we have $P^c(s, a, s') = P(s, a, s')$ and $\hat{P}_\delta^c(s, a, s') = \hat{P}_\delta(s, a, s')$ for any $(s, a, s') \in S \setminus \{c\} \times A \times S$ and $P^c(c, a, s_I) = \hat{P}_\delta^c(c, a, s_I) = 1$ for any $a \in \mathcal{A}(c)$, thus

$$\|P^c - \hat{P}_\delta^c\|_\infty = \|P - \hat{P}_\delta\|_\infty \leq \varepsilon. \quad (66)$$

For any $(s, a, s') \in S \setminus \{c\} \times A \times S$, we have $P_{\text{mr}}^c(s, a, s') = P(s, a, s')$ and $\hat{P}_{\text{mr},\delta}^c(s, a, s') = \hat{P}(s, a, s')$, and, for any $a \in \mathcal{A}(c)$, we have $P_{\text{mr}}^c(c, a, s_{\text{eff}}) = \text{Pr}_{\mathcal{M},c}^{\min}(\diamond E)$, $P_{\text{mr}}^c(c, a, s_{\text{noeff}}) = 1 - \text{Pr}_{\mathcal{M},c}^{\min}(\diamond E)$, $\hat{P}_{\text{mr},\delta}^c(c, a, s_{\text{eff}}) = \underline{x}_{\mathcal{M},\delta}^{\min,*}(c)$, and $\hat{P}_{\text{mr},\delta}^c(c, a, s_{\text{noeff}}) = 1 - \bar{x}_{\mathcal{M},\delta}^{\min,*}(c)$. Thus, for any $\epsilon > 0$, there is a mixing time $T = T_{\mathcal{M},\epsilon,c}^{\min}$ such that, by Lemma 6,

$$\|P_{\text{mr}}^c - \hat{P}_{\text{mr},\delta}^c\|_\infty \leq \max\{|\text{Pr}_{\mathcal{M},c}^{\min}(\diamond E) - \underline{x}_{\mathcal{M},\delta}^{\min,*}(c)|, |\text{Pr}_{\mathcal{M},c}^{\min}(\diamond E) - (1 - \bar{x}_{\mathcal{M},\delta}^{\min,*}(c))|\} \\ \leq T d_{\max} \varepsilon + \epsilon. \quad (67)$$

$\square$

Theorem 1. We first show (9). By Hoeffding's inequality [Hoeffding, 1963] and Bool's inequality, for any $\delta_{\mathcal{M}} \in (0, 1)$ and any $(s, a) \in S \setminus E \times A$, if

$$N(s, a) \geq \frac{2}{\varepsilon^2} \ln \frac{2|\text{Tr}(\mathcal{M})|}{\delta_{\mathcal{M}}}, \quad (68)$$

then we have $\mathbb{P}_{\mathcal{D}}(\|P - \hat{P}_\delta\|_\infty \leq \varepsilon) \geq 1 - \delta_{\mathcal{M}}$, where $\delta = \delta_{\mathcal{M}}/|\text{Tr}(\mathcal{M})|$. We consider the two cases $\Delta_c > 0$ and $\Delta_c < 0$:

1. Suppose $\Delta_c > 0$. Then, let $\hat{\Delta}_c = \underline{x}_{\mathcal{M}}^{\min,*}(c) - \bar{x}_{\mathcal{M}[c]}^{\max,*}(s_I)$. We have

$$|\hat{\Delta}_c - \Delta_c| \leq |\underline{x}_{\mathcal{M}}^{\min,*}(c) - \text{Pr}_{\mathcal{M},c}^{\min}(\diamond E)| + |\bar{x}_{\mathcal{M}[c]}^{\max,*}(s_I) - \text{Pr}_{\mathcal{M}[c]}^{\max}(\diamond E)|,$$

by Lemma 8, for any $\epsilon > 0$, there is $T = \max\{T_{\mathcal{M},\epsilon,c}^{\min}, T_{\mathcal{M}^{[c]},\epsilon,s_I}^{\max}\}$ such that

$$\begin{aligned} & \leq |(f_{\mathcal{M},\delta}^{\min,\min})^T[\underline{x}_0](c) - \Pr_{\mathcal{M},c}^{\min}(\diamond E)| + |(f_{\mathcal{M}^{[c]},\delta}^{\max,\max})^T[\bar{x}_0](s_I) - \Pr_{\mathcal{M}^{[c]}}^{\max}(\diamond E)| \\ & \leq |(f_{\mathcal{M},\delta}^{\min,\min})^T[\underline{x}_0](c) - \Pr_{\mathcal{M},c}^{\pi_{\min}}(\diamond \leq^T E) + \Pr_{\mathcal{M},c}^{\pi_{\min}}(\diamond \leq^T E) - \Pr_{\mathcal{M},c}^{\min}(\diamond E)| \\ & \quad + |(f_{\mathcal{M}^{[c]},\delta}^{\max,\max})^T[\bar{x}_0](s_I) - \Pr_{\mathcal{M}^{[c]}}^{\pi_{\max}}(\diamond \leq^T E) + \Pr_{\mathcal{M}^{[c]}}^{\pi_{\max}}(\diamond \leq^T E) - \Pr_{\mathcal{M}^{[c]}}^{\max}(\diamond E)|, \end{aligned}$$

by Lemma 6 and (64) in Lemma 9, we have

$$\leq 2Td_{\max}\epsilon + 2\epsilon, \quad (69)$$

where $\underline{x}_0 : S \rightarrow [0, 1]$ and $\bar{x}_0 : S \rightarrow [0, 1]$ are lower and upper bounds for $\Pr_{\mathcal{M},c}^{\min}(\diamond E)$ and $\Pr_{\mathcal{M}^{[c]}}^{\max}(\diamond E)$, respectively, such that $\underline{x}_0 \leq f_{\mathcal{M},\delta}^{\min,\min}[\underline{x}_0]$ and $\bar{x}_0 \geq f_{\mathcal{M},\delta}^{\max,\max}[\bar{x}_0]$, and $\pi_{\min}$ and $\pi_{\max}$ are the policies such that $\Pr_{\mathcal{M},\cdot}^{\pi_{\min}}(\diamond E) = \Pr_{\mathcal{M},\cdot}^{\min}(\diamond E)$ and $\Pr_{\mathcal{M}^{[c]},\cdot}^{\pi_{\max}}(\diamond E) = \Pr_{\mathcal{M}^{[c]},\cdot}^{\max}(\diamond E)$, respectively. Since $|\hat{\Delta}_c - \Delta_c| < |\Delta_c|$ implies that the sign of $\hat{\Delta}_c$ is equal to the sign of Δ_c , we choose $\epsilon = \frac{\Delta_c}{8}$ and $\epsilon = \frac{\Delta_c}{8Td_{\max}}$, resulting $N(s, a)$ that satisfies (70) implies $\hat{\Delta}_c > 0$:

$$N(s, a) \geq 128 \frac{d_{\max}^2 T^2}{\Delta_c^2} \ln \frac{2|\text{Tr}(\mathcal{M})|}{\delta_{\mathcal{M}}}. \quad (70)$$

2. We can prove that (70) is a sufficient condition for $\bar{x}_{\mathcal{M}}^{\min,*}(c) < \underline{x}_{\mathcal{M}^{[c]}}^{\max,*}(s_I)$ when $\Delta_c < 0$ in the same way as in case 1.

