Markov Decision Processes with Unobserved Confounders: A Causal Approach

Junzhe Zhang
Department of Computer Science
Purdue University
zhang745@purdue.edu

Elias Bareinboim
Department of Computer Science
Purdue University
eb@purdue.edu

September 19, 2016

Abstract

Markov decision processes (MDPs) constitute one of the most general frameworks for modeling decision-making under uncertainty, being used in multiple fields, including economics, medicine, and engineering. The goal of the agent in an MDP setting is to learn more about the environment so as to optimize a certain criterion. This task is pursued through the exploration of the environment by actively performing interventions (i.e., through the randomization of its actions), which contrasts with the agent passively observing the environment and not exerting any control over it (i.e., through random sampling). The existence of unobserved confounders, namely, unmeasured variables affecting both the action and the outcome or both the action and the state variables, implies that these two data-collection modes (passive and active) will in general not coincide. It is clear that by performing interventions, any potential inclination (intuition) of the agent will be ignored, which will imply a loss of information and failure to achieve an optimal behavior. In this paper, we formalize this observation and study its conceptual and algorithmic implications. We first demonstrate that standard algorithms may act sub-optimally when unobserved confounders are present. We then propose a systematic method to enhance these algorithms using causal inference theory and leveraging observational data. We formally and empirically show that this new approach produces superior results than current state-of-the-art MDP algorithms.

Introduction

Markov Decision Processes (MDPs) is one of the most general formalisms for modeling sequential decision-making under uncertainty within the reinforcement learning paradigm (Puterman, 1994; Sutton and Barto, 1998; Szepesvári, 2010; Bubeck and Cesa-Bianchi, 2012). In the MDP framework, the environment is modeled as a set of states and actions, where actions encode the autonomy of the agent to perform causal interventions. Agents learn about the environment by performing actions while trying to optimize a certain optimality criterion. One possible criterion is the minimization of the sample complexity of exploration, which is the number of timesteps for the agent to converge to the optimal policy. Here, the optimal policy is the one that maximizes the optimality criterion (e.g. cumulative reward). We will show that the definition of "optimal" is somewhat more involved when unobserved confounders are taken into account, which has conceptual and algorithmic implications.

The rich literature of MDPs encompasses a number of representations and algorithms for the different assumptions about the underlying data-generating model, including factored MDP (Dietterich, 1998), relational MDPs (Van Otterlo, 2009), Semi-Markov Decision Process (SMDP) (Puterman, 1994), Partially Observable Markov Decision Process (POMDP)

(Smallwood and Sondik, 1973; Ellis, Jiang, and Corotis, 1995; Singh, Jaakkola, and Jordan, 1994). For a survey, see (Sutton and Barto, 1998; Szepesvári, 2010).

Our work touches on another dimension of MDPs that has not been fully explored yet, the existence of unobserved confounders (UCs, for short). To understand the pervasiveness of the confounding problem, we first note that the goal of the randomization of the treatment assignment, as used in the causal inference literature, is precisely to eliminate the influence of unobserved confounders – factors that simultaneously affect the treatment (action) and the outcome (reward), but are unknown a priori in the analysis and not measured (Fisher, 1951; Pearl, 2000). [1]

The use of randomization in the actions' selection is a central component of the exploratory nature of RL algorithms and represents a distinguishing feature of the RL framework that contrasts with other modes of learning (e.g., supervised). Recently, (Bareinboim, Forney, and Pearl, 2015) noted a subtle property of the use of randomization in the context of Multi-Armed Bandits (MABs). Standard procedures based on randomization do not always reach an optimal behavior, and the agent's natural decision (without external intervention) is necessary for convergence. Perhaps surprisingly, there is more to the issue of confounding than simply randomizing when selecting actions. Bareinboim, Forney, and Pearl (2015) then explained that the natural decision of an agent (e.g., physician) without an external intervener (e.g., MAB algorithm) contains information about the UCs that is washed out by standard randomization. This loss of information can potentially mislead the agent in the evaluation of the underlying rewards' distribution and search for an optimal policy. They formalized and proposed a general solution to the problem of UCs in the context of MABs. [2]

The recent advances in the treatment of UCs in MABs do not directly translate to MDPs for different reasons. First, the type of confounding in MDPs is qualitatively different than in MABs since they not only affect the action and outcome, but can also affect state and outcome variables, or their combination, which require special treatment. Also, as opposed to MABs, the agent in an MDP setting cannot simply maximize the expected reward at each round, but instead has to evaluate policy's performance in the long-term. Finally, the interventions considered in MABs and MDPs are different – the former is atomic while the latter is conditional. These interventions entail different evaluations since conditioning may open up different back-door paths (Pearl, 2000, Ch. 3) and require a more refined analysis.

To the best of our knowledge, no systematic treatment for handling UCs in the context of MDPs has been developed. Further, no MDP algorithm has appropriately treated different data-collection modes – i.e., passively interacting with the environment (without intervention) versus interacting with the environment through active interventions (randomizing the actions). [3] In this paper, we explicitly acknowledge, formalize, and then exploit these different data-collection modes to solve MDP with UCs (MDPUC, for short). Specifically, our contributions are as follow:

1. We show that standard MDP algorithms are not guaranteed to learn an optimal policy in the presence of unobserved confounders in a general class of models.

2. We represent the MDP problem in causal language and compare two sets of candidate policies: experimental and counterfactual. We prove that a strategy that explores counterfactual policies outperforms standard procedures, which consider only experimental policies.

[1] Confounding represents a major challenge in tasks where policy-learning is required, but performing experiments is not feasible (Simpson, 1951; Pearl, 2000; Bareinboim and Pearl, 2016).

[2] UCs are automatically avoided in many tasks in the RL literature – e.g., off-policy evaluation is valid since randomization neutralizes the effect of the UCs, which makes agents interchangeable (Szepesvári, 2010; Li et al., 2011). Here, we are interested in new learning opportunities opened up by UCs in some general settings.

[3] This dichotomy is related in behavioral modeling to the notions of "natural" and "controlled" environments (Willems, 1989).

3. We propose a simple modification to empower standard MDP algorithms so as they can search in the space of counterfactual policies. We run simulations and show that the new algorithm is both efficient, stable, and outperforms state-of-the-art procedures.

Preliminaries

In the remainder of this section, we review the basic machinery used throughout the paper.

Definition 1. (MDP (Bertsekas and Tsitsiklis, 1995)). A Markov decision process (MDP) is a tuple $\langle S, X, T, R \rangle$ in which S is a finite set of states, X a finite set of actions, T a transition function defined as $T : S \times X \times S \to [0, 1]$, and R a reward function defined as $R : S \times X \to \mathbb{R}$.

Let $Y^{(t)}, X^{(t)}, S^{(t)}$ denote corresponding variables for the reward, action, and state at round $t \in \mathbb{N}^+$. A policy π is a mapping that assigns weights to each action $x \in X$ given state $s \in S$. A deterministic policy π is a function defined as $\pi : S \to X$. A stochastic policy is defined as $\pi : S \times X \to [0, 1]$ such that $\sum_{x \in X} \pi(s, x) = 1$, for $s \in S$. We focus here on infinite-horizon discounted MDPs where the goal is to maximize the discounted cumulative reward $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t Y^{(t)}]$ with discount factor $\gamma \in [0, 1)$.

The language of structural causal models will play a central role in the analysis of MDPs since it will allow the articulation of concepts such as confounding, observational and experimental distributions, and counterfactuals (Pearl, 2000). We introduce key causal concepts and notation next.

Definition 2. (SCM (Pearl, 2000, pp. 203-205)). A Structural Causal Model (SCM) is a tuple $\langle U, V, F, P(u) \rangle$ in which U is a set of exogenous (unobserved) variables, V is a set of endogenous (observed) variables, F is a set of structural equations such that for each $V_i \in V$, $f_i \in F : V_i \leftarrow f_i(PA_i, U_i)$, where $U_i \subseteq U$, $PA_i \subseteq V \backslash V_i$, and $P(u)$ is a probability distribution over U.

Each SCM M has an associated causal diagram G, where the nodes represent the endogenous variables V and the edges represent the functional relationships (the arguments of the structural equations in F). Within the structural semantics, performing an action $X = x$ is represented through the do-operator, $do(X = x)$, which encodes the operation of replacing the original equation of X by the constant x and induces a submodel M_x (with equations $F_x = \{X \leftarrow x\} \cup F \setminus \{f_x\}$). The effect of $do(X = x)$ on a variable Y is described probabilistically as $P(Y_{X=x} = y)$. Similarly, performing an action $X = \pi(z)$ is represented as $do(X = \pi(z))$, where π is a policy function that takes z as an argument. Let $P(Y_{x=\pi})$ denote the effect of $do(X = \pi(z))$ on a variable Y when π decides for x and the argument z is dropped for simplicity. [4] We are finally ready to define counterfactuals.

Definition 3. (Counterfactuals (Pearl, 2000, pp. 204)). Given a SCM M and X and Y two subsets of endogenous variables in V, the counterfactual sentence "The value that Y would have obtained, had X been x (in situation $U = u$)" is interpreted as denoting the potential response $Y_x(u)$ – the solution for Y of the set of equations F_x in submodel M_x, where $F_x = \{X \leftarrow x\} \cup F \setminus \{f_x\}$.

We use capital letters to represent variables and small letters to their values, $P(y_x|z)$ to represent $P(Y_{X=x} = y|Z = z)$, and $X^{([i,j])}$ to represent the sequence starting at $X^{(i)}$ and going until $X^{(j)}$ (i.e., $(X^{(i)}, X^{(i+1)}, \ldots, X^{(j)})$). We use the vertical line to represent evaluation, e.g., $P(Y_x = y|z)\big|_{x=\pi}$ represents $\sum_{x \in X} P(Y_x = y|z) I\{x = \pi(z)\}$, where $I\{\cdot\}$ is the indicator function.

[4] For a detailed discussion on the properties of structural models, we refer readers to (Pearl, 2000, Ch. 7).

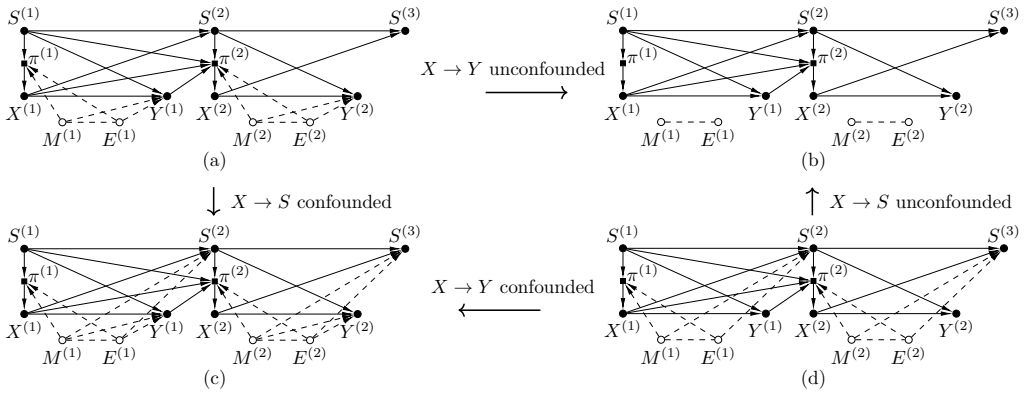

Figure 1: (a) MDPUC instance with $x^{(t)} \to y^{(t)}$ confounded. (b) Graphical representation for an unconfounded MDP instance. (c) MDPUC instance with both $x^{(t)} \to s^{(t+1)}$ and $x^{(t)} \to y^{(t)}$ confouned. (d) MDPUC instance with $x^{(t)} \to s^{(t+1)}$ confounded.

Practical Challenges Due to Unobservable Confounders

We discuss in this section practical challenges presented to MDPs due to the existence of unobserved confounders.

Medical treatment. A physician treats patients with a serious disease who have to visit the hospital regularly. At the t-th visit, the physician measures the patient's corticosteroid level $S^{(t)} \in \{0, 1\}$, where 0 stands for a low and 1 for a high level of corticosteroid. She then decides whether to give or not the drug to the patient, respectively, $X^{(t)} = \{1, 0\}$, and then measures an overall health score $Y^{(t)} \in \{1, 0\}$ (i.e., "healthy" and "not healthy"). The goal is to maximize the cumulative health score of the patient in the long run. Let the discount factor be $\gamma = 0.99$. This problem can be modelled as an MDP where the optimality criterion is to maximize the cumulative discounted reward.

In reality, the patient's health score $Y^{(t)}$ is affected not only by $X^{(t)}$ and $S^{(t)}$, but also by confounders such as the patient's mood $M^{(t)} \in \{0, 1\}$ (0 for positive, 1 for negative) and socioeconomic status (SES) $E^{(t)} \in \{0, 1\}$ (0 for wealthy, 1 for poor). The physician decides whether to give the drug by a criterion, which is computed (consciously or subconsciously) by a structural equation, for example, $X^{(t)} \leftarrow \pi^{ndt}(S^{(t)}, M^{(t)}, E^{(t)})$. [5] Despite affecting the physician's decision, the values of $M^{(t)}, E^{(t)}$ are not recorded in the hospital's database. [6] In other words, the agent's "natural" decision (i.e., without any external intervention) $X^{(t)}$ is reached taking as input variables $S^{(t)}, M^{(t)}, E^{(t)}$, but only the decision ($X^{(t)}$) and the state ($S^{(t)}$) are recorded. The graphical representation of this process is depicted in Fig. 1(a). We can see that the causal relation between $X^{(t)}$ and $Y^{(t)}$ (i.e., arrow from $X^{(t)}$ to $Y^{(t)}$) is confounded (Pearl, 2000, Ch. 6) by the UCs $M^{(t)}$ and $E^{(t)}$.

As an AI researcher, we decide to run a battery of experiments using well-known MDP algorithms (e.g., Delayed-Q-Learning, SARSA, MORMAX), which graphically amounts to replacing the function π that selects the action at a given step. We include two baseline policies for comparison: 1. a policy that follows the physician's decision-making process described above, which we call *ndt* (for "natural decision theory"), and 2. a policy where treatment is picked at random, which we call *random*. Fig. 2 shows the cumulative reward and average reward per episode of these experiments. Somewhat surprisingly, we realize that none of the algorithms is able to learn a reasonable policy – the results coincide with

[5] The full parametrization of this structural model with the reward function $P(Y^{(t)} = 1 \mid S^{(t)}, M^{(t)}, E^{(t)}, X^{(t)})$ and transition function $P(S^{(t+1)} = 0 \mid S^{(t)}, X^{(t)})$ is provided in Appendix 1.

[6] It is usual the case that an intricate combination of factors lead to a decision in settings involving humans. In practice, however, these factors are not always fully known or easily articulable by the decision maker herself, which engender the MDPUC problem.

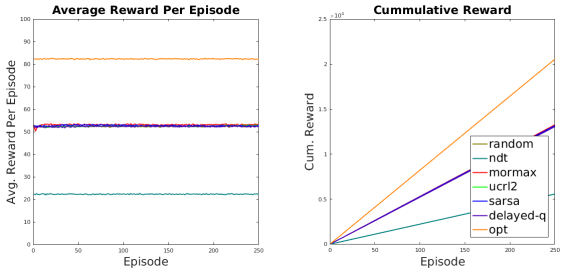

Figure 2: Average reward per episode plot and cumulative reward plot (right) for patient treatment example.

the random policy. Moreover, the *ndt* policy performs worst than all other policies, i.e., the physician is incapable of outperforming random guess. [7]

The experimental results suggest that algorithms that employ standard randomization are unable to converge to some acceptable policy, which raises the question of whether it is feasible in these scenarios to achieve a better performance than random guessing. It is worth highlighting at this point the relationship between MDPUCs and other MDP settings:

1. **Compared to MDPs.** If all variables affecting the agent's decisions and rewards (confounders) are recorded (in our example, $M^{(t)}$ and $E^{(t)}$), it would be possible to control for all the biases and the problem would reduce to standard MDP learning. In practice, however, admitting the existence of UCs is the most relaxed scenario since, alternatively, the modeler would need to know a priori *all* factors that make up the agent's decision, which is a strong requirement in many settings.

2. **Compared to POMDPs.** POMDPs are MDPs with partial or no information of the state variables, where the partial observation does not summarize all the trajectories that led to the present state (i.e., non-Markovian). MDPUCs are MDPs with local unobserved variables that confound the relationships between actions $X^{(t)}$, effects $Y^{(t)}$, and states $S^{(t+1)}$. There are two key differences between POMDPs and MDPUCs. First, POMDPs do not necessarily imply the confounding – i.e., it is possible that no knowledge about the state variable is available, but there exists still no unobserved confounding in the system. Second, the Markovian property holds in MDPUCs (Lemma 1), while it is clearly violated in POMDPs. In fact, POMDPs and MDPUCs are complementary and cover orthogonal dimensions of the modeling space.

We will discuss throughout the paper different scenarios involving UCs. For instance, in the context of a social intervention (e.g., job training program) that we describe in the appendix (Heckman, 1992), there exist no UCs between $X^{(t)}$ and $Y^{(t)}$, but between $X^{(t)}$ and state $S^{(t)}$ (Fig. 1(d)). In fact, we will focus on the general scenario shown in Fig. 1(c), where there are UCs among the three types of variables $(X^{(t)}, Y^{(t)}, S^{(t)})$. Standard algorithms are guaranteed to perform optimally and cannot be improved in settings where UCs do not exist, which are summarized by the model in Fig. 1(b).

MDPUC as a Causal Inference Problem

In this section, we study two classes of policies – the first is called F_{exp} and encompasses policies that decide actions based on the state information and the experimental distribution;

[7] Examples of this kind are the very reason the FDA requires the execution of rigorous clinical trials where the treatment allocation is randomized so as causal effects can be computed without medical biases (due to UCs). In this example, however, both the FDA (random policy) and more sophisticated, adaptive MDP strategies perform no better than chance.

the second is called F_{ctf} and encompasses polices that decide actions based on both the state information and the counterfactual distribution. We compare the performance of polices from both classes and show that whenever UCs are present, agents should search for an optimal policy within F_{ctf} instead of F_{exp}.

In order to properly account for UCs and compare the performance of polices in F_{exp} and F_{ctf}, we will represent MDPUCs using causal formalism and re-express some important RL notions (e.g., value functions and state-action values) in its language. We start by defining MDPUCs and necessary toolkits for analyzing F_{exp}.

Definition 4. A Markov Decision Process with Unobserved Confounders (MDPUCs) is an augmented SCM M (Def. 2) with finite action domain X, state domain S, and binary reward Y:

1. $\gamma \in [0, 1)$ is the discount factor.
2. $U^{(t)} \in U$ is the exogenous variable (i.e., unobserved confounder) at round t.
3. $V^{(t)} = X^{(t)} \cup Y^{(t)} \cup S^{(t)}$ is the set of endogenous (observed) variables at round t, where $X^{(t)} \in X$, $Y^{(t)} \in Y$, and $S^{(t)} \in S$.
4. $F = \{f_x, f_y, f_s\}$ is the set of structural equations relative to V such that $X^{(t)} \leftarrow f_x(s^{(t)}, u^{(t)})$, $Y^{(t)} \leftarrow f_y(x^{(t)}, s^{(t)}, u^{(t)})$, and $S^{(t)} \leftarrow f_s(x^{(t-1)}, s^{(t-1)}, u^{(t-1)})$.
5. $P(u)$ encodes the probability distribution over the exogenous variables U.

Definition 5. Let F_{exp} denote a set of functions between the current state $s^{(t)}$ and the action $x^{(t)}$. Formally, F_{exp} is defined as $F_{exp} = \{\pi \mid \pi : S \to X\}$.

For a standard MDP model taking as input the current state, Filar and Vrieze (2012) showed that an optimal policy π^* must be contained in F_{exp}. Generally, MDP algorithms compute the optimal policy by learning *how good* it is for an agent to be in a certain state or perform a certain action, which is encoded through the value and state-action value functions (Van Otterlo and Wiering, 2012). In order to understand optimality in the presence of UCs, we re-write these functions in terms of MDPUCs.

Definition 6. Given a MDPUC model $M\langle\gamma, U, X, Y, S, F, P(u)\rangle$, an arbitrary deterministic policy π, the value function starting from state $s^{(t)}$ and thereafter following policy π is defined as:

$$V^{\pi}(s^{(t)}) = \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t)}\Big] \tag{1}$$

The state-action value function starting from state $s^{(t)}$, taking action $x^{(t)}$, and thereafter following policy π is defined as:

$$Q^{\pi}(s^{(t)}, x^{(t)}) = \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k Y^{(t+k)}_{x^{(k)}, x^{([t+1,t+k])}=\pi} \mid s^{(t)}, x^{(t)}\Big] \tag{2}$$

One convenient property of value and state-action value functions is that they can be written recursively, which form the basis for most MDP learning algorithms and are often referred as the Bellman Equation. We derive below the recursive expressions for these functions for a given MDPUC instance.

Theorem 1. *Given a MDPUC model $M\langle\gamma, U, X, Y, S, F, P(u)\rangle$, for any policy $\pi \in F_{exp}$, state $s^{(t)}$, and action $x^{(t)}$, the value function $V^{\pi}(s^{(t)})$ can be recursively written as:*

$$\mathbb{E}\Big[Y^{(t)}_{x^{(t)}} \mid s^{(t)}\Big]\Big|_{x^{(t)}=\pi} + \gamma \sum_{s^{(t+1)} \in S} P\Big(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}\Big)\Big|_{x^{(t)}=\pi} V^{\pi}(s^{(t+1)}) \tag{3}$$

The state-action value $Q^{\pi}(s^{(t)}, x^{(t)})$ can be recursively written as:

$$\mathbb{E}\Big[Y^{(t)}_{x^{(t)}} \mid s^{(t)}\Big] + \gamma \sum_{s^{(t+1)} \in S} P\Big(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}\Big) V^{\pi}(s^{(t+1)}) \tag{4}$$

The crucial step in the proof is to show that the Markovian property holds in MDPUCs.

Lemma 1 (Markovian Property in MDPUCs)**.** *For a MDPUC model $M = \langle \gamma, U, X, Y, S, F, P(u) \rangle$, a policy $\pi \in F_{exp}$ and a starting state $s^{(t)}$, the agent performs actions $do(X^{(t)} = x^{(t)})$ at round t and $do(X^{([t+1,t+k])} = \pi)$ afterwards ($k \in \mathbb{Z}_+$), the following statement holds:*

$$P\left(Y^{t+k}_{x^{(t)}, x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right) = P\left(Y^{t+k}_{x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}\right)$$

Lemma 1 implies that all previous states and actions can be best summarized by the current state. We note that Theorem 1 coincide with Bellman Equation representation for standard MDPs. However, we are not aware of any proof of Theorem 1 with the treatment of UCs. Our analysis depends on SCMs and three axioms of counterfactuals Pearl (2000, Sec. 7.3.1). We invite readers to check Appendix 2 for details.

Counterfactual Policies

We consider in this section more elaborated counterfactual sentences of the form $Y_{X=x}|X = x'$, which can be read as "given that $X = x'$, what would the value of Y be had X been x (contrary to the fact)," where $x \neq x'$. In contrast with the experimental counterfactuals discussed above in the context of F_{exp}, these more involved counterfactual quantities are not in general estimable from data, except for some special conditions (Pearl, 2000, Ch. 9).

Somewhat surprisingly, Bareinboim, Forney, and Pearl (2015) noted in the context of MABs that if the decision flow is interrupted just before the agent executes decision x', and then X is randomized conditioned on x', the counterfactual above can in fact be evaluated. The agent's intuition x' is the decision that it would be taken had the system not been submitted to an intervention, which encodes information about the state of the UCs. In fact, standard randomization procedures wash this information out. To avoid this problem, we consider in the sequel a class of counterfactual policies that are sensible to, and incorporate the notion of intuition.

Definition 7. Let F_{ctf} denote a set of functions between the current state $s^{(t)}$, intuition $x'^{(t)}$, and the action $x^{(t)}$. Formally, F_{ctf} is defined as $F_{ctf} = \{\pi | \pi : S \times X \to X\}$.

In the sequel, we accommodate the notion of intuition $x'^{(t)}$ in the value and state-value functions.

Definition 8. Given an MDPUC instance $M\langle \gamma, U, X, Y, S, F, P(u) \rangle$ and an arbitrary deterministic policy function π, the value function starting from state $s^{(t)}$, intuition $x'^{(t)}$, and thereafter following policy π is defined as:

$$V^{\pi}(s^{(t)}, x'^{(t)}) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t)}, x'^{(t)}] \tag{5}$$

The state-action value starting from state $s^{(t)}$, intuition $x'^{(t)}$, taking action $x^{(t)}$, and thereafter following policy π is defined as:

$$Q^{\pi}(s^{(t)}, x'^{(t)}, x^{(t)}) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k Y^{(t+k)}_{x^{(t)}, x^{([t+1,t+k])}=\pi} \mid s^{(t)}, x'^{(t)}] \tag{6}$$

Since F_{ctf} operates with a richer context than F_{exp}, we can easily prove the following theorem:

Theorem 2. *Given an MDPUC instance $M\langle \gamma, U, X, Y, S, F, P(u) \rangle$, let $\pi^*_{exp} = \arg\max_{\pi \in F_{exp}} V^{\pi}(s^{(t)})$ and $\pi^*_{ctf} = \arg\max_{\pi \in F_{ctf}} V^{\pi}(s^{(t)}, x'^{(t)})$. For any state $s^{(t)}$, the following statement holds:*

$$V^{\pi^*_{exp}}(s^{(t)}) \leq V^{\pi^*_{ctf}}(s^{(t)}) \tag{7}$$

More specifically, the equality does not always hold. If UCs are not present, the equality holds.

Theorem 2 states that whenever UCs are not present, algorithms that are not sensitive to the notion of intuition will perform equally well as the ones that are. [8] More strongly, it provides a guarantee that whenever UCs exist and the intuition is available, one should search in the space of counterfactual policies F_{ctf}. To leverage this fact, we derive recursive expressions for Eqs. 5 and 6 so as to allow a more efficient exploration of this search space.