Next, we show (10). We consider the two cases $\Delta_{\text{mr},c} > 0$ and $\Delta_{\text{mr},c} < 0$:

1. Suppose $\Delta_{\text{mr},c} > 0$. Then, let $\hat{\Delta}_{\text{mr},c} = \underline{x}_{\mathcal{M}}^{\min,*}(c) - \bar{x}_{\mathcal{M}^{[c]}}^{\max,*}(s_I)$. By Lemmas 6 and 8 and (65) in Lemma 9, for any $\epsilon > 0$, there is $T_{\text{mr}} = \min\{T > 0 \mid T \geq \max\{T_{\mathcal{M},\epsilon,c}^{\min}, T_{\mathcal{M}^{[c]},\epsilon,s_I}^{\max}\}\}$ with $\epsilon = \frac{|\Delta_{\text{mr},c}|}{8(Td_{\max}+1)}$ such that

$$\begin{aligned} |\Delta_{\text{mr},c} - \hat{\Delta}_{\text{mr},c}| & \leq 2T_{\text{mr}}d_{\max}(T_{\mathcal{M},\epsilon,c}^{\min}d_{\max}\epsilon + \epsilon) + 2\epsilon \\ & \leq 2T_{\text{mr}}^2d_{\max}^2\epsilon + 2(T_{\text{mr}}d_{\max} + 1)\epsilon. \end{aligned} \quad (71)$$

Since $|\hat{\Delta}_{\text{mr},c} - \Delta_{\text{mr},c}| < |\Delta_{\text{mr},c}|$ implies that the sign of $\hat{\Delta}_{\text{mr},c}$ is equal to the sign of $\Delta_{\text{mr},c}$, we choose $\epsilon = \frac{\Delta_{\text{mr},c}}{8T_{\text{mr}}^2d_{\max}^2}$, resulting $N(s, a)$ that satisfies (72) implies $\hat{\Delta}_c > 0$:

$$N(s, a) \geq 128 \frac{d_{\max}^4 T_{\text{rm}}^4}{\Delta_{\text{rm},c}^2} \ln \frac{2|\text{Tr}(\mathcal{M})|}{\delta_{\mathcal{M}}}. \quad (72)$$

2. We can prove that (72) is a sufficient condition for $\bar{x}_{\mathcal{M}}^{\min,*}(c) < \underline{x}_{\mathcal{M}^{[c]}}^{\max,*}(s_I)$ when $\Delta_{\text{rm},c} < 0$ in the same way as in case 1.

□

Lemma 10. For any MDP $\mathcal{M}$, any $\delta > 0$, any multi-sample $\mathcal{D}$ of transitions, and any lower and upper bounds $\underline{x}, \bar{x} : S \rightarrow [0, 1]$ for $\Pr_{\mathcal{M},s}^{\text{opt}}(\diamond E)$ with $\text{opt} \in \{\min, \max\}$, if $P \geq \hat{P}_{\delta}$ holds, then the updated functions $f_{\mathcal{M},\delta}^{\text{opt},\min}[\underline{x}]$ and $f_{\mathcal{M},\delta}^{\text{opt},\max}[\bar{x}]$ obtained from (4), respectively, are also lower and upper bounds for $\Pr_{\mathcal{M},s}^{\text{opt}}(\diamond E)$.

Proof. For any $s \in S$, we have

$$\begin{aligned} \Pr_{\mathcal{M},s}^{\text{opt}}(\diamond E) & \geq \text{opt}_{a \in \mathcal{A}(s)} \sum_{s' \in \text{Post}(s,a)} P(s, a, s') \underline{x}(s') \\ & = \text{opt}_{a \in \mathcal{A}(s)} \sum_{s' \in \text{Post}(s,a)} \hat{P}_{\delta}(s, a, s') \underline{x}(s') + \sum_{s' \in \text{Post}(s,a)} \left(P(s, a, s') - \hat{P}_{\delta}(s, a, s') \right) \underline{x}(s'), \end{aligned} \quad (73)$$

since $P(s, a, s') > \hat{P}_{\delta}(s, a, s')$, we have

$$\geq \text{opt}_{a \in \mathcal{A}(s)} \sum_{s' \in \text{Post}(s,a)} \hat{P}_{\delta}(s, a, s') \underline{x}(s') + \min_{s' \in \text{Post}(s,a)} \underline{x}(s') \left(1 - \sum_{s' \in \text{Post}(s,a)} \hat{P}_{\delta}(s, a, s') \right). \quad (74)$$

Likewise, for any $s \in S$, we have $\Pr_{\mathcal{M},s}^{\text{opt}}(\diamond E) \leq f_{\mathcal{M},\delta}^{\text{opt},\max}[\bar{x}](s)$.

□

In the following, we provide a key lemma to guarantee the correctness of Algorithm 1.

Lemma 11. For any MDP $\mathcal{M}$, any $\delta_{\mathcal{M}} > 0$, and any multi-sample $\mathcal{D}$ of transitions, if $\hat{P}_{\delta} \leq P$, then $C_{\top}$, $C_{\perp}$, $C_{?}$ obtained from each iteration in Algorithm 1, respectively, satisfy the following properties:

1. $\forall c \in C_{\top}, \{c\} \models (\mathbf{C1})$,
2. $\forall c \in C_{\perp}, \{c\} \not\models (\mathbf{C1})$.
3. $\forall c \in C_{?}, |\Pr_{\mathcal{M}}^{\min}(\Diamond E \mid \Diamond c) - \Pr_{\mathcal{M}}^{\max}(\Diamond E \mid \neg \Diamond c)| \leq \tau$.

Moreover, if $\hat{P}_{\delta} \leq P$, $C^* = \{c \in C_{\top} \mid \exists \rho \text{ s.t. } \rho \models \neg C_{\top} \cup c\}$ obtained from Line 15 in Algorithm 1 satisfies that, for any SPR cause $C \subseteq S \setminus C_{?}$,

$$\min_{\pi} \Pr_{\mathcal{M}}^{\pi}(\Diamond C^* \mid \Diamond E) - \Pr_{\mathcal{M}}^{\pi}(\Diamond C \mid \Diamond E) \geq 0. \quad (75)$$

Proof. First, we show properties 1 and 2. Under $\hat{P}_{\delta} \leq P$, by Lemma 10, if $\underline{x}_{\mathcal{M}}^{\min}(c) > \bar{x}_{\mathcal{M}^{[c]}}^{\max}(s_I)$ (resp., $\bar{x}_{\mathcal{M}}^{\min}(c) < \underline{x}_{\mathcal{M}^{[c]}}^{\max}(s_I)$) in Line 7, then we have $\Pr_{\mathcal{M},c}^{\min}(\Diamond E) > \Pr_{\mathcal{M}^{[c]}}^{\max}(\Diamond E)$ (resp., $\Pr_{\mathcal{M},c}^{\min}(\Diamond E) < \Pr_{\mathcal{M}^{[c]}}^{\max}(\Diamond E)$). Thus, by Proposition 1, c is a causal state (resp., a non-causal state). Hence, we have $\{c\} \models (\mathbf{C1})$ for any $c \in C_{\top}$ and $\{c\} \not\models (\mathbf{C1})$ for any $c \in C_{\perp}$. Next, we show property 3. For any $\tau > 0$ and any $c \in C_{?}$, under under $\hat{P}_{\delta} \leq P$, if we have

$$[\underline{x}_{\mathcal{M}}^{\min,*}(c) - \bar{x}_{\mathcal{M}^{[c]}}^{\max,*}(s_I), \bar{x}_{\mathcal{M}}^{\min,*}(c) - \underline{x}_{\mathcal{M}^{[c]}}^{\max,*}(s_I)] \subseteq [-\tau, \tau], \quad (76)$$

then, by Lemma 10, we have

$$|\Pr_{\mathcal{M}}^{\min}(\Diamond E \mid \Diamond c) - \Pr_{\mathcal{M}}^{\max}(\Diamond E \mid \neg \Diamond c)| \leq \tau. \quad (77)$$

Then we show (75), under $\hat{P}_{\delta} \leq P$. In Line 14 in Algorithm 1, we have $S \setminus C_{?} = C_{\top} \cup (C_{\perp} \cup C_{\perp}^{\text{pre}})$. Consider any SPR cause $C \subseteq S \setminus C_{?}$. We then have $C \subseteq C_{\top}$ by property 2 and Proposition 1. Thus, by the construction of C^* , we have

$$\forall \rho \in \text{InfPath}, \rho \models \Diamond C \implies \Diamond C^*, \quad (78)$$

and hence, for any policy π ,

$$\Pr_{\mathcal{M}}^{\pi}(\Diamond C \wedge \Diamond E) \leq \Pr_{\mathcal{M}}^{\pi}(\Diamond C^* \wedge \Diamond E). \quad (79)$$

Therefore, C^* satisfies (75). $\square$

Proof of Theorem 3. For any $c \in C_{\top}$, there exists an iteration number i such that c is assigned to $C_{\top}$ at i -th iteration and transition data $\mathcal{D}_i$ sampled up to i -th iteration. Then, by Lemma 11, we have

$$\begin{aligned} \mathbb{P}_{\mathcal{D}_i}(\{c\} \models (\mathbf{C1})) &\geq \mathbb{P}_{\mathcal{D}_i}(\{c\} \models (\mathbf{C1}) \mid \hat{P}_{\delta} \leq P) \mathbb{P}_{\mathcal{D}}(\hat{P}_{\delta} \leq P) \\ &= \mathbb{P}_{\mathcal{D}_i}(\hat{P}_{\delta} \leq P), \end{aligned} \quad (80)$$

where $\mathbb{P}_{\mathcal{D}_i}$ is the probability measure for $\mathcal{D}_i$. By Lemma 4, we have $\mathbb{P}_{\mathcal{D}_i}(\hat{P}_{\delta} \leq P) \geq 1 - \delta_{\mathcal{M}}$ and hence,

$$\geq 1 - \delta_{\mathcal{M}}. \quad (81)$$

We analogously derive properties 2 and 3 in Theorem 3 by Lemmas 4 and 11. $\square$