Theorem 3. *For any policy $\pi \in F_{ctf}$, state $s^{(t)}$, intuition $x'^{(t)}$, and action $x^{(t)}$, the value function $V^{\pi}(s^{(t)}, x'^{(t)})$ can be recursively written as:*

$$\mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right]\Big|_{x^{(t)}=\pi} + \gamma \sum_{s^{(t+1)} \in S} \sum_{x'^{(t+1)} \in X} P\left(s^{(t+1)}_{x^{(t)}}, x'^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right)\Big|_{x^{(t)}=\pi} V^{\pi}(s^{(t+1)}, x'^{(t+1)}) \tag{8}$$

The state-action value function $Q^{\pi}(s^{(t)}, x'^{(t)}, x^{(t)})$ can be recursively written as:

$$\mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right] + \gamma \sum_{s^{(t+1)} \in S} \sum_{x'^{(t+1)} \in X} P\left(s^{(t+1)}_{x^{(t)}}, x'^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right) V^{\pi}(s^{(t+1)}, x'^{(t+1)}) \tag{9}$$

The crucial step of the proof is to show that the Markovian property still holds in counterfactual settings.

Lemma 2 (Counterfactual Markovian Property)**.** *For a MDPUC model $M = \langle \gamma, U, X, Y, S, F, P(u) \rangle$, a policy $\pi \in F_{ctf}$, a starting state $s^{(t)}$, and an intuition $x'^{(t)}$, the agent performs actions $do(X^{(t)} = x^{(t)})$ at round t and $do(X^{[t+1,t+k]} = \pi)$ afterwards ($k \in \mathbb{Z}_+$), the following statement holds:*

$$P\left(Y^{t+k}_{x^{(t)}, x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, x'^{(t+1)}_{x^{(t)}}, s^{(t)}, x'^{(t)}\right) = P\left(Y^{t+k}_{x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}, x'^{(t+1)}\right)$$

Proofs are provided in Appendix 2. Lemma 2 implies that all previous states, actions and intuitions can be best summarized by the current state and intuition. We note an interesting feature that follows from Theorem 3. The agent's intuition $x'^{(t)}$ and the state $s^{(t)}$ have the same syntactic form in the recursive expressions of the value and action-value functions (Eqs. 8 and 9). In practice, therefore, as long as the counterfactual quantities $\mathbb{E}\big[Y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\big]$ and $P\big(s^{(t+1)}_{x^{(t)}}, x'^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\big)$ can be empirically evaluated, we can leverage state-of-art MDP algorithms to learn an optimal counterfactual policy in F_{ctf} by simply operating on the augmented state $s'^{(t)}$ encompassing both state and intuition (i.e., $s'^{(t)} = (s^{(t)}, x'^{(t)})$).

Applications and Experiments

Our goal in this section is to operationalize an intent-specific randomization strategy based on Theorem 3 in MDP online learning scenarios.

We take MORMAX (Szita and Szepesvári, 2010), a state-of-art MDP online learning algorithm, as an example and apply the state augmentation which goes as follows:

1. Given an MDPUC instance $M\langle \gamma, U, X, Y, S, F, P(u) \rangle$, we first translate M to a standard MDP instance $M'\langle S', X', T, R \rangle$, where the augmented state $S' = S \times X$, $X' = X$, $T = P(s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)})$, and $R = P(y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)})$.

2. We populate the transition function and reward function table with observational samples, namely [9]: $P(s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x^{(t)}) = P(s^{(t+1)}, x^{(t+1)} \mid s^{(t)}, x^{(t)})$ and $P(y^{((t)}_{x^{(t)}} \mid s^{(t)}, x^{(t)}) = P(y^{(t)} \mid s^{(t)}, x^{(t)})$. Mark the prepopulated entries as known.

[8] It is immediate to see that whenever UCs do not exist, the counterfactual distributions reduce to their experimental counterparts following the independence $Y_X \perp\!\!\!\perp X'$ (Pearl, 2000, Ch. 7).

[9] We follow the seeding rationale introduced in (Bareinboim, Forney, and Pearl, 2015) where constraints across distributions are exploited whenever intuition and decision agree, which follow from the consistency property (Pearl, 2000, pp. 229).

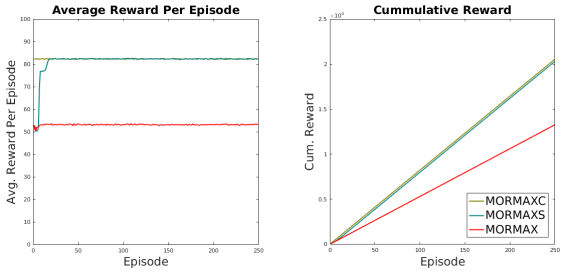

Figure 3: Simulation results for Experiment 1 comparing MORMAX (search within F_{exp}), MORMAXS (search within F_{ctf}), and MORMAXC (search within F_{ctf} while leveraging observational data through seeding).

3. Finally, we call MORMAX with the augmented MDP instance M' and transferred samples.

Candidate Algorithms. We compare three variants of the MORMAX algorithm: vanilla (called MORMAX), MORMAX modified with steps 1 and 3 to consider counterfactual policies (MORMAXS), and MORMAX searching within the counterfactual space and seeding, which include all steps described above (MORMAXC).

Evaluation Metrics. The performance is evaluated with standard metrics: (1) the cumulative reward per episode averaging over 800 runs, and (2) the cumulative reward for 250 episodes. Our metrics compare algorithms' policies to the optimal policy computed by standard MDP planning algorithms (value iteration and policy iteration) assuming all confounders are available, though these variables are not directly available to the agent. We believe this is fair for our examples since it allows the comparison of our algorithm against a truly optimal policy with full access to the UCs.

We performed experiments across multiple parametrizations and settings for both the medical treatment and the job training program (Appendix 1). We provide next one of such parametrizations.

Experiment 1: "Incapable Doctor." This parametrization represents the scenario where the transition probability function is the same under observational and experimental conditions, but the reward function is confounded by $M^{(t)}$ and $E^{(t)}$. As shown in Fig. 2, the natural inclination of the physician led to a policy worse than random guessing while standard MDP algorithms' performed similarly to the randomized policy (i.e., picking the treatment at random at each round).

Furthermore, the results shown in Fig. 3 support the causal approach. Specifically, the simulation reveals an improvement in cumulative reward obtained by MORMAXC (2.0574×10^4) compared to MORMAXS (2.0335×10^4). We can also see from the average reward per episode graph that MORMAXC shows a faster convergence rate than MORMAXS, which corroborates with the view that the sample transferring procedure and leveraging observational data can be helpful. Surprisingly, MORMAXC is able to converge to an optimal policy in the very beginning. The standard MORMAX, predictably, is not a competitor and experiences a relatively low cumulative reward (1.3278×10^4).

Overall, these results confirm that algorithms with the augmented state, which search for the optimal counterfactual policy, converge to a higher expected return; the samples transferring procedure allows algorithms to converge at a faster pace. These conclusions are not unique to this specific setting, but also replicate across a wide range of parametrizations (Appendix 1, Supplementary Material).

Conclusion

We studied the problem of finding optimal policies for MDPs when unobserved confounders are present (MDPUC). We showed that MDPUCs can be found in practical settings and represent a natural formulation for decision problems when unobserved confounders (UCs) exist. Using causal semantics, we acknowledged the existence of two classes of policies – experimental (F_{exp}) and counterfactual (F_{ctf}). We then showed that agents should search for an optimal policy within the counterfactual class (F_{ctf}) instead of the experimental one (F_{exp}) whenever UCs are present (Thm. 2). Through a syntactic transformation of the state variable allowed by Thm. 3, we operationalized this search strategy (Alg. 1) and showed that it improves state-of-the-art algorithms both in terms of speed and convergence.

References

Bareinboim, E., and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* 113:7345–7352.

Bareinboim, E.; Forney, A.; and Pearl, J. 2015. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, 1342–1350.

Bertsekas, D. P., and Tsitsiklis, J. N. 1995. Neuro-dynamic programming: an overview. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1, 560–564. IEEE.

Bubeck, S., and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* 5:1–122.

Dietterich, T. G. 1998. The maxq method for hierarchical reinforcement learning. In *International Conference on Machine Learning (ICML)*, 118–126.

Ellis, H.; Jiang, M.; and Corotis, R. B. 1995. Inspection, maintenance, and repair with partial observability. *Journal of Infrastructure Systems* 1(2):92–99.

Filar, J., and Vrieze, K. 2012. *Competitive Markov decision processes*. Springer Science & Business Media.

Fisher, R. 1951. *The Design of Experiments*. Edinburgh: Oliver and Boyd, 6th edition.

Heckman, J. 1992. Randomization and social policy evaluation. In Manski, C., and Garfinkle, I., eds., *Evaluations: Welfare and Training Programs*. Cambridge, MA: Harvard University Press. 201–230.

Li, L.; Chu, W.; Langford, J.; and Wang, X. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, 297–306.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press. 2nd edition, 2009.

Puterman, M. L. 1994. Markov decision processes: Discrete stochastic dynamic programming.

Simpson, E. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* 13:238–241.

Singh, S. P.; Jaakkola, T. S.; and Jordan, M. I. 1994. Learning without state-estimation in partially observable markovian decision processes. In *ICML*, 284–292.

Smallwood, R. D., and Sondik, E. J. 1973. The optimal control of partially observable markov processes over a finite horizon. *Operations Research* 21(5):1071–1088.

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

Szepesvári, C. 2010. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Szita, I., and Szepesvári, C. 2010. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 1031–1038.

Van Otterlo, M., and Wiering, M. 2012. Reinforcement learning and markov decision processes. In *Reinforcement Learning*. Springer. 3–42.

Van Otterlo, M. 2009. *The logic of adaptive behavior*. Ios Press.

Willems, J. C. 1989. *Models for Dynamics*. Wiesbaden: Vieweg+Teubner Verlag. 171–269.

Appendix A. Experimental Results and Applications

In this section, we discuss the performance of the algorithms and provide the full parametrizations of the medical treatment discussed in Sec. 2 of the paper and the job training program (Heckman, 1992). We also discuss simulations across a wide range of parametrizations for the general case when both the transition and reward probabilities are confounded.

Medical Treatment

Recall that the patient's health score is affected by the patient's mood $M^{(t)} \in \{0, 1\}$ (0 for positive, 1 for negative) and socioeconomic status (SES) $E^{(t)} \in \{0, 1\}$ (0 for wealthy, 1 for poor) at each time step. Further, the patient has an equal chance of having bad/good mood and financial difficulties, i.e., $P(M^{(t)} = 0) = \frac{1}{2}$, $P(E^{(t)} = 0) = \frac{1}{2}$. The physician's own policy is defined as $X^{(t)} \leftarrow \pi^{ndt}(S^{(t)}, M^{(t)}, E^{(t)}) = S^{(t)} \oplus M^{(t)} \oplus E^{(t)}$, where $\oplus$ represents the exclusive OR operator.

The reward probability function $P(Y^{(t)} \mid S^{(t)}, M^{(t)}, E^{(t)}, X^{(t)})$ and the transition probability function $P(S^{(t+1)} \mid S^{(t)}, X^{(t)})$ are provided in Tables 1 and 2. The entries encode the probabilities for $Y^{(t)} = 1$. The doctor's natural choice of action (i.e., following π^{ndt}) are indicated by asterisks.

	$S^{(t)}=0$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	$*0.2$	0.9	0.8	$*0.3$
$X^{(t)}=1$	0.9	$*0.2$	$*0.3$	0.8
	$S^{(t)}=1$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	0.7	$*0.2$	$*0.1$	0.8
$X^{(t)}=1$	$*0.2$	0.7	0.8	$*0.1$

Table 1: Reward probability table for health score $Y^{(t)} = 1$, which is $P(Y^{(t)} = 1 \mid S^{(t)}, M^{(t)}, E^{(t)}, X^{(t)})$. The doctor's natural choice under $S^{(t)}, M^{(t)}, E^{(t)}$ are indicated by asterisks.

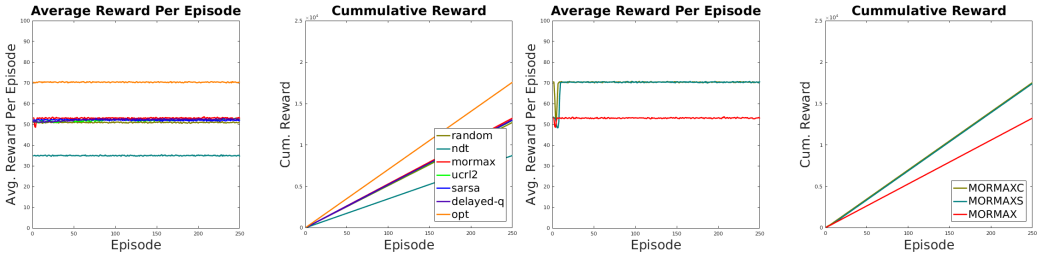

Figure 4: (a, b) Average reward per episode plot and cumulative reward plot (right) for patient treatment example. (c, d) Simulation results comparing standard Mormax (Mormax), S-Empowered Mormax (MormaxS), and CausalMormax (MormaxC)

	$S^{(t)}=0$	$S^{(t)}=1$
$X^{(t)}=0$	0.9	0.3
$X^{(t)}=1$	0.7	0.8

Table 2: The transition probability table $P(S^{(t+1)}=0 \mid S^{(t)}, X^{(t)})$.

Social training program

We describe next a training program that is a prototypical social intervention (Heckman, 1992). There is a government sponsored training center offering job training where the attendant is instructed to determine whether the salary of a given applicant is below or above a certain threshold (named status $S^{(t)}$), and then decide whether to accept or reject this applicant ($X^{(t)}$). The salary status changes periodically and is measured at each time she/he visits the center – the same ($S^{(t+1)}$) is known to be affected by the previous status $S^{(t)}$ and decision $X^{(t)}$. To evaluate the performance of the training center, the government estimates an overall score $Y^{(t)}$, which is a function of the center previous decisions ($X^{(t)}$) and the applicants' salary status ($S^{(t)}$). The goal is to maximize the center's cumulative score over a long period. Clearly, this setting can be modeled as an MDP problem where the optimal criterion is to maximize the cumulative discounted reward.

In reality, however, this is just part of the story since the applicant's salary status is also affected by certain unobserved confounders (UCs). For instance, her health level $M^{(t)}$ and family's socioeconomic condition $E^{(t)}$ also affects $S^{(t)}$. The center's attendant tries to assess this information to make a "more informed decision" – i.e., action $X^{(t)}$ is a function of all these factors (i.e., $M^{(t)}, E^{(t)}, S^{(t)}$), but does not record the UCs in the center's database. The main difference between this setting (Fig. 1(d), paper) and the medical treatment (Fig. 1(a), paper) is that the edge connecting $X^{(t)}$ and $S^{(t+1)}$ are confounded by the unobserved variables $M^{(t)}$ and $E^{(t)}$ (instead of $X^{(t)}$ and $Y^{(t)}$), so the estimation of the transition probability requires a more refined treatment.

	$S^{(t)}=0$	$S^{(t)}=1$
$X^{(t)}=0$	0.3	0.9
$X^{(t)}=1$	0.6	0.7

Table 3: The reward probability table for government score is $P(Y^{(t)}=1|S^{(t)}, X^{(t)})$.

The reward probability and transition probability functions are given in Tables 3 and 4. We first perform experiments based on standard MDP solvers and the results are shown in Fig. 4(a,b). Similarly to the medical example, the optimal policy *opt* can be computed using standard MDP algorithms when $M^{(t)}$ and $E^{(t)}$ are observed, which we will use as a baseline for comparison. The experimental results indicate that the attendant's policy

(i.e., ndt) is worse than randomized-based strategies; also, they show that standard MDP algorithms perform no better than a purely randomized policies in F_{exp}. We then run the empowered MORMAX strategy as described in the paper and obtain the results shown in Fig. 4(c,d). The results support the superiority of intent-based randomization strategy – MORMAXC converges faster than MORMAXS while pure MORMAX does not converge.

	$S^{(t)}=0$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$D^{(t)}=0$	$D^{(t)}=1$	$D^{(t)}=0$	$D^{(t)}=1$
$X^{(t)}=0$	0.7	$*0.2$	$*0.4$	0.8
$X^{(t)}=1$	$*0.2$	0.7	0.8	$*0.4$
	$S^{(t)}=1$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$D^{(t)}=0$	$D^{(t)}=1$	$D^{(t)}=0$	$D^{(t)}=1$
$X^{(t)}=0$	$*0.3$	0.6	0.9	$*0.2$
$X^{(t)}=1$	0.6	$*0.3$	$*0.2$	0.9

Table 4: Transition probabilities $P(S^{(t+1)}=0 \mid S^{(t)}, M^{(t)}, E^{(t)}, X^{(t)})$.

General Experiments

In this section, we perform experiments across five parametrizations describing qualitatively different relationships between observational (ndt), experimental (F_{exp}), and counterfactual (F_{ctf}) distributions. Our simulations consider the general setting where transition probability function and reward function are both confounded (Fig. 1(c), paper). Overall, the experimental results confirm that counterfactual strategies that belongs in F_{ctf} perform better than experimental ones (from F_{exp}). The experimental procedure follows the same guideline as described in the paper. The transition probability distribution remains the same for all parametrizations and is described in Table 5, where $\{M^{(t)}, E^{(t)}\}$ confound the function between the action $X^{(t)}$ and the state variable $S^{(t)}$.

	$S^{(t)}=0$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	$*0.5$	0.9	0.9	$*0.4$
$X^{(t)}=1$	0.9	$*0.5$	$*0.4$	0.9
	$S^{(t)}=1$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	0.9	$*0.1$	$*0.2$	0.8
$X^{(t)}=1$	$*0.1$	0.9	0.8	$*0.2$

Table 5: Transition probabilities for $S^{(t+1)}=0$, which is $P(S^{(t+1)}=0|S^{(t)}, M^{(t)}, E^{(t)}, X^{(t)})$. The agent's natural choice of treatment under $S^{(t)}, M^{(t)}, E^{(t)}$ are indicated by asterisks.

We will use the label "S-powered" to denote algorithms that use intuition-based randomization strategy, and "causal" to denote algorithms that use both intuition-based randomization strategy and seeding (transfer) of observational data. We report the experimental results below comparing the following algorithms: standard MORMAX, S-Powered MORMAX (MORMAXS), CausalMORMAX (MORMAXC), standard UCRL2, S-powered UCRL2 (UCRL2S), Causal UCLR2 (UCRL2C), Delayed-Q learning, S-powered Delayed-Q Learning, standard SARSA, and S-powered SARSA (SARSAS).

Experiment 1 ("Incapable Agent"): The reward function is described in Table 6. In this parametrization, the agent's behavior is similar to the physician described in the med-

ical treatment example. However, learning becomes harder because of the UCs between action $X^{(t)}$ and $S^{(t)}$. Simulation results are appended after this section, see Figs. 5 and 6. Intuition powered agents consistently outperform standard MDP algorithms. Agents with transferred observational samples (UCRL2C and MORMAXC) demonstrate faster convergence rate then agents without transferred samples (UCRL2S, MORMAXS, Delayed-QS and SARSAS).

	$S^{(t)}=0$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	$*0.2$	0.9	0.8	$*0.3$
$X^{(t)}=1$	0.9	$*0.2$	$*0.3$	0.8
	$S^{(t)}=1$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	0.7	$*0.2$	$*0.1$	0.8
$X^{(t)}=1$	$*0.2$	0.7	0.8	$*0.1$

Table 6: Incapable Agent

Experiment 2 ("Capable Agent"): The reward function is described in Table 7. In this parametrization, the "natural" policy of the agent (π^{ndt}) is already operating at the optimal level. As expected, standard MDP algorithm wash out the positive intuition of the agent that follows π^{ndt}, so it performs no better than random guessing. The simulation results are shown in Figs. 7 and 8. The results suggests that intuition powered algorithms demonstrate consistent improvements over standard MDP algorithms. Agents with transferred observational samples (UCRL2C and MORMAXC) that leverage the correct intuition demonstrate faster convergence rate then agents without transferred samples (UCRL2S, MORMAXS, Delayed-QS and SARSAS).

	$S^{(t)}=0$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	$*0.9$	0.2	0.3	$*0.8$
$X^{(t)}=1$	0.2	$*0.9$	$*0.8$	0.3
	$S^{(t)}=1$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	0.2	$*0.7$	$*0.6$	0.3
$X^{(t)}=1$	$*0.7$	0.2	0.3	$*0.6$

Table 7: Capable Agent

Experiment 3 ("Paradoxical Switching"): The reward function is shown in Table 8. In this parametrization, the value function estimated based on the observational samples suggests opposite action compared with the reality because of the presence of UCs. If the agent naively transfers observational samples as if they were obtained through randomizations, it could cause significant negative effect on the agent's performance. The simulation result are shown in Figs. 9 and 10, which suggest that intuition powered algorithms demonstrate consistent improvements over standard MDP algorithms. Perhaps surprisingly to some, agents with transferred observational samples (UCRL2C and MORMAXC) demonstrate faster convergence regardless of potential negative transfer effect. This seems to suggest that our transfer strategy is robust against confounding bias.

	$S^{(t)}=0$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	$*0.3$	0.6	0.6	$*0.2$
$X^{(t)}=1$	0.8	$*0.2$	$*0.1$	0.9
	$S^{(t)}=1$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	0.8	$*0.1$	$*0.1$	0.8
$X^{(t)}=1$	$*0.2$	0.6	0.6	$*0.3$

Table 8: Paradoxical Switching

Experiment 4 ("Sometimes Switching"): The reward function is shown in Table 9. In this parametrization, the confounding bias still exists based on the parametrization, but it is irrelevant for the action's choice. The simulation results are shown in Figs. 11 and 12. We can see that intuition-powered algorithms achieve similar performance as standard MDP algorithms since it work as if the UCs are not there. Interestingly, this simulation suggests that our strategy also performs well in simpler settings and can be generally applied when UCs are not known to be present.

	$S^{(t)}=0$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	$*0.6$	0.7	0.7	$*0.4$
$X^{(t)}=1$	0.1	$*0.1$	$*0.2$	0.15
	$S^{(t)}=1$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	0.2	$*0.3$	$*0.2$	0.4
$X^{(t)}=1$	$*0.6$	0.4	0.7	$*0.5$

Table 9: Sometimes Switching

Experiment 5 ("Non-optimal"): The reward function is described in Table 10. In this parametrization, the decision is a function of the state and the UCs, but by construction, the specific numbers imply probabilistic independence. This implies that the intuition does not capture any additional information about the reward and state distributions. The simulation results are shown in Figs. 13 and 14 and demonstrate that this is a challenging parametrization for all MDP algorithms.

	$S^{(t)}=0$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	$*0.3$	0.6	0.6	$*0.7$
$X^{(t)}=1$	0.8	$*0.3$	$*0.3$	0.2
	$S^{(t)}=1$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	0.5	$*0.4$	$*0.6$	0.4
$X^{(t)}=1$	$*0.2$	0.9	0.1	$*0.7$