Proof of Theorem 2. By Lemma 11 and Def. 1, each SPR cause included by $S \setminus C_{?}$ is a subset of $C_{\top}$ under $\hat{P}_{\delta} \leq P$. Hence, for any multi-sample $\mathcal{D}$ of transitions, we have

$$\begin{aligned} \mathbb{P}_{\mathcal{D}}(C^* \text{ satisfies (5)}) &\geq \mathbb{P}_{\mathcal{D}}(C^* \text{ satisfies (5)} \mid \hat{P}_{\delta} \leq P) \mathbb{P}_{\mathcal{D}}(\hat{P}_{\delta} \leq P) \\ &= \mathbb{P}_{\mathcal{D}}(\hat{P}_{\delta} \leq P), \end{aligned} \quad (82)$$

where $\mathbb{P}_{\mathcal{D}}$ is the probability measure for $\mathcal{D}$. By Lemma 4, we have $\mathbb{P}_{\mathcal{D}}(\hat{P}_{\delta} \leq P) \geq 1 - \delta_{\mathcal{M}}$ and hence,

$$\geq 1 - \delta_{\mathcal{M}}. \quad (83)$$

$\square$

Moreover, the following Propositions 3 and 4 guarantee that fixed point operations in Lines 10 to 16 in Algorithm 1 converge if $\hat{P}_\delta \leq P$, and Algorithm 1 converges as the number of transition samples goes to infinity under mild conditions.

For any MDP $\mathcal{M}$, a pair (S', A') of a subset of states $S' \subseteq S$ and a subset of actions $A' \subseteq \bigcup_{s \in S'} \mathcal{A}(s)$ such that $A' \neq \emptyset$ is called an end component (EC) if and only if, (i) for any $s \in S'$ and any $a \in A' \cap \mathcal{A}(s)$, we have $\{s' \in S \mid P(s, a, s') > 0\} \subseteq S'$, and (ii) for any $s, s' \in S'$, there exists a path from s to s' under A' . An EC $\mathcal{E} = (S', A')$ is called a maximal EC (MEC) if there is no other EC $\mathcal{E}' = (S'', A'')$ such that $S' \subset S''$ or $A' \subset A''$. We say that an MEC (S', A') is a singleton MEC if and only if $|S'| = 1$. Moreover, an MEC (S', A') is called a bottom MEC (BMEC) if and only if $A' = \bigcup_{s \in S'} \mathcal{A}(s)$. We assume that all MECs of $\mathcal{M}$ are composed of singleton MECs and BMECs¹.

Our restart-based modification for MDPs, as defined in Def. 4, may introduce a non-singleton and non-bottom MEC. So, we compress the MEC into a single state in a standard way [De Alfaro, 1998] if such an MEC exists in $\mathcal{M}^{[c]}$. The following proposition characterizes the MECs in $\mathcal{M}^{[c]}$.

Proposition 2. *For any MDP $\mathcal{M}$ and any $c \in S$, if there exists a non-bottom and non-singleton MEC (S', A') in the modified MDP $\mathcal{M}^{[c]}$, then we have $s_I, c \in S'$.*

Proof. We prove by contradiction. Suppose that there exists a non-bottom MEC (S', A') in $\mathcal{M}^{[c]}$ and $A' \neq \bigcup_{s \in S'} \mathcal{A}(s)$ such that $|S'| > 1$ and (1) $c \notin S'$ or (2) $s_I \notin S'$ but $c \in S'$.

1. Since every MEC in $\mathcal{M}$ is either a singleton MEC or a BMEC, and all transitions from states other than c on $\mathcal{M}^{[c]}$ are identical to those on $\mathcal{M}$ by Def. 4, for any $S'' \subseteq S \setminus \{c\}$ and any $A' \subseteq \bigcup_{s \in S''} \mathcal{A}(s)$ such that $|S''| > 1$, if (S'', A') is an MEC, then it is a BMEC. This contradicts the assumption.
2. By Def. 4, if $c \in S'$, then $s_I \in S'$. However, this contradicts the assumption.

□

With a slight abuse of notation, we also denote the modified MDP after the state compression by $\mathcal{M}^{[c]} = (S, A, s_I, P^c)$. In Propositions 3 and 4, we treat the reduced version as the modified MDP.

Proposition 3. *For any MDP $\mathcal{M}$, any multi-sample $\mathcal{D}$ of transitions, and any $c \in S$, $\underline{x}_{\mathcal{M}}^{\min}$, $\bar{x}_{\mathcal{M}}^{\min}$, $\underline{x}_{\mathcal{M}^{[c]}}^{\max}$, and $\bar{x}_{\mathcal{M}^{[c]}}^{\max}$ converge to fixed points with probability at least $1 - \delta_{\mathcal{M}}$, respectively.*

Assumption 1. *For any MDP $\mathcal{M}$, any transition $(s, a, s') \in S \times A \times S$, if s is reachable from s_I , $a \in \mathcal{A}(s)$, and $P(s, a, s') < 1$, then, in Algorithm 1, $|\mathcal{D}_{s,a,s'}| \rightarrow \infty$ with probability 1 as the number of times transitions are sampled goes to ∞ , where $\mathcal{D}_{s,a,s'}$ denotes the number of samples of (s, a, s') .*

Proposition 4. *For any MDP $\mathcal{M}$ and any $\delta > 0$, under Assumption 1, as the number of multi-sample $\mathcal{D}$ goes to infinity, with probability 1, Algorithm 1 converges and C^* that satisfies (5) when $\tau = 0$.*

Proof of Proposition 3. Let $\{\underline{x}_{\mathcal{M}}^{\min,(n)}\}_{n=0}^{\infty}$ and $\{\bar{x}_{\mathcal{M}}^{\min,(n)}\}_{n=0}^{\infty}$ be the sequences derived by repeatedly applying $f_{\mathcal{M},\delta}^{\min,\min}$ and $f_{\mathcal{M},\delta}^{\min,\max}$ defined by (4), respectively, to the functions initialized in Lines 6 and 7, that is, $\underline{x}_{\mathcal{M}}^{\min,(i+1)} = f_{\mathcal{M},\delta}^{\min,\min}[\underline{x}_{\mathcal{M}}^{\min,(i)}]$ and $\bar{x}_{\mathcal{M}}^{\min,(i+1)} = f_{\mathcal{M},\delta}^{\min,\max}[\bar{x}_{\mathcal{M}}^{\min,(i)}]$ for any $i = 0, \dots$. By Lemma 7, under $\hat{P}_\delta \leq P$, there exists $j > 0$ such that the sequences $\{\underline{x}_{\mathcal{M}}^{\min,(n)}\}_{n=j}^{\infty}$ and $\{\bar{x}_{\mathcal{M}}^{\min,(n)}\}_{n=j}^{\infty}$ are monotonically increasing and decreasing, that is, $\underline{x}_{\mathcal{M}}^{\min,(i)} \leq \bar{x}_{\mathcal{M}}^{\min,(i+1)}$ and $\bar{x}_{\mathcal{M}}^{\min,(i)} \geq \underline{x}_{\mathcal{M}}^{\min,(i+1)}$ for any $i \geq j$. Thus, $\underline{x}_{\mathcal{M}}^{\min,(\infty)}$ and $\bar{x}_{\mathcal{M}}^{\min,(\infty)}$ are fixed points $\underline{x}_{\mathcal{M}}^{\min,*}$ and $\bar{x}_{\mathcal{M}}^{\min,*}$ of $f_{\mathcal{M},\delta}^{\min,\min}$ and $f_{\mathcal{M},\delta}^{\min,\max}$, respectively, since S is finite and $[0, 1]^{|S|}$ are closed. Thus, by Lemma 4, for each sequence $\{\underline{x}_{\mathcal{M}}^{\min,(n)}\}_n = \{\underline{x}_{\mathcal{M}}^{\min,(n)}\}_n, \{\bar{x}_{\mathcal{M}}^{\min,(n)}\}_n$, we have

$$\begin{aligned} \mathbb{P}_{\mathcal{D}}(\lim_{n \rightarrow \infty} \underline{x}_{\mathcal{M}}^{\min,(n)} = \underline{x}_{\mathcal{M}}^{\min,*}) &\geq \mathbb{P}_{\mathcal{D}}(\lim_{n \rightarrow \infty} \underline{x}_{\mathcal{M}}^{\min,(n)} = \underline{x}_{\mathcal{M}}^{\min,*} \mid \hat{P}_\delta \leq P) \mathbb{P}_{\mathcal{D}}(\hat{P}_\delta \leq P) \\ &\geq 1 - \delta_{\mathcal{M}}, \end{aligned} \tag{84}$$

where $\mathbb{P}_{\mathcal{D}}$ is the probability measure for $\mathcal{D}$.

¹Any non-bottom MEC can be replaced with a single state [De Alfaro, 1998, Haddad and Monmege, 2018]

In the same way, we can show that, for each sequence $\{x_{\mathcal{M}^{[c]}}^{\max,(n)}\}_n = \{\underline{x}_{\mathcal{M}^{[c]}}^{\max,(n)}\}_n, \{\bar{x}_{\mathcal{M}^{[c]}}^{\max,(n)}\}_n$, we have

$$\mathbb{P}_{\mathcal{D}}(\lim_{n \rightarrow \infty} x_{\mathcal{M}^{[c]}}^{\max,(n)} = x_{\mathcal{M}^{[c]}}^{\max,*} \geq \mathbb{P}_{\mathcal{D}}(\lim_{n \rightarrow \infty} x_{\mathcal{M}^{[c]}}^{\max,(n)} = x_{\mathcal{M}^{[c]}}^{\max,*} \mid \hat{P}_{\delta} \leq P) \mathbb{P}_{\mathcal{D}}(\hat{P}_{\delta} \leq P) \quad (85)$$

$$\geq 1 - \delta_{\mathcal{M}}. \quad (86)$$

□

Proof of Proposition 4. By law of large numbers and Assumption 1, $\hat{P}_{\delta}$ converges pointwise to P with probability 1 as $|\mathcal{D}|$ goes to ∞ . Hence, as $|\mathcal{D}|$ goes to ∞ , $f_{\mathcal{M},\delta}^{\min,\min}$ and $f_{\mathcal{M},\delta}^{\min,\max}$ defined by (4), respectively, converge to the following operator $f_{\mathcal{M}}^{\min} : [0, 1]^S \rightarrow [0, 1]^S$ with probability 1: for any $x : S \rightarrow [0, 1]$ and any $s \in S$,

$$f_{\mathcal{M}}^{\min}[x](s) = \min_{a \in \mathcal{A}(s)} \sum_{s' \in \text{Post}(s,a)} P(s, a, s')x(s'). \quad (87)$$

Likewise, when $|\mathcal{D}|$ goes to ∞ , for any $c \in S$, $f_{\mathcal{M}^{[c]},\delta}^{\max,\min}$ and $f_{\mathcal{M}^{[c]},\delta}^{\max,\max}$ defined by (4), respectively, are identical to the following operator $f_{\mathcal{M}^{[c]}}^{\max} : [0, 1]^S \rightarrow [0, 1]^S$ with probability 1: for any $x : S \rightarrow [0, 1]$ and any $s \in S$,

$$f_{\mathcal{M}^{[c]}}^{\max}[x](s) = \max_{a \in \mathcal{A}(s)} \sum_{s' \in \text{Post}(s,a)} P^c(s, a, s')x(s'), \quad (88)$$

where P^c is the transition probability function of $\mathcal{M}^{[c]}$. Since the sets of states of $\mathcal{M}$ are composed of non-ECs and BMECs, $\underline{x}_{\mathcal{M}}^{\min}(c)$ and $\bar{x}_{\mathcal{M}}^{\min}(c)$ converge to $\Pr_{\mathcal{M},c}^{\min}(\Diamond E)$ by Propositions 2 and 3 in [Haddad and Monmege, 2018]. Likewise, $\underline{x}_{\mathcal{M}^{[c]}}^{\max}(s_I)$ and $\bar{x}_{\mathcal{M}^{[c]}}^{\max}(s_I)$ converge to $\Pr_{\mathcal{M}^{[c]}}^{\max}(\Diamond E)$ by Propositions 4 and 5 in [Haddad and Monmege, 2018]. Thus, when $\tau = 0$ and $|\mathcal{D}|$ goes to ∞ , by Proposition 1, we have,

$$C_{\top} = \{c \in S \setminus C_{\perp}^{\text{pre}} \mid \{c\} \models (\mathbf{C1})\}, \quad (89)$$

$$C_{\perp} = \{c \in S \setminus C_{\perp}^{\text{pre}} \mid \{c\} \not\models (\mathbf{C1})\} \cup C_{\perp}^{\text{pre}}, \quad (90)$$

with probability 1, where $C_{\perp}^{\text{pre}} = \{c \in S \mid \Pi_{\mathcal{M},c} = \emptyset\} \cup E$. Therefore, with probability 1, Algorithm 1 converges and, by Lemma 11, outputs C^* that satisfies (5). □

We show that the soundness of our restart-based MDP modification remains intact even when considering the existential-quantifier variant of SPR causes.

Definition 6 (Existential SPR Cause). *For any MDP $\mathcal{M}$, the set of terminal states $E \subseteq S$, and a nonempty subset $C \subseteq S \setminus E$, we say that C is an existential strict probability-raising (eSPR) cause for E on $\mathcal{M}$ if and only if the following condition hold:*

$$\forall c \in C, \exists \pi \text{ s.t. } \Pr_{\mathcal{M}}^{\pi}(\neg CUc) \in (0, 1) \text{ and } \Pr_{\mathcal{M}}^{\pi}(\Diamond E \mid \neg CUc) > \Pr_{\mathcal{M}}^{\pi}(\Diamond E). \quad (91)$$

The difference of Def. 6 from Def. 1 comes from only the existential quantifier for the probability-raising condition (91). Intuitively, the condition requires that, for each state $c \in C$, there exists a policy under which traversal of c increases the likelihood of reaching E .

In the following, we denote the minimal and maximal probability on $\mathcal{M}$ over a set of policies Π by $\Pr_{\mathcal{M},s}^{\min\Pi}$ and $\Pr_{\mathcal{M},s}^{\max\Pi}$, respectively.

Proposition 5 (Soundness of Restart-based Modification for Def. 6). *For any MDP $\mathcal{M}$ and any $c \in S$, if $\Pi_{\mathcal{M},c} = \emptyset$, then c is not causal, otherwise, for any $\Pi \subseteq \{\pi : S \times A \rightarrow [0, 1] \mid \Pr_{\mathcal{M}}^{\pi}(\neg \Diamond c) > 0\}$, we have the following properties:*

1. *If $\Pr_{\mathcal{M},c}^{\max}(\Diamond E) > \Pr_{\mathcal{M}^{[c]}}^{\min\Pi}(\Diamond E)$, then $\{c\}$ is an eSPR cause.*
2. *If $\Pr_{\mathcal{M},c}^{\max}(\Diamond E) < \Pr_{\mathcal{M}^{[c]}}^{\min\Pi}(\Diamond E)$, then $\{c\}$ is not an eSPR cause.*

where $\mathcal{M}^{[c]}$ is defined by Def. 4 and $\Pi_{\mathcal{M},c}$ is defined by (7).

Proof. Suppose $\Pi_{\mathcal{M},c} = \emptyset$. Then, $\{c\}$ does not satisfy (91) in Def. 6. Next, we consider $\Pi_{\mathcal{M},c} \neq \emptyset$.