Table 10: Non-optimal

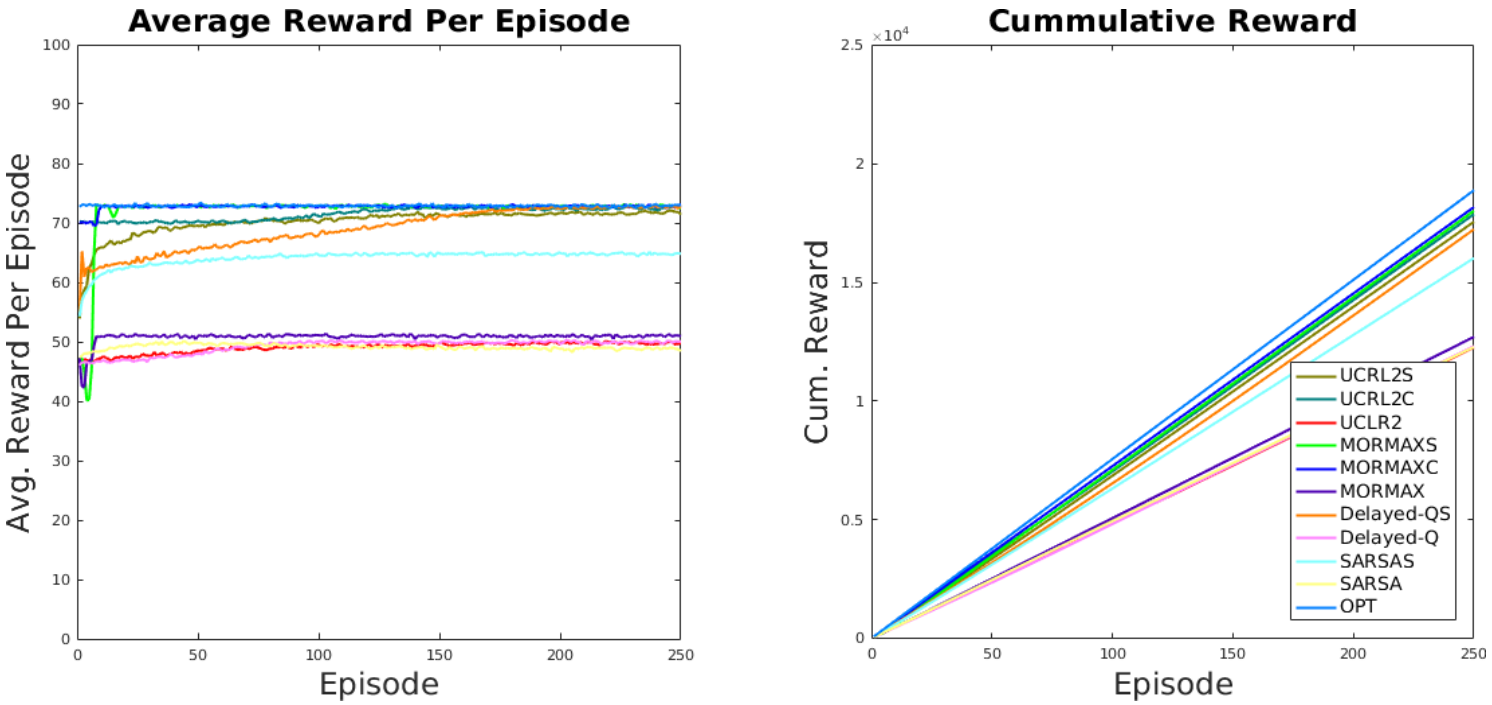

Figure 5: Incapable Agent: Cumulative Reward and Average Return per Episode

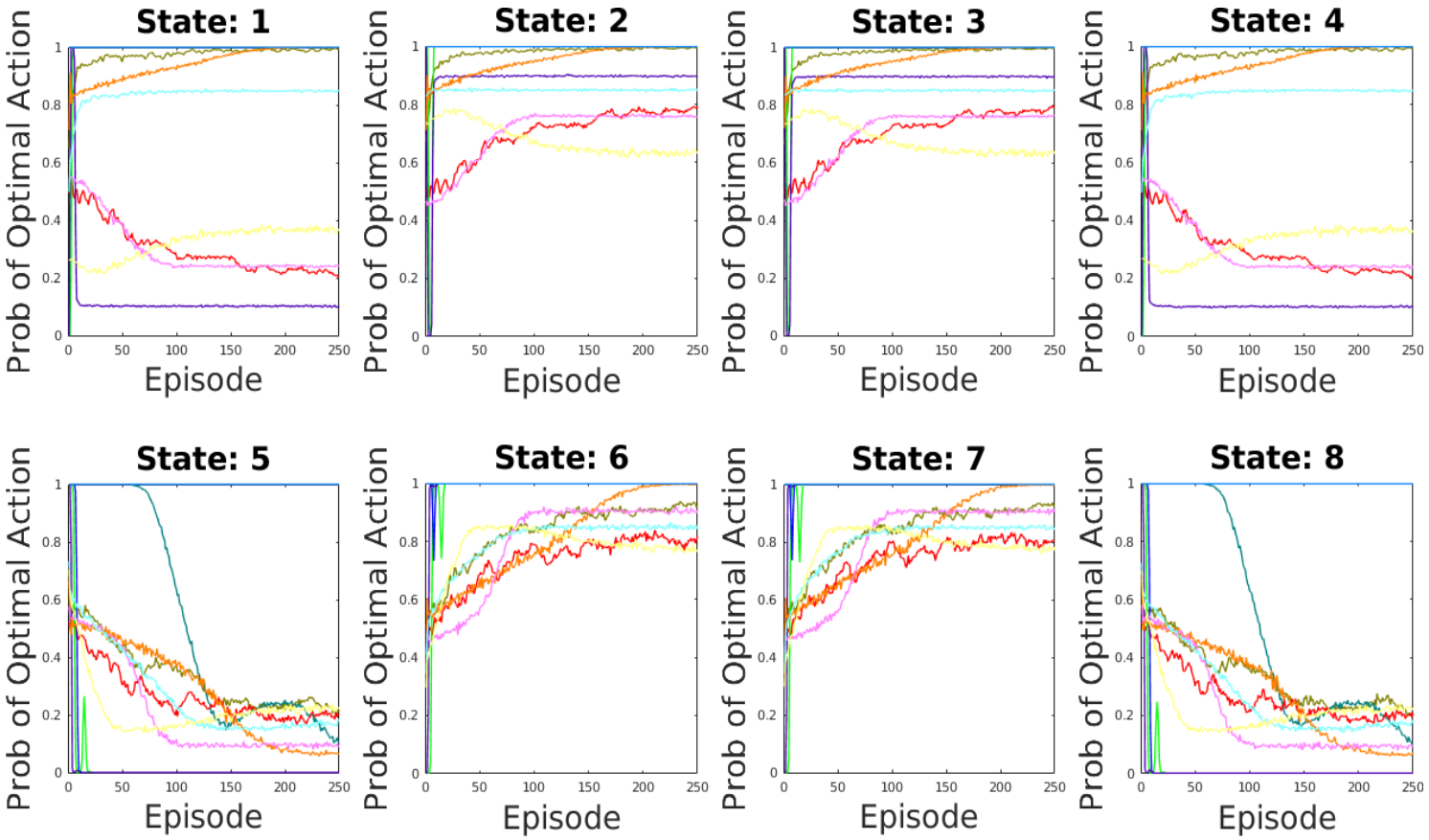

Figure 6: Incapable Agent: Probability of Selecting Optimal Action at given state