1. Suppose $\Pr_{\mathcal{M},c}^{\max}(\diamond E) > \Pr_{\mathcal{M}^{[c]}}^{\min}(\diamond E)$. Then, there exists a two-memory policy $\pi_{2\text{mem}} : \text{FinPath} \rightarrow A$ such that $\pi_{2\text{mem}}$ is identical to a memoryless policy $\pi_{\max}$ with $\Pr_{\mathcal{M},c}^{\max}(\diamond E) = \Pr_{\mathcal{M},c}^{\max}(\diamond E)$ after hitting c , and identical to a memoryless policy $\pi_{\min}$ with $\Pr_{\mathcal{M}^{[c]}}^{\min}(\diamond E) = \Pr_{\mathcal{M}^{[c]}}^{\min}(\diamond E)$ before hitting c . Moreover, there exists a policy π such that π is identical to a memoryless policy $\pi' \in \Pi_{\mathcal{M},c}$ before hitting c , and identical $\pi_{\max}$ after hitting c . By Lemma 3, for any $\lambda \in (0, 1)$, there exists a policy π_λ such that, for any measurable set $\mathcal{T} \subseteq \text{InfPath}$, we have

$$\Pr_{\mathcal{M}}^{\pi_\lambda}(\mathcal{T}) = \lambda \Pr_{\mathcal{M}}^{\pi_{2\text{mem}}}(\mathcal{T}) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\mathcal{T}), \quad (92)$$

Hence,

$$\Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond c) \geq (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\diamond c) > 0, \quad (93)$$

$$\Pr_{\mathcal{M}}^{\pi_\lambda}(\neg \diamond c) \geq (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\neg \diamond c) > 0. \quad (94)$$

We have

$$\begin{aligned} \Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E \mid \diamond c) - \Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E) &= (1 - \Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond c))(\Pr_{\mathcal{M},c}^{\pi_\lambda}(\diamond E) - \Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E \mid \neg \diamond c)) \\ &= (1 - \Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond c))(\Pr_{\mathcal{M},c}^{\max}(\diamond E) - \Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E \mid \neg \diamond c)) \end{aligned} \quad (95)$$

We consider the two cases.

- (a) $\Pr_{\mathcal{M}}^{\pi}(\diamond E \mid \neg \diamond c) = \Pr_{\mathcal{M}}^{\pi_{\min}}(\diamond E \mid \neg \diamond c)$: note that $\Pr_{\mathcal{M}}^{\pi}$ we have

$$\Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E \mid \neg \diamond c) = \Pr_{\mathcal{M}}^{\pi_{\min}}(\diamond E \mid \neg \diamond c), \quad (96)$$

by Lemma 1,

$$= \Pr_{\mathcal{M}^{[c]}}^{\min}(\diamond E). \quad (97)$$

Thus, by the assumption, $\Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E \mid \diamond c) > \Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E)$.

- (b) $\Pr_{\mathcal{M}}^{\pi}(\diamond E \mid \neg \diamond c) > \Pr_{\mathcal{M}}^{\pi_{\min}}(\diamond E \mid \neg \diamond c)$: then,

$$\begin{aligned} \Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E \mid \neg \diamond c) &= \frac{\lambda \Pr_{\mathcal{M}}^{\pi_{2\text{mem}}}(\diamond E \wedge \neg \diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\diamond E \wedge \neg \diamond c)}{\lambda \Pr_{\mathcal{M}}^{\pi_{2\text{mem}}}(\neg \diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\neg \diamond c)} \\ &= \frac{\lambda \Pr_{\mathcal{M}}^{\pi_{\min}}(\diamond E \wedge \neg \diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\diamond E \wedge \neg \diamond c)}{\lambda \Pr_{\mathcal{M}}^{\pi_{\min}}(\neg \diamond c) + (1 - \lambda) \Pr_{\mathcal{M}}^{\pi}(\neg \diamond c)} \end{aligned} \quad (98)$$

is continuous with respect to λ . Moreover, as $\lambda \rightarrow 1$, we have $\Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E \mid \neg \diamond c) = \Pr_{\mathcal{M}}^{\pi_{\min}}(\diamond E \mid \neg \diamond c)$. So, by Lemma 1, $\Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E \mid \neg \diamond c) = \Pr_{\mathcal{M}^{[c]}}^{\min}(\diamond E)$ as $\lambda \rightarrow 1$. Hence, combined with the assumption, there exists $\lambda < 1$ such that $\Pr_{\mathcal{M},c}^{\max}(\diamond E) > \Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E \mid \neg \diamond c)$. Thus, $\Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E \mid \diamond c) > \Pr_{\mathcal{M}}^{\pi_\lambda}(\diamond E)$.

2. Suppose that $\Pr_{\mathcal{M},c}^{\max}(\diamond E) < \Pr_{\mathcal{M}^{[c]}}^{\min}(\diamond E)$. Then, by Lemma 1, for any $\pi \in \Pi_{\mathcal{M},c}$, there exists a policy $\tilde{\pi}$ such that $\Pr_{\mathcal{M}^{[c]}}^{\tilde{\pi}}(\diamond E) = \Pr_{\mathcal{M}}^{\pi}(\diamond E \mid \neg \diamond c)$

$$\Pr_{\mathcal{M}}^{\pi}(\diamond E \mid \diamond c) - \Pr_{\mathcal{M}}^{\pi}(\diamond E) = (1 - \Pr_{\mathcal{M}}^{\pi}(\diamond c))(\Pr_{\mathcal{M},c}^{\pi}(\diamond E) - \Pr_{\mathcal{M}^{[c]}}^{\tilde{\pi}}(\diamond E)),$$

hence,

$$\begin{aligned} &\leq (1 - \Pr_{\mathcal{M}}^{\pi}(\diamond c))(\Pr_{\mathcal{M},c}^{\max}(\diamond E) - \Pr_{\mathcal{M}^{[c]}}^{\min}(\diamond E)) \\ &< 0. \end{aligned} \quad (99)$$

□

Example of Π in Proposition 5 based on MEC decomposition We can provide an example of Π for case 2 in Proposition 5 as follows. We assume that every MEC in $\mathcal{M}^{[c]}$ is either a singleton MEC or a BMEC. The goal is to exclude policies π such that $\Pr_{\mathcal{M}}^{\pi}(\diamond c) = 1$ by forbidding actions that keep the process inside the MEC containing c in $\mathcal{M}^{[c]}$. By Proposition 2, any non-singleton, non-bottom MEC in $\mathcal{M}^{[c]}$ must contains both s_I and c . Hence, preventing the policy from choosing the "staying" actions in the MEC rules out the almost-sure reachability to c from s_I . For each $c \in S$, we denote by (S_c, A_c) the MEC in $\mathcal{M}^{[c]}$ such that $c \in S_c$. Then, we choose Π as the set of policies that never choose actions in A_c when the current state is in S_c :

$$\Pi = \{\pi : \text{FinPath} \times A \rightarrow [0, 1] \mid \forall \rho \in \text{FinPath}, \text{last}(\rho) \in S_c \implies \sum_{a \in A_c \cap A(s)} \pi(\rho, a) = 0\}. \quad (100)$$

B INITIALIZATION AND EARLY STOPPING FOR ALGORITHM 1

Our restart-based modification defined by Def. 4 sometimes induces a slow computation for the upper bounds of $\Pr_{\mathcal{M}^{[c]}}^{\max}(\Diamond E)$ if the initialized function x for $f_{\mathcal{M}^{[c]},\delta}^{\max,\max}$ defined by (4) is loose, e.g., $x(s) = 0$ for any $s \in S_{\mathcal{M}}^{\max}$ and $x(s) = 1$ for any other states $s \notin S_{\mathcal{M}}^{\max}$. This is because the optimized policy under a loosely initialized function tends to make a behavior: repeatedly reaching c and resetting to s_I with a high probability. To mitigate this issue, we introduce a two-step initialization using a loosely initialized function. Given an MDP $\mathcal{M}$ and $\delta \in (0, 1)$, we consider the following two operators $g_{\mathcal{M}^{[c]},\delta}, F_{\mathcal{M}^{[c]},\delta} : [0, 1]^S \rightarrow [0, 1]^S$. For any $x \in [0, 1]^S$ and any $c \in S \setminus E$, $g_{\mathcal{M}^{[c]},\delta}[x](s) = 0$ for any $s \in S_{\mathcal{M}}^{\max} \cup \{c\}$, $g_{\mathcal{M}^{[c]},\delta}[x](s) = 1$ for any $s \in E$, and $g_{\mathcal{M}^{[c]},\delta}[x](s) = f_{\mathcal{M}^{[c]},\delta}^{\max,\max}[x](s)$ defined by (4) for any $s \in S \setminus (E \cup S_{\mathcal{M}}^{\max} \cup \{c\})$. For any $x \in [0, 1]^S$ and any $s \in S$, $F_{\mathcal{M}^{[c]},\delta}[x](s) = \max\{x(s), f_{\mathcal{M}^{[c]},\delta}^{\max,\max}[x](s)\}$. By definition, for any $x \in [0, 1]^S$, we have

$$g_{\mathcal{M}^{[c]},\delta}[x](s) \leq f_{\mathcal{M}^{[c]},\delta}^{\max,\max}[x](s), \quad (101)$$

$$F_{\mathcal{M}^{[c]},\delta}[x](s) \geq x(s). \quad (102)$$

Intuitively, the fixed point of g for an initial function x acts as an upper bound of $\Pr_{\mathcal{M}^{[c]}}^{\max}(\Diamond E \wedge \neg \Diamond c)$. This alleviates the aforementioned issue. However, the fixed point of g is smaller than the fixed point of $f_{\mathcal{M},\delta}^{\max,\max}$. So, we increase the fixed point function x' of $g_{\mathcal{M}^{[c]},\delta}$ so that we have $x' \geq f_{\mathcal{M},\delta}^{\max,\max}[x']$, i.e., a pre-fixpoint of $f_{\mathcal{M},\delta}^{\max,\max}$ by applying $F_{\mathcal{M}^{[c]},\delta}$. We summarize the initialization with an early stopping in Algorithm 2. Algorithm 2 can be used to implement Line 5 in Algorithm 1. After this computation, we resume Line 6 in Algorithm 1.

Algorithm 2: Initialization and Early Stopping for Algorithm 1

Input: Unknown MDP $\mathcal{M}$, $\delta_{\mathcal{M}} > 0$, $k > 0$, $\tau \geq 0$, $N > 0$.

// Initialize the lower and upper bounds

1 $\forall s, c \in S, \underline{x}_{\mathcal{M}}^{\min}(s) \leftarrow 0, \bar{x}_{\mathcal{M}}^{\min}(s) \leftarrow 1, \underline{x}_{\mathcal{M}^{[c]}}^{\max}(s) \leftarrow 0, \bar{x}_{\mathcal{M}^{[c]}}^{\max}(s) \leftarrow 1.$

2 $\forall s \in E, \forall s' \in S_{\mathcal{M}}^{\min}, \underline{x}_{\mathcal{M}}^{\min}(s) \leftarrow 1, \bar{x}_{\mathcal{M}}^{\min}(s') \leftarrow 0.$

3 $\forall c \in S, \forall s \in E, \forall s' \in S_{\mathcal{M}^{[c]}}^{\max}, \underline{x}_{\mathcal{M}^{[c]}}^{\max}(s) \leftarrow 1, \bar{x}_{\mathcal{M}^{[c]}}^{\max}(s') \leftarrow 0.$

// Bound updates and early stopping for non-causal states

4 **for** $i = 0, \dots, N$ **do**

5 $\underline{x}_{\mathcal{M}}^{\min} \leftarrow f_{\mathcal{M},\delta}^{\min,\min}[\underline{x}_{\mathcal{M}}^{\min}]$

6 $\bar{x}_{\mathcal{M}}^{\min} \leftarrow f_{\mathcal{M},\delta}^{\min,\max}[\bar{x}_{\mathcal{M}}^{\min}]$

7 **for** $c \in S \setminus (E \cup C_{\perp} \cup C_{\perp}^{\text{pre}})$ **do**

8 $\underline{x}_{\mathcal{M}^{[c]}}^{\max} \leftarrow f_{\mathcal{M}^{[c]},\delta}^{\max,\min}[\underline{x}_{\mathcal{M}^{[c]}}^{\max}]$

9 $\bar{x}_{\mathcal{M}^{[c]}}^{\max} \leftarrow g_{\mathcal{M}^{[c]},\delta}[\bar{x}_{\mathcal{M}^{[c]}}^{\max}]$

10 **if** $\bar{x}_{\mathcal{M}}^{\min}(c) < \underline{x}_{\mathcal{M}^{[c]}}^{\max}(s_I)$ **then**

11 Add $c \rightarrow \bar{C}_{\perp}$.

// Re-compute initial upper bound for maximal reachability in $\mathcal{M}^{[c]}$

12 **repeat**

13 $\bar{x}_{\mathcal{M}^{[c]}}^{\max} \leftarrow F_{\mathcal{M}^{[c]},\delta}[\bar{x}_{\mathcal{M}^{[c]}}^{\max}].$

14 **until** $\forall s \in S, \bar{x}_{\mathcal{M}^{[c]}}^{\max}(s) \geq f_{\mathcal{M},\delta}^{\max,\max}[\bar{x}_{\mathcal{M}^{[c]}}^{\max}](s);$

// Bound updates and early stopping for causal and non-causal states

15 **for** $i = 0, \dots, N$ **do**

16 $\underline{x}_{\mathcal{M}}^{\min} \leftarrow f_{\mathcal{M},\delta}^{\min,\min}[\underline{x}_{\mathcal{M}}^{\min}]$

17 $\bar{x}_{\mathcal{M}}^{\min} \leftarrow f_{\mathcal{M},\delta}^{\min,\max}[\bar{x}_{\mathcal{M}}^{\min}]$

18 **for** $c \in S \setminus (E \cup C_{\perp} \cup C_{\perp}^{\text{pre}})$ **do**

19 $\underline{x}_{\mathcal{M}^{[c]}}^{\max} \leftarrow f_{\mathcal{M}^{[c]},\delta}^{\max,\min}[\underline{x}_{\mathcal{M}^{[c]}}^{\max}]$

20 $\bar{x}_{\mathcal{M}^{[c]}}^{\max} \leftarrow g_{\mathcal{M}^{[c]},\delta}[\bar{x}_{\mathcal{M}^{[c]}}^{\max}]$

21 **if** $\bar{x}_{\mathcal{M}}^{\min}(c) < \underline{x}_{\mathcal{M}^{[c]}}^{\max}(s_I)$ **then**

22 Add $c \rightarrow \bar{C}_{\perp}$.

23 **else if** $\underline{x}_{\mathcal{M}}^{\min}(c) > \bar{x}_{\mathcal{M}^{[c]}}^{\max}(s_I)$ **then**

24 Add $c \rightarrow \bar{C}_{\top}$.

C DETAILS OF BASELINE 1

We provide an algorithmic summary of Baseline1, which adapts the model-checking procedure of [Baier et al., 2022] to the unknown-MDP setting by combining the transformed MDP of Def. 3 with the interval (two-sided) value iteration of [Ashok et al., 2019]. The procedure shares the same interval value iteration as our method (Algorithm 1) and differs only in the transformation used to check the probability-raising condition **(C1)**. $\mathcal{M}_{\text{mr}}^{[c]}$ requires the computation of $\Pr_{\mathcal{M},c}^{\min}(\Diamond E)$ for the transition from c to E . So, we estimate the lower bound $\hat{P}_{\text{mr},\delta}^c : S \times A \times S \rightarrow [0, 1]$ of the transition probability function of $\mathcal{M}_{\text{mr}}^{[c]}$ as

$$\hat{P}_{\text{mr},\delta}^c(s, a, s') = \begin{cases} \hat{P}_\delta(s, a, s') & \text{if } s \neq c, \\ \underline{x}_{\mathcal{M}}^{\min,*}(c) & \text{if } s = c, s' = s_{\text{eff}}, \\ 1 - \underline{x}_{\mathcal{M}}^{\min,*}(c) & \text{if } s = c, s' = s_{\text{noeff}}, \\ 0 & \text{otherwise,} \end{cases} \quad (103)$$

For each candidate state $c \in S \setminus E$, Baseline1 proceeds as Algorithm 3.

Algorithm 3: Baseline 1

Input: Unknown MDP $\mathcal{M}$, $\delta_{\mathcal{M}} > 0$, $k > 0$, $\tau \geq 0$.