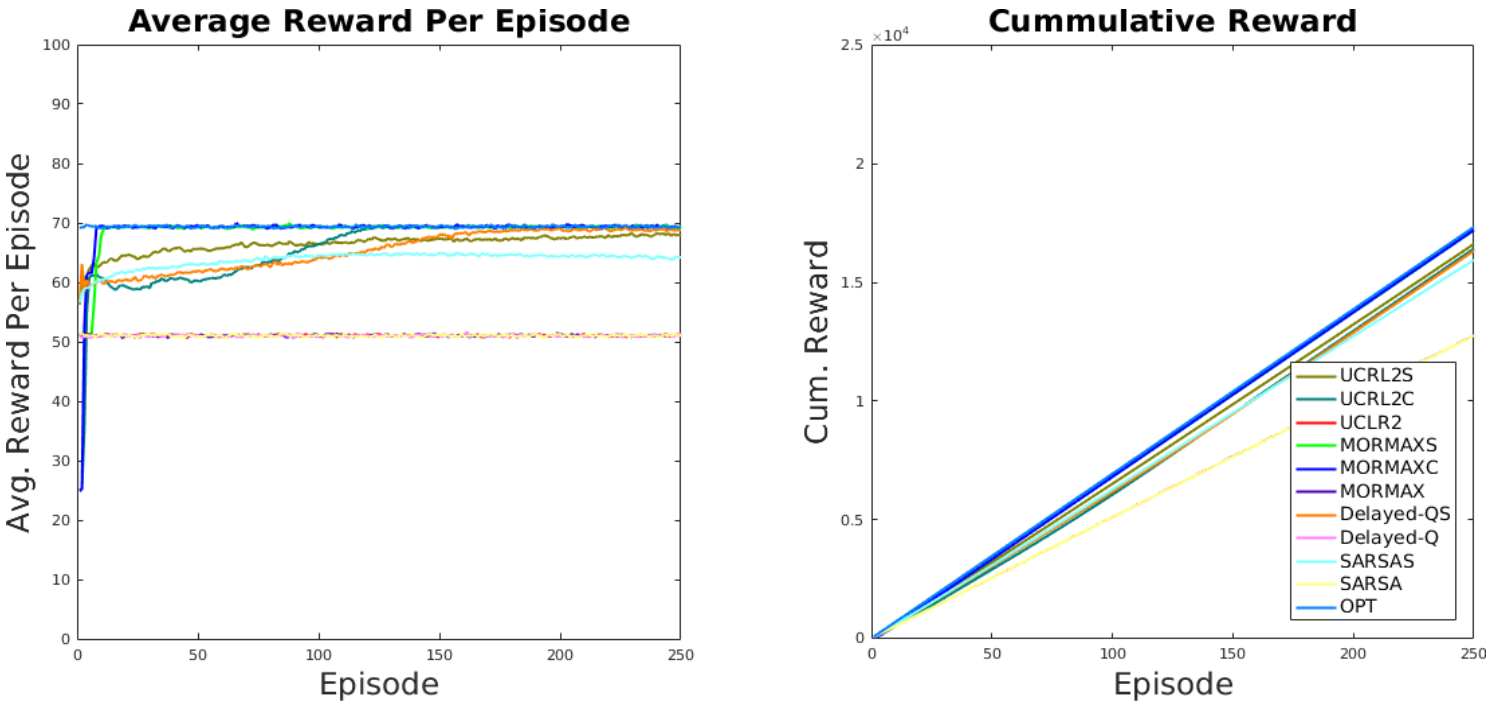

Figure 7: Capable Agent: Cumulative Reward and Average Return per Episode

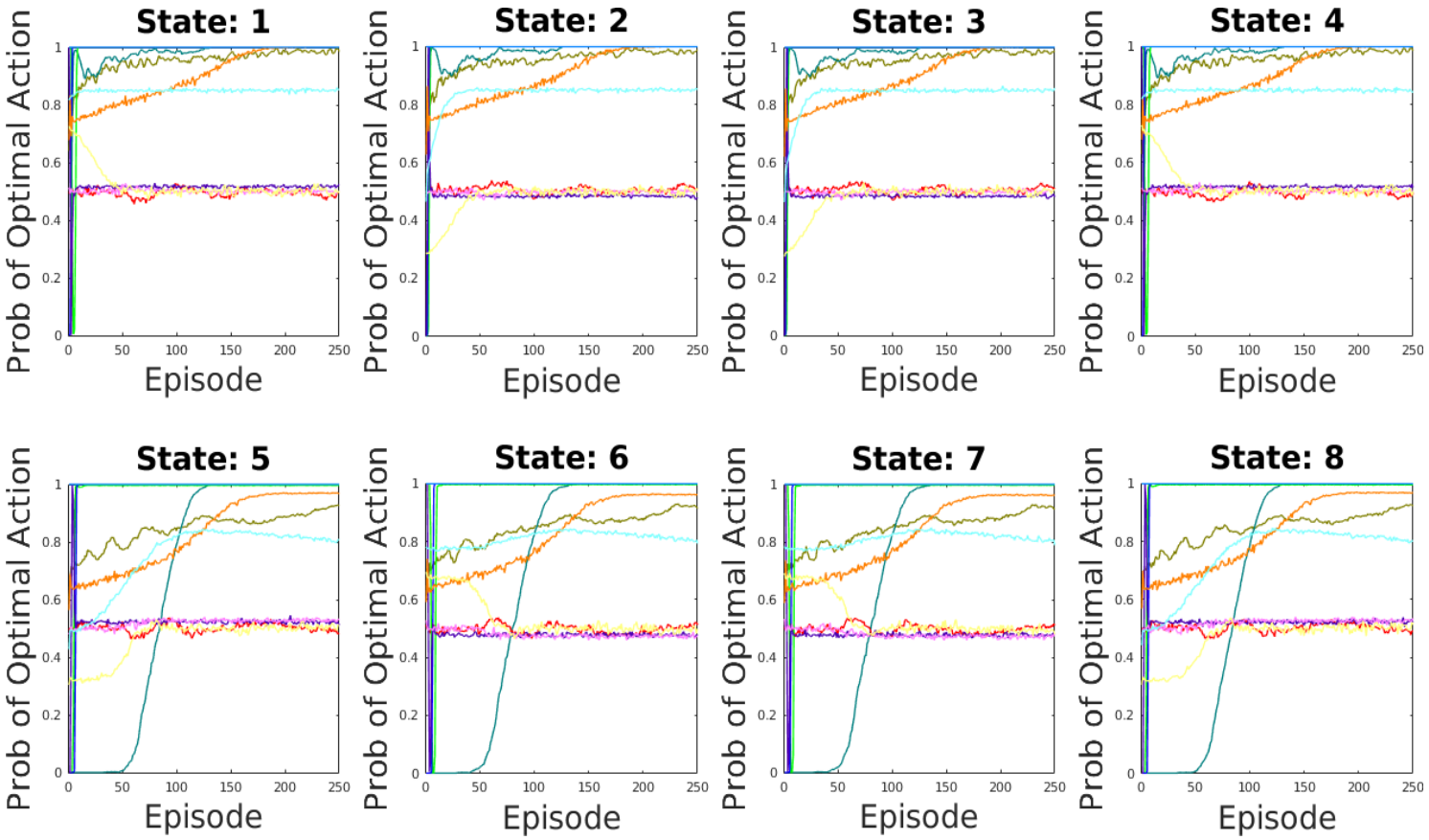

Figure 8: Capable Agent: Probability of Selecting Optimal Action at given state

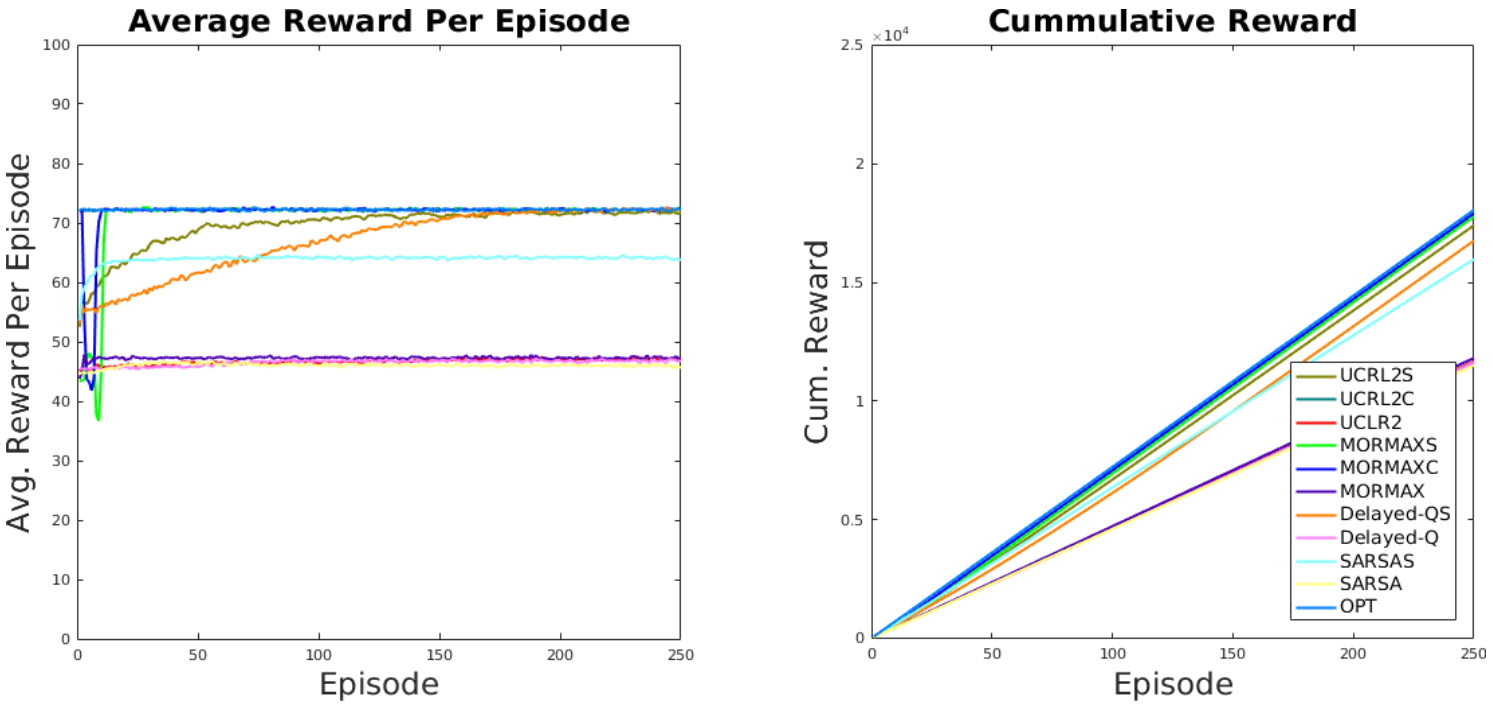

Figure 9: Paradoxical Switching: Cumulative Reward and Average Return per Episode

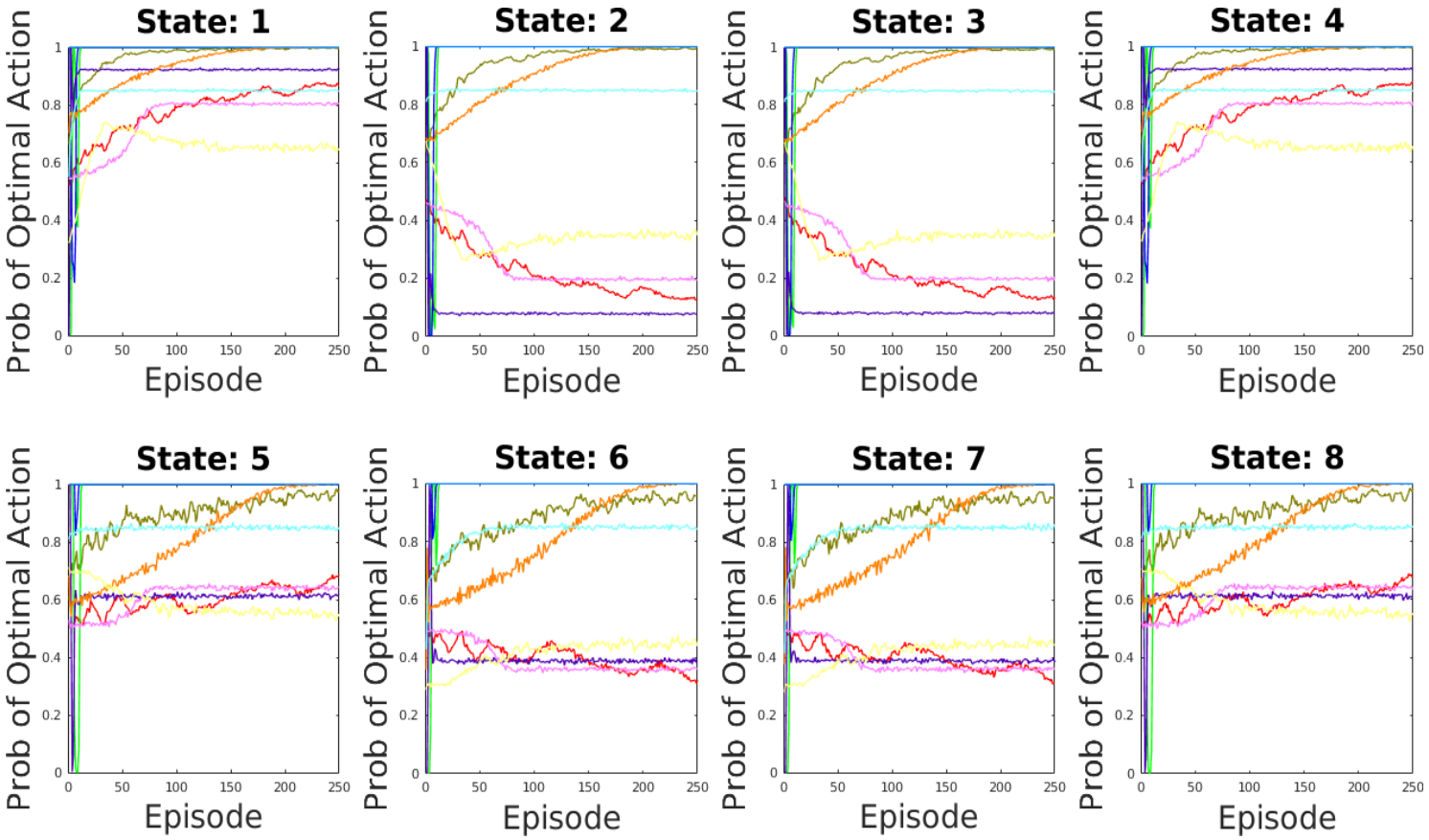

Figure 10: Paradoxical Switching: Probability of Selecting Optimal Action at given state

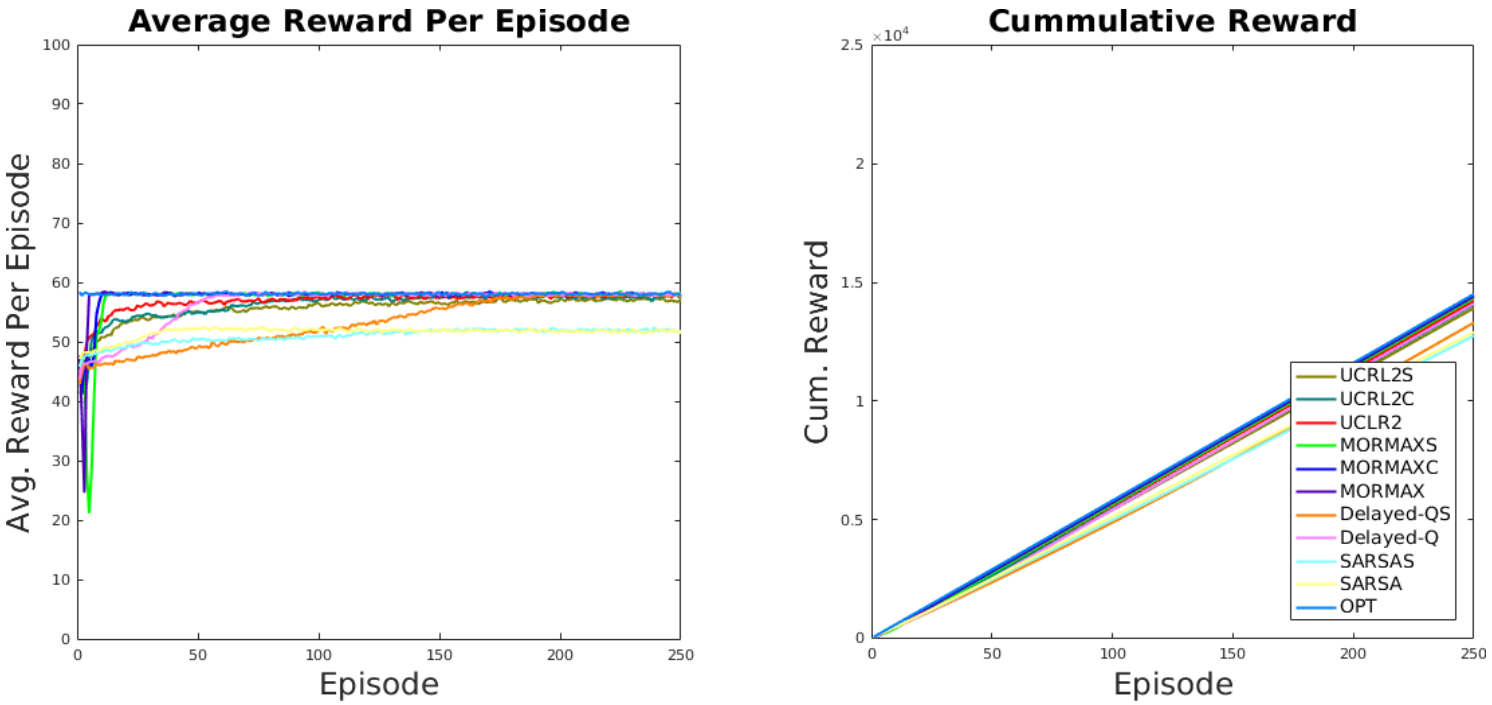

Figure 11: Sometimes Switching: Cumulative Reward and Average Return per Episode

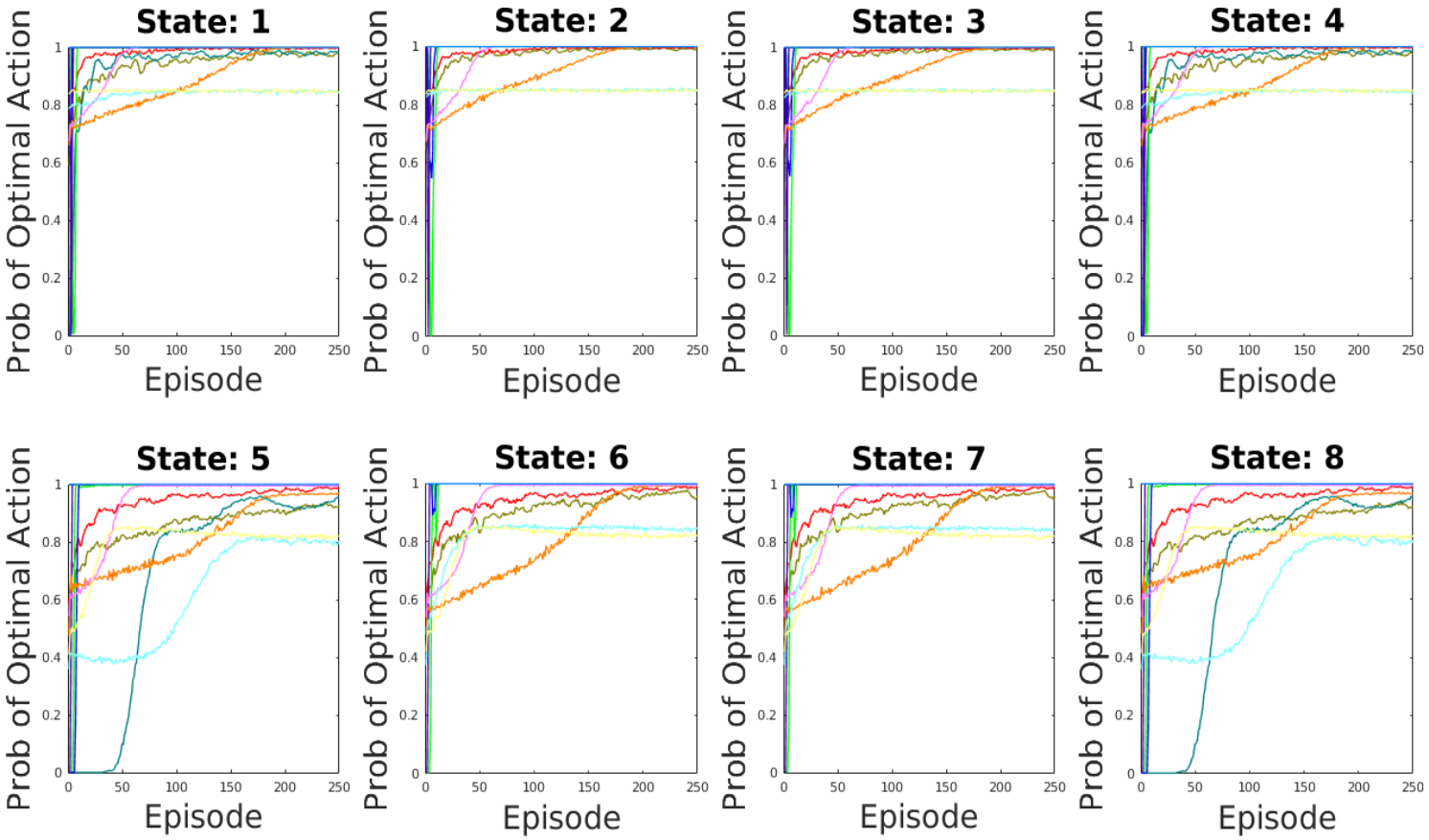

Figure 12: Sometimes Switching: Probability of Selecting Optimal Action at given state