Output: $C^* \subset S$

```

1  $\mathcal{D}, C_{\top}, C_{\perp}, C_{?} \leftarrow \emptyset$ , and  $\delta \leftarrow \frac{\delta_{\mathcal{M}}}{\text{Tr}(\mathcal{M})}$ .
2 repeat
3   Sample  $k$  transitions in  $\mathcal{M}$ , add them to  $\mathcal{D}$ , and compute  $\hat{P}_\delta$ .
4   Initialize lower and upper bounds of  $\Pr_{\mathcal{M}}^{\min}(\Diamond E)$ :  $[\underline{x}_{\mathcal{M}}^{\min}, \bar{x}_{\mathcal{M}}^{\min}]$ .
5   Compute the fixed points of  $\underline{x}_{\mathcal{M}}^{\min}$  and  $\bar{x}_{\mathcal{M}}^{\min}$  for (4).
6   For all  $c \in S \setminus E$ , construct  $\mathcal{M}_{\text{mr}}^{[c]}$  by (103) and initialize lower and upper bounds of  $\Pr_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max}(\Diamond E)$ :  $[\underline{x}_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max}, \bar{x}_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max}]$ .
7   Compute the fixed points of  $[\underline{x}_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max}, \bar{x}_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max}]$  for (4).
8   for  $c \in S \setminus (C_{\top} \cup C_{\perp} \cup C_{?} \cup E)$  do
9     Add  $c$  to  $C_{\top}$  if  $\underline{x}_{\mathcal{M}}^{\min,*}(c) > \bar{x}_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max,*}(s_I)$ .
10    Add  $c$  to  $C_{\perp}$  if  $\bar{x}_{\mathcal{M}}^{\min,*}(c) < \underline{x}_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max,*}(s_I)$ .
11    if  $[\underline{x}_{\mathcal{M}}^{\min,*}(c) - \bar{x}_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max,*}(s_I), \bar{x}_{\mathcal{M}}^{\min,*}(c) - \underline{x}_{\mathcal{M}_{\text{mr}}^{[c]}}^{\max,*}(s_I)] \subseteq [-\tau, \tau]$  and  $c \notin C_{\top} \cup C_{\perp}$  then
12      Add  $c$  to  $C_{\top}$  if 3-(a) in Lemma 2 in [Baier et al., 2022] is satisfied, otherwise add  $c$  to  $C_{\perp}$ .
13 until  $S \setminus (C_{\top} \cup C_{\perp} \cup C_{?} \cup E) = \emptyset$ ;
14 if  $(C_{\top}, C_{\perp}, C_{?})$  remains unchanged under  $\mathcal{D}$  then
15   Return  $C^* = \{c \in C_{\top} \mid \exists \rho \text{ s.t. } \rho \models \neg C_{\top} \cup c\}$ .
16 Go back to Line 4 with  $C_{\top}, C_{\perp}, C_{?} \leftarrow \emptyset$ .
```

D DETAILED EXPERIMENTAL RESULTS

Table 3 reports the accuracies of states identified as causal/non-causal during learning, that is, $C_{\top}$ and $C_{\perp}$, and the final outputs C^* for the nondeterministic path-planning examples under $\delta_{\mathcal{M}} \in \{0.05, 0.1\}$ and $(p_1, p_2) \in \{(0.4, 0.5), (0.3, 0.6), (0.2, 0.7)\}$. The results are obtained for our method and the baselines. Our method and Baseline1 produced correct results; however, Baseline1 left (3, 5) and (4, 5) unclassified. The number of iterations required by our method increases as the gap between p_1 and p_2 decreases, because the gap between reachability probabilities conditioned on visiting a candidate state and bypassing it also decreases. Baseline2 converged fastest in all experiments, but its accuracy dropped below $1 - \delta_{\mathcal{M}}$ when $(p_1, p_2) \in \{(0.4, 0.5), (0.3, 0.6)\}$.

Figs. 4a-4f show, for the same parameter settings, the numbers of causal and non-causal states correctly identified at each iteration by our method and Baseline1. In all cases, our method identified both causal and non-causal states more quickly. Baseline1 left (5, 3) and (5, 4) unclassified in all experiments, and it also left (6, 7), (7, 7), and (7, 8) unclassified when $(p_1, p_2) = (0.4, 0.5)$.

Figs. 4g-4i show the number of states labeled as undecided at termination and the number of iterations required by our method. As expected, the number of undecided states increases with τ . Nevertheless, the required number of iterations, and thus the number of transition samples, is substantially reduced at the expense of more undecided states.

Table 3: Means and standard deviations (stds.) of the accuracies of the identified causal states, non-causal states, and final SPR causes for the nondeterministic path-planning example, as well as of the number of iterations each method required to obtain the final outputs. In all nondeterministic planning experiments, Baseline1 timed out.

	$(p_1, p_2) = (0.4, 0.5), \delta_{\mathcal{M}} = 0.05$				$(p_1, p_2) = (0.3, 0.6), \delta_{\mathcal{M}} = 0.05$				$(p_1, p_2) = (0.2, 0.7), \delta_{\mathcal{M}} = 0.05$			
	$C_{\top}$	$C_{\perp}$	C^*	Iter	$C_{\top}$	$C_{\perp}$	C^*	Iter	$C_{\top}$	$C_{\perp}$	C^*	Iter
Ours	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1449 ± 188	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	284 ± 33	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	168 ± 5
Baseline1	1.0 ± 0.0	1.0 ± 0.0	–	TO	1.0 ± 0.0	1.0 ± 0.0	–	TO	1.0 ± 0.0	1.0 ± 0.0	–	TO
Baseline2	–	–	0.25 ± 0.09	55 ± 69	–	–	0.75 ± 0.10	23 ± 14	–	–	0.95 ± 0.05	18 ± 2
	$(p_1, p_2) = (0.4, 0.5), \delta_{\mathcal{M}} = 0.1$				$(p_1, p_2) = (0.3, 0.6), \delta_{\mathcal{M}} = 0.1$				$(p_1, p_2) = (0.2, 0.7), \delta_{\mathcal{M}} = 0.1$			
	$C_{\top}$	$C_{\perp}$	C^*	Iter	$C_{\top}$	$C_{\perp}$	C^*	Iter	$C_{\top}$	$C_{\perp}$	C^*	Iter
Ours	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1338 ± 148	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	267 ± 334	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	156 ± 5
Baseline1	1.0 ± 0.0	1.0 ± 0.0	–	TO	1.0 ± 0.0	1.0 ± 0.0	–	TO	1.0 ± 0.0	1.0 ± 0.0	–	TO
Baseline2	–	–	0.20 ± 0.09	44 ± 61	–	–	0.85 ± 0.08	31 ± 18	–	–	1.0 ± 0.0	18 ± 6

To show how causal states affect the system behavior corresponding to Fig. 2b, we depict two states in the SPR cause obtained from our method in Fig. 5. Each arrow represents the executed actions at the locations. Fig. 5a shows that the collision probability is increased if the blue robot tries to pick up the item when the green robot is still in the top-left side. Moreover, Fig. 5b indicates that the collision probability is increased if the green robot reenters the left area when the blue robot is delivering the item.

E PARAMETERS FOR EXPERIMENTS AND REINFORCEMENT LEARNING

We summarize the parameter settings in our experiments.

E.1 NONDETERMINISTIC PLANNING SCENARIO

Environmental parameters are as follows:

- (p_1, p_2) : $(0.2, 0.7)$, $(0.3, 0.6)$, $(0.4, 0.5)$.
- Transition probability to the intended direction: 0.9.
- Transition probability to the perpendicular directions: 0.05.
- Number of states: 81.
- Number of experimental runs: 20.

Hyperparameters in Algorithms 1 and 2 are as follows.

- k : 5×10^4 .
- $\delta_{\mathcal{M}}$: 0.05, 0.1.
- τ : 0, 0.01, 0.05, 0.1, 0.3, 0.5.
- N : 500.

E.2 DELIVERY IN WAREHOUSE

Environmental parameters are as follows:

- Transition probability to the intended direction when not carrying the item: 1.
- Transition probability to the intended direction when carrying the item: 0.6.
- Transition probability to the same cell when not carrying the item: 0.
- Transition probability to the same cell when carrying the item: 0.4.