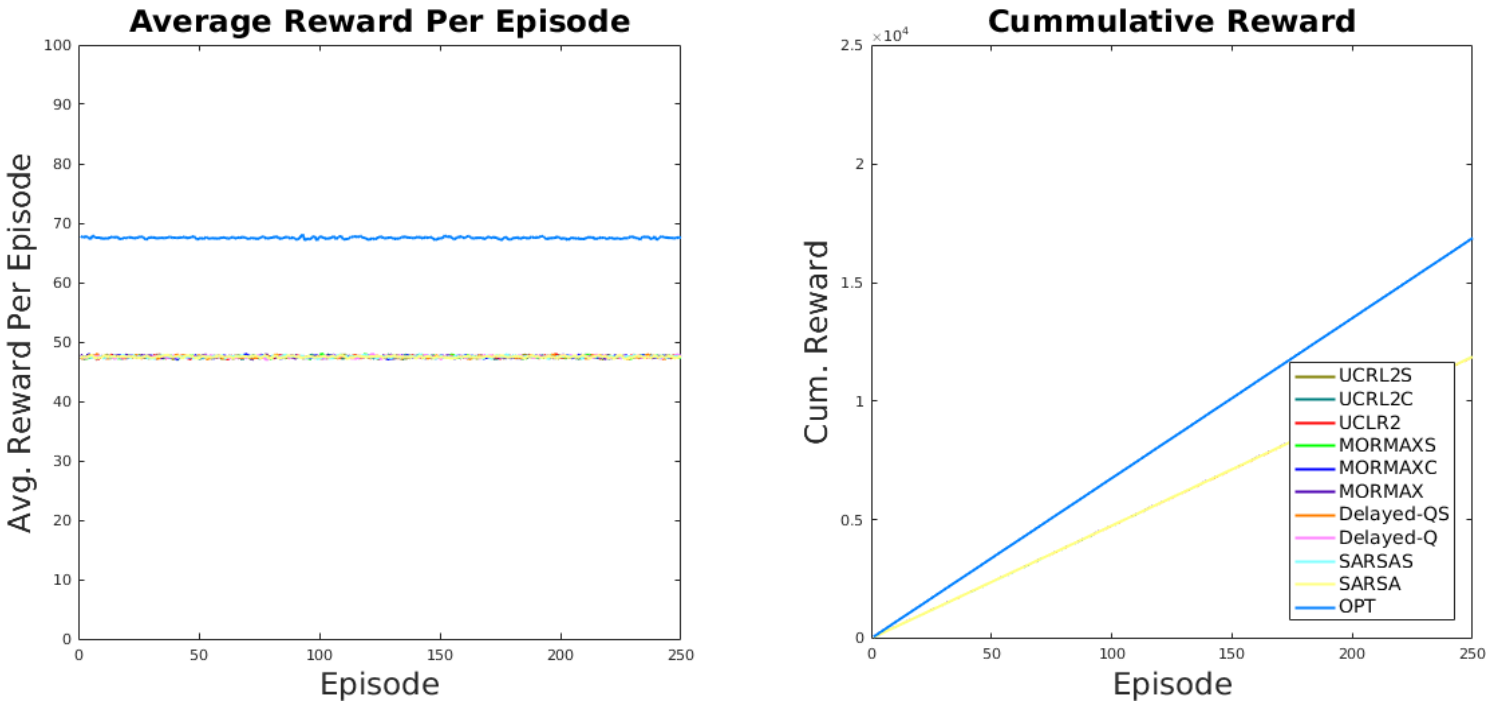

Figure 13: Non-Optimal: Cumulative Reward and Average Return per Episode

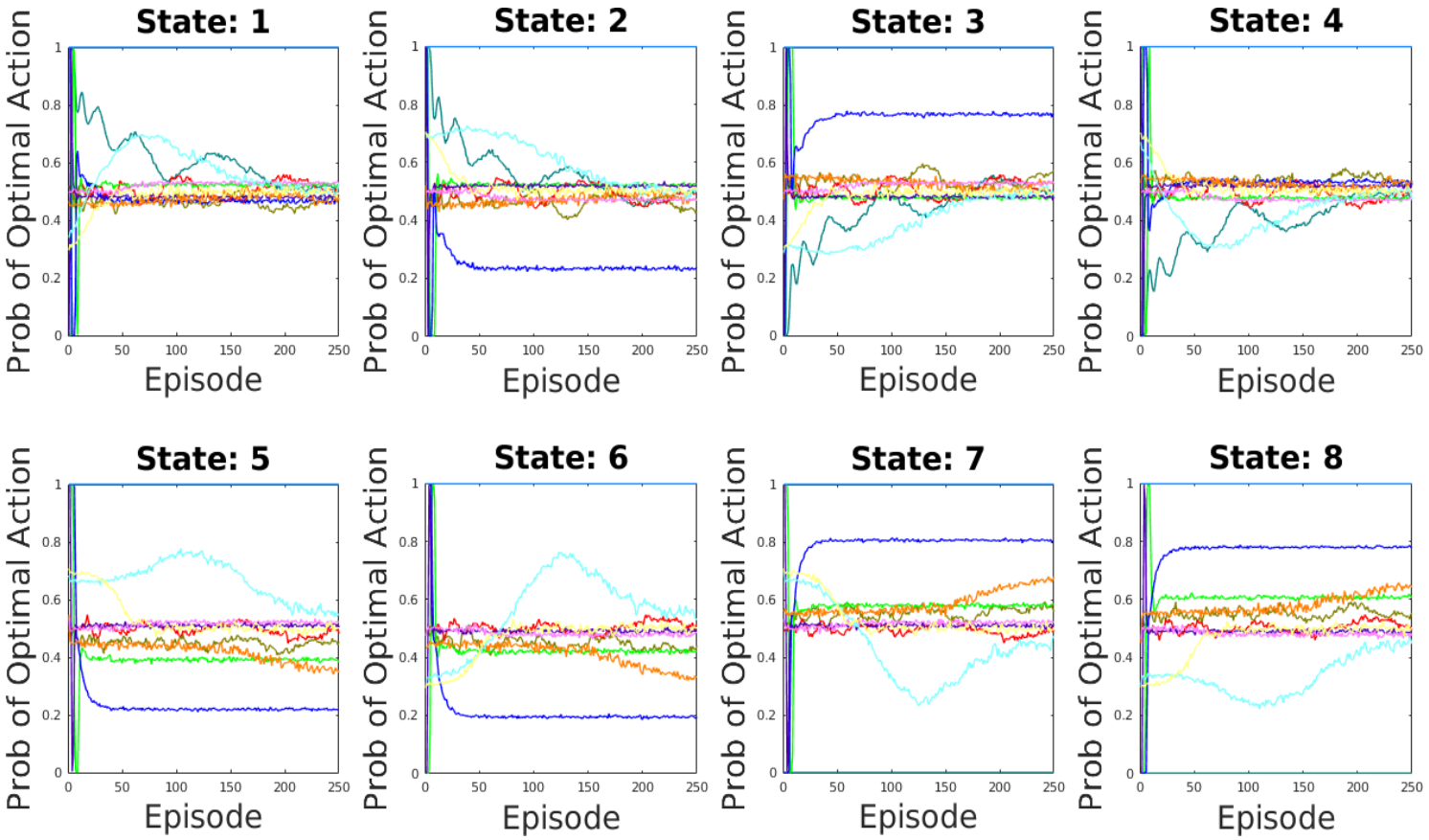

Figure 14: Non-Optimal: Probability of Selecting Optimal Action at given state

Appendix B. Proofs

We start by introducing the notations and theorems used throughout the proofs. Recall that we use $X^{([i,j])}$ to represent the sequence $(X^{(i)}, X^{(i+1)}, \ldots, X^{(j)})$. $X^{([i,j])}$ is an empty sequence if $i > j$.

Lemma 3. *(Conditional Interventions, Pearl (2000, Sec. 4.2)) Let $P\big(y_{x=f(z)}|, z\big)$ denote the effect of $do(X = f(x))$ on a variable Y given $Z = z$. Then, $P\big(y|do(x = f(z)), z\big)$ is equivalent to:*

$$P\big(y_{x=f(z)}|z\big) = P\big(y_x|z\big)\big|_{x=\pi(z)}$$

Proof. Assuming that $f(z)$ is a function that takes an argument z and decides for x. When $Z = z$ is fixed, X is also fixed with the value $x = f(z)$. The effect $P\big(y_{x=f(z)}|z\big)$ is equivalent to the effect of an atomic intervention $P\big(y_x|z\big)$ with fixed value $x = f(z)$. We then have:

$$P\big(y_{x=f(z)}|z\big) = \sum_{x\in X} P\big(y_x|z\big)I\{x = f(z)\} = P\big(y_x|z\big)\big|_{x=f(z)}$$

□

Axiom 1. *(The Axioms of Counterfactuals, Pearl (2000, Sec. 7.3.1)) In all causal models, composition, effectiveness and reversibility properties hold.*
Composition: *For any three sets of endogenous variables X, Y, and W in a causal model, we have:*

$$W_x = w \Rightarrow Y_{xw} = Y_x$$

Effectiveness: *For any all sets of variables X and W,*

$$X_{xw} = x$$

Reversibility: *For any two variables Y and W and any set of variables X,*

$$(Y_{xw} = y)\&(W_{xy} = w) \Rightarrow Y_x = y$$

Composition, effectiveness and reversibility are proved to be sound and complete in all causal models. To translated the assumptions embodied in the graphical model into the language of counterfactuals, we also introduces two rules: exclusion restrictions and independence restrictions.

Lemma 4. *(Exclusion restrictions, Pearl (2000, Sec. 7.3.2)) For every variables Y having parents PA_Y for every set of variables $Z \subset V$, V disjoint of PA_Y, we have:*

$$Y_{pa_Y} = Y_{pa_Y,Z}$$

Lemma 5. *(Independence restrictions, Pearl (2000, Sec. 7.3.2)) If $Z_1, \ldots, Z_k$ is any set of nodes in V not connected to Y via paths containing only U variables, we have:*

$$\left(Y_{pa_Y} \perp\!\!\!\perp Z_{1pa_{Z_1}}, \ldots, Z_{kpa_{Z_k}}\right)$$

We also introduce the concept of C-component (**?**), which will be useful in our later proofs.

Definition 9. (C-component, **?**) Let G be a causal diagram such that a subset of its bidirectional arcs forms a spanning tree over all vertices in G. Then G is a C-component.

We start off by proving a probabilistic decomposition based on C-components which holds in all SCMs. This decomposition implies a more generalized and stronger independence relations suggested by the independence restrictions rule.

Theorem 4. *Given a SCM $M\langle U, V, F, P(u)\rangle$, assume that V is partitioned into k C-components $S^{(1)}, S^{(2)}, \ldots, S^{(k)}$ and denote $N^{(j)}$ the set of U variables that are parents of those variables in $S^{(j)}$. Let v_{pa} denote the sequence $\{v^{(i)}_{pa^{(i)}} | \forall V^{(i)} \in V, \ , V^{(i)} = v^{(i)}\}$ and $s^{(j)}_{pa}$ the sequence $\{v^{(i)}_{pa^{(i)}} | \forall V^{(i)} \in S_j, \ , V^{(i)} = v^{(i)}\}$. $P(v_{pa})$ can be decomposed into a product of $P(s^{(j)}_{pa})$'s:*

$$P(v_{pa}) = \prod_{j=1}^{k} P(s^{(j)}_{pa}) \tag{10}$$

Proof. Recall that $V^{(i)}_{pa^{(i)}}$ encodes the operation that fixes the value $PA^{(i)} = pa^{(i)}$ and decides the value $V^{(i)} = f(pa^{(i)}, N^{(j)})$. The randomness of variable $V^{(i)}$ is fully encoded by the exogenous variable $N^{(j)}$. Thus, $P(s^{(j)}_{pa})$ can be written as follows:

$$P(s^{(j)}_{pa}) = \sum_{\{N^{(j)}=n^{(j)}\}} \prod_{\{V^{(i)} \in S^{(j)}, V^{(i)}=v^{(i)}\}} I\{v^{(i)} = f(pa^{(i)}, n^{(j)})\} P(n^{(j)}) \tag{11}$$