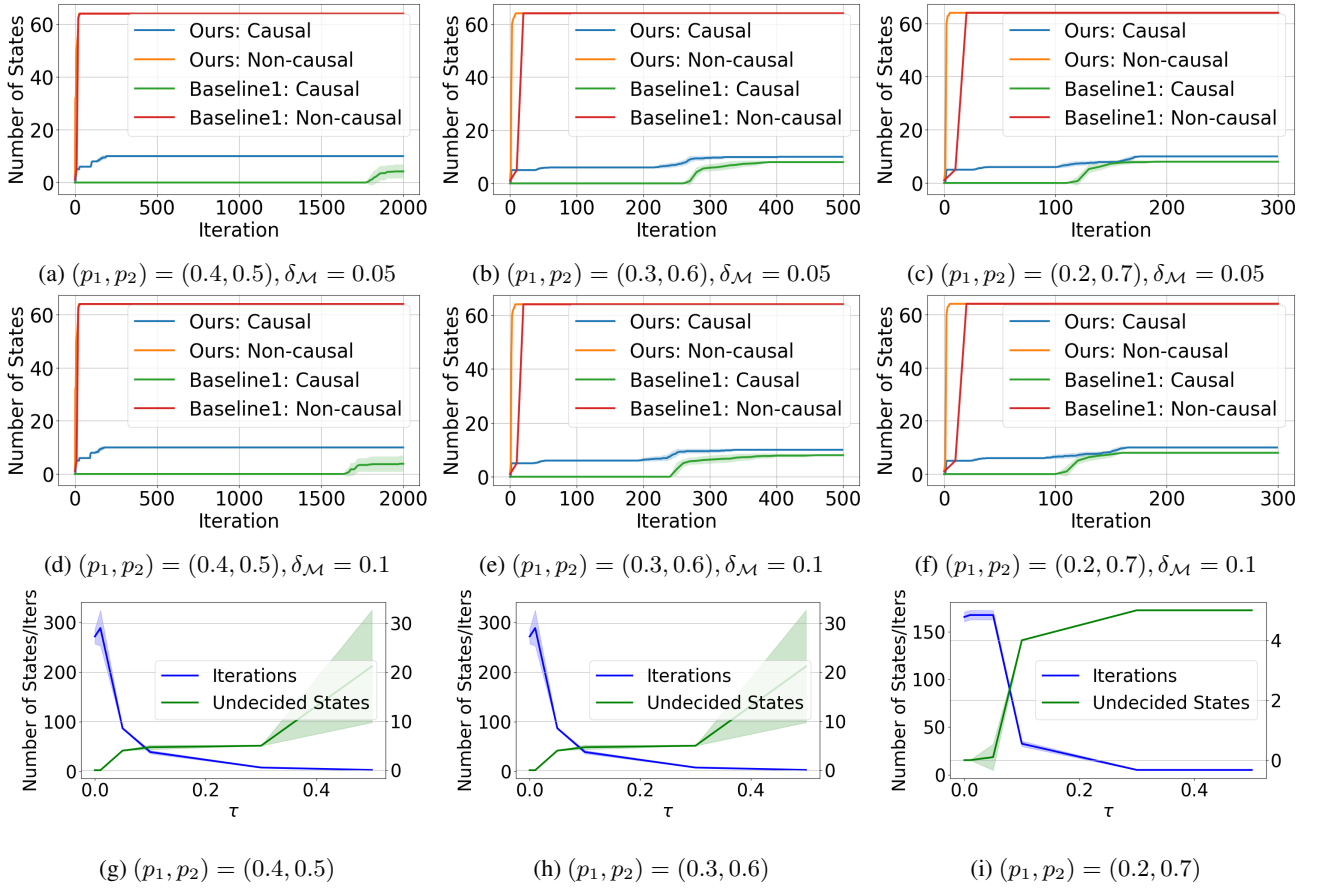


Figure 4: (a)–(f) show, for the nondeterministic planning setting, the mean numbers of causal and non-causal states identified by our method and Baseline1, respectively. (g)–(i) show, under $\delta_{\mathcal{M}} = 0.05$, the mean number of undecided states, $|C_2|$, and the mean number of iterations for our method. In (a)–(i), shaded areas indicate the standard deviation.

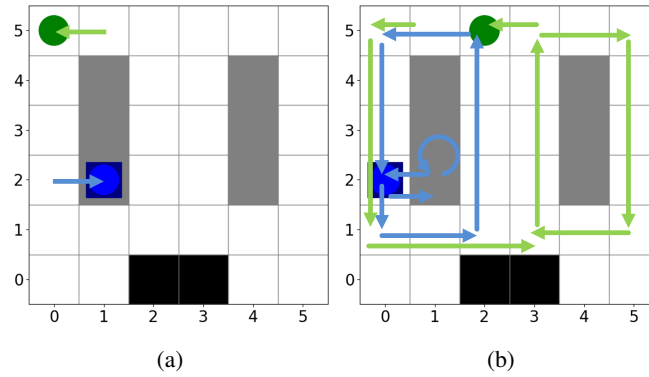


Figure 5: Two example states in the obtained SPR cause on the scenario shown in Fig. 2b identified by our method.

- Number of states: 2808.
- Number of experimental runs: 20.

Hyperparameters in Algorithms 1 and 2 are as follows.

- $k: 5 \times 10^5$.
- $\delta_{\mathcal{M}}: 0.1$.
- $\tau: 0$.
- $N: 500$.

E.3 SETTINGS FOR REINFORCEMENT LEARNING IN WAREHOUSE DELIVERY

We describe the experimental parameter settings for Q-learning in Section 4.2. The task is specified by the linear temporal logic formula $\varphi = \Diamond(\text{Pickup_Item} \wedge \Diamond\text{Reach_Goal}) \wedge \Box\neg\text{Collision}$ [Baier and Katoen, 2008]. The learning process is executed on the product space of the MDP and the deterministic Büchi automaton (DBA) translated from the task, that is, the set of states $S^\times = S \times Q$, where S is the set of states of the MDP and Q is the set of states of the DBA. Let $F^\times = S \times F$ be the set of accepting states in the product space, where F is the set of accepting states of the DBA. The value function $Q : S^\times \times A \rightarrow \mathbb{R}$ to be learned is defined as follows:

$$Q^*(s^\times, a) = \sup_{\pi \in \Pi^{\text{det}}} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(S_t^\times) \mid S_0^\times = s^\times, A_0 = a \right], \quad (104)$$

where the discount factor is given by $\gamma = 0.99$, Π^{det} denotes the set of deterministic policies $\pi : S^\times \rightarrow A$, and $\mathbb{E}_\pi$ denote the expectation under the policy π . The reward function $R : S^\times \rightarrow \{0, 1\}$ is defined as follows:

$$R(s^\times) = \begin{cases} 1 & \text{if } s^\times \in F^\times, \\ 0 & \text{otherwise,} \end{cases} \quad (105)$$

The update rule of the Q-learning is given by

$$Q(s_t^\times, a_t) \leftarrow Q(s_t^\times, a_t) + \alpha \left(R(s_t^\times) + \lambda \max_{a_{t+1} \in A(s_{t+1}^\times)} Q(s_{t+1}^\times, a_{t+1}) - Q(s_t^\times, a_t) \right), \quad (106)$$

where the learning rate $\alpha = 0.1$. During learning, we used the epsilon-greedy method to choose actions under the following scheduling for ε : For each learning step t , we compute $\varepsilon \in [0, 1]$ as

$$\varepsilon = \max\{\eta^t \varepsilon_0, \varepsilon_{\min}\}, \quad (107)$$

where $\eta = 0.995$, $\varepsilon = 1$, and $\varepsilon_{\min} = 0.1$. We conducted 10000 episodes to learn the deterministic controller $\pi : S^\times \rightarrow A$.

F APPLICATION SCENARIOS

We describe two concrete scenarios in which the identified causal states are useful.

Permissive robotic planning with unsafe terminal states. A cause defined by Def. 1 is a set of states whose visitation increases the probability of reaching bad outcomes regardless of how the permissive controller resolves its actions, and the non-vacuity condition (C2) removes states reachable only through other causal states. Here, the nondeterminism arises from the controller’s permissiveness, while the probabilistic transitions arise from the noisy environment or probabilistic events. The result is a compact set of “robustly responsible” states, which can guide local repair of the permissive controller toward avoiding the cause. In contrast, a cause defined by Def. 6 is a set of states whose visitation increases the probability of reaching bad outcomes under at least one way the controller resolves its actions, offering a set of “potentially responsible” states. This set is also useful for understanding the source of bad outcomes and for repairing the controller accordingly.

Linear temporal logic planning with a fixed controller and an adversary. Treating the accepting end components of a linear temporal logic task as the terminal set E , a cause defined by Def. 1 identifies states whose visitation increases the task-satisfaction probability no matter what actions the other agent takes. Here, the nondeterminism arises from the other agent's action selection, while the probabilistic transitions arise from the controller's action selection or the noisy environment. Under this definition, the other agent is treated adversarially, so the cause is a set of "robustly responsible" states that promote task satisfaction even in the worst case, which helps identify favorable conditions for achieving the task and repair the policy to be more robust against adversarial behavior. In contrast, a cause defined by Def. 6 identifies states whose visitation increases the task-satisfaction probability under at least one action selection of the other agent, which is meaningful when the other agent is cooperative: such "potentially responsible" states promote task satisfaction provided that the cooperative agent acts suitably, and they suggest favorable conditions to be encouraged through coordination.