By definition, $P(v_{pa})$ equals to:

$$\begin{aligned}
P(v_{pa}) &= \sum_{\{U=(n^{(1)},\ldots,n^{(k)})\}} \prod_{\{V^{(j)} \in V, V^{(j)}=v^{(j)}\}} I\{v^{(j)} = f(pa^{(j)}, n^{(j)})\} \prod_{j=1}^{k} P(n^{(j)}) \\
&= \underbrace{\sum_{\{N^{(i)}=n^{(j)}\}} \prod_{\{V^{(j)} \in S^{(j)}, V^{(j)}=v^{(j)}\}} I\{v^{(j)} = f(pa^{(j)}, n^{(j)})\} P(n^{(i)})}_{\text{Part 1}} \\
&\cdot \underbrace{\sum_{\{U \setminus N^{(i)} = u \setminus n^{(i)}\}} \prod_{\{V^{(j)} \in V \setminus S^{(j)}, V^{(j)}=v^{(j)}\}} I\{v^{(j)} = f(pa^{(j)}, n^{(j)})\} \prod_{j=1, j\neq i}^{k} P(n^{(j)})}_{\text{Part 2}}
\end{aligned}$$

where $u \setminus n^{(i)} = (n^{(1)}, \ldots, n^{(i-1)}, n^{(i+1)}, \ldots, n^{(k)})$. The last step holds because for variables in $S^{(j)}$, they are only affected by $N^{(j)}$, we can move $S^{(j)}$ and $N^{(j)}$ outside the summation of $U \setminus N^{(i)}$. Note that part 1 is exactly $P(s^{(i)}_{pa})$ defined in Equation 11. We have that:

$$P(v_{pa}) = P(s^{(i)}_{pa}) \underbrace{\sum_{\{U \setminus N^{(i)} = u \setminus n^{(i)}\}} \prod_{\{V^{(j)} \in V \setminus S^{(j)}, V^{(j)}=v^{(j)}\}} I\{v^{(j)} = f(pa^{(j)}, n^{(j)})\} \prod_{j=1, j\neq i}^{k} P(n^{(j)})}_{\text{Part 2}}$$

By recursively applying the above process on the part 2 for remaining $k-1$ C-components $S^{(1)}, S^{(2)}, \ldots, S^{(i-1)}, S^{(i+1)}, \ldots, S^{(k)}$, we have that:

$$P(v_{pa}) = \prod_{j=1}^{k} P(s^{(j)}_{pa})$$

which proves the statement. □

Lemma 6. *Given a SCM $M\langle U, V, F, P(u)\rangle$, assume that V is partitioned into k C-components $S^{(1)}, S^{(2)}, \ldots, S^{(k)}$. Let V_{pa} denote the set $\{V^{(i)}_{pa^{(i)}} | \forall V^{(i)} \in V\}$ and $S^{(j)}_{pa}$ denote the set*

$\{V^{(i)}_{pa^{(i)}} | \forall V^{(i)} \in S^{(j)}\}$ *where* $pa^{(i)}$ *is an arbitrary value for* $PA^{(i)}$*. The following independence relation holds:*

$$(\forall i \in \{1, \ldots, k\}) \quad \left(S^{(i)}_{pa} \perp\!\!\!\perp V_{pa} \backslash S^{(i)}_{pa} \right) \tag{12}$$

where $V_{pa} \backslash S^{(i)}_{pa}$ *is the set difference between* V_{pa} *and* $S^{(i)}_{pa}$*.*

Proof. The independence relation 12 is implied by the decomposition 10. □

We now shift the gear and focus on MDPUCs. The general procedure of proof is following: We first find a set of equivalent events; We derive a set of independence relations with the equivalent events and Lemma 6; Finally, we prove Theorem 1, 2, and 3 with derived independence relations.

Lemma 7. *Given a MDPUC model* $M\langle \gamma, U, X, Y, S, F, P(u) \rangle$*, starting from state* $S^{(t)}$*, for any* $\forall s^{(t)}, \ldots, s^{(t+k)} \in S, x^{(t)}, \ldots, x^{(t+k)} \in X, k \in \mathbb{Z}_+$*, following events are equivalent:*

$$S^{(t+k)}_{x^{([t,t+k-1])}, s^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-1])}, s^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)} \tag{13}$$

$$S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t)} = s^{(t)} \tag{14}$$

Proof. **Step 1: 13 ⇒ 14.** We will prove this by induction.
Base Case: When $k = 1$, for the event 13, we have

$$\begin{aligned}
&S^{(t+1)}_{x^{(t)}, s^{(t)}} = s^{(t+1)}, S^{(t)}_{x^{(t)}} = s^{(t)} \\
&\Rightarrow S^{(t+1)}_{x^{(t)}} = s^{(t+1)}, S^{(t)}_{x^{(t)}} = s^{(t)} \quad \text{By composition} \\
&\Rightarrow S^{(t+1)}_{x^{(t)}} = s^{(t+1)}, S^{(t)} = s^{(t)} \quad \text{By exclusion restrictions}
\end{aligned}$$

The last step is the event 14 when $k = 1$.
Induction Step: Suppose the event 13 ⇒ 14 is true for case $k - 1$, we want to show that it still holds for case k. By composition, we have that:

$$S^{(t+k-1)}_{x^{([t,t+k-1])}, s^{([t,t+k-2])}} = s^{(t+k-1)} \Rightarrow S^{(t+k)}_{x^{([t,t+k-1])}, s^{([t,t+k-1])}} = S^{(t+k)}_{x^{([t,t+k-1])}, s^{([t,t+k-2])}}$$

By continuing the above process for conditions $S^{(t+k-2)}_{x^{([t,t+k-1])}, s^{([t,t+k-3])}} = s^{(t+k-2)}, \ldots, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)}$, we have that:

$$S^{(t+k)}_{x^{([t,t+k-1])}, s^{([t,t+k-1])}} = S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}$$

Since for $\forall k \geq 1$, $S^{(t+k-1)}$ has all of its parents fixed, by exclusion restrictions, we have that:

$$\begin{aligned}
&S^{(t+k-1)}_{x^{([t,t+k-1])}, s^{([t,t+k-2])}} = S^{(t+k-1)}_{x^{([t,t+k-2])}, s^{([t,t+k-2])}} \\
&S^{(t+k-2)}_{x^{([t,t+k-1])}, s^{([t,t+k-3])}} = S^{(t+k-2)}_{x^{([t,t+k-2])}, s^{([t,t+k-3])}} \\
&\vdots \\
&S^{(t)}_{x^{([t,t+k-1])}} = S^{(t)}_{x^{([t,t+k-2])}}
\end{aligned}$$

This leads to the condition 13 for case $k - 1$:

$$S^{(t+k-1)}_{x^{([t,t+k-2])}, s^{([t,t+k-2])}} = s^{(t+k-1)}, S^{(t+k-2)}_{x^{([t,t+k-2])}, s^{([t,t+k-3])}} = s^{(t+k-2)}, \ldots, S^{(t)}_{x^{([t,t+k-2])}} = s^{(t)}$$

Since the statement for case $k - 1$ holds, we have that:

$$S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t)} = s^{(t)}$$

Conclusion: By the principle of induction, the statement is true or all $k \in \mathbb{Z}_+$.

Step 2: 14 $\Rightarrow$ 13. We first consider that given $S^{(t)} = s^{(t)}$, $S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}$ can be written as:

$$\begin{aligned}
&S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t)} = s^{(t)} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)} \quad \text{By exclusion restrictions} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k-1])},s^{(t)}} = s^{(t+k)}, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)} \quad \text{By composition} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k-1])},s^{(t)}} = s^{(t+k)}, S^{(t)} = s^{(t)} \quad \text{By exclusion restrictions}
\end{aligned}$$

Given $S^{(t)} = s^{(t)}$, apply the above process to variables $S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, \ldots, S^{(t+1)}_{x^{(t)}} = s^{(t+1)}$, we have that:

$$S^{(t+k)}_{x^{([t,t+k-1])},s^{(t)}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-2])},s^{(t)}} = s^{(t+k-1)}, \ldots, S^{(t+1)}_{x^{(t)},s^{(t)}} = s^{(t+1)}, S^{(t)} = s^{(t)}$$

Since the variable $S^{(t+1)}_{x^{(t)},s^{(t)}}$ has all its parents fixed, it is subjected to exclusion restrictions. we can use $S^{(t+1)}_{x^{(t)},s^{(t)}}$ as $S^{(t)}$ and repeat our previous process. Given $S^{(t+1)}_{x^{(t)},s^{(t)}} = s^{(t+1)}$, $S^{(t+k)}_{x^{([t,t+k-1])},s^{(t)}} = s^{(t+k)}$ can be written as:

$$\begin{aligned}
&S^{(t+k)}_{x^{([t,t+k-1])},s^{(t)}} = s^{(t+k)}, S^{(t+1)}_{x^{(t)},s^{(t)}} = s^{(t+1)} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k-1])},s^{(t)}} = s^{(t+k)}, S^{(t+1)}_{x^{([t,t+k-1])},s^{(t)}} = s^{(t+1)} \quad \text{By exclusion restrictions} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+1])}} = s^{(t+k)}, S^{(t+1)}_{x^{([t,t+k-1])},s^{(t)}} = s^{(t+1)} \quad \text{By composition} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+1])}} = s^{(t+k)}, S^{(t+1)}_{x^{(t)},s^{(t)}} = s^{(t+1)} \quad \text{By exclusion restrictions}
\end{aligned}$$

Given $S^{(t+1)}_{x^{(t)},s^{(t)}} = s^{(t+1)}$, apply the above process to variables $S^{(t+k)}_{x^{([t,t+k-1])},s^{(t)}} = s^{(t+k)}, \ldots, S^{(t+2)}_{x^{([t,t+1]},s^{(t)}} = s^{(t+2)}$, we have that:

$$S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-2])},s^{([t,t+1])}} = s^{(t+k-1)}, \ldots, S^{(t+2)}_{x^{([t,t+1])},s^{([t,t+1])}} = s^{(t+2)}, \ldots, S^{(t)} = s^{(t)}$$

Now the variable $S^{(t+2)}_{x^{([t,t+1])},s^{([t,t+1])}}$ has all its parents fixed, we can again repeat our previous process by using $S^{(t+1)}_{x^{(t)},s^{(t)}}$ as $S^{(t)}$. Continue this procedure for all variables, in the end, we have:

$$S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-2])},s^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t+1)}_{x^{(t)},s^{(t)}} = s^{(t+1)}, S^{(t)} = s^{(t)}$$

Note that for all $\forall k \in \mathbb{Z}_+$, the variable $S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}} = s^{(t+k)}$ has all of its parents fixed. By exclusion restrictions, we have:

$$S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-1])},s^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t)}_{x^{([1,t+k-1])}} = s^{(t)}$$

which is the event 13. □

Lemma 8. *Given a MDPUC model $M\langle\gamma, U, X, Y, S, F, P(u)\rangle$, starting from state $S^{(t)}$, for any $\forall s^{(t)}, \ldots, s^{(t+k)} \in S, x^{(t)}, \ldots, x^{(t+k)} \in X, k \in \mathbb{Z}_+$, following events are equivalent:*

$$S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-1])},s^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \tag{15}$$

$$S^{(t+k)}_{x^{([t+1,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t+1,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t+1)} = s^{(t+1)}, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \tag{16}$$

Proof. **Step 1: 15 ⇒ 16.** By Lemma 7, event 15 is equivalent to:

$$S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t+1)}_{x^{(t)}} = s^{(t+1)}, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$

Given $S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$, the variable $S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}$ can be written as:

$$\begin{aligned}
&S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)}, X^{(t)}_{x^{([t+1,t+k-1])},s^{(t)}} = x^{(t)} \quad \text{By exclusion restrictions} \\
&\Rightarrow S^{(t+k)}_{x^{([t+1,t+k-1])},s^{(t)}} = s^{(t+k)}, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)}, X^{(t)}_{x^{([t+1,t+k-1])},s^{(t)}} = x^{(t)} \quad \text{By composition} \\
&\Rightarrow S^{(t+k)}_{x^{([t+1,t+k-1])},s^{(t)}} = s^{(t+k)}, S^{(t)}_{x^{([t+1,t+k-1])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \quad \text{By exclusion restrictions} \\
&\Rightarrow S^{(t+k)}_{x^{([t+1,t+k-1])}} = s^{(t+k)}, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \quad \text{By composition and exclusion restrictions}
\end{aligned}$$

By applying the above steps to variables $S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t+1)}_{x^{(t)}} = s^{(t+1)}$, we have that:

$$S^{(t+k)}_{x^{([t+1,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t+1,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t+1)} = s^{(t+1)}, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$

which is the event 16

Step 4: 16 ⇒ 15. As for the event 16, given $S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$, the variable $S^{(t+k)}_{x^{([t+1,t+k-1])}} = s^{(t+k)}$ can be written as:

$$\begin{aligned}
&S^{(t+k)}_{x^{([t+1,t+k-1])}} = s^{(t+k)}, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \\
&\Rightarrow S^{(t+k)}_{x^{([t+1,t+k-1])}} = s^{(t+k)}, S^{(t)}_{x^{([t+1,t+k-1])}} = s^{(t)}, X^{(t)}_{x^{([t+1,t+k-1])},s^{(t)}} = x^{(t)} \quad \text{By exclusion restrictions} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k-1])},s^{(t)}} = s^{(t+k)}, S^{(t)}_{x^{([t+1,t+k-1])}} = s^{(t)}, X^{(t)}_{x^{([t+1,t+k-1])},s^{(t)}} = x^{(t)} \quad \text{By composition} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k-1])},s^{(t)}} = s^{(t+k)}, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \quad \text{By exclusion restrictions} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \quad \text{By composition} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \quad \text{By exclusion restrictions}
\end{aligned}$$

By applying the above steps to variables $S^{(t+k)}_{x^{([t+1,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t+1,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t+1)} = s^{(t+1)}$, we have that:

$$S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t+1)}_{x^{(t)}} = s^{(t+1)}, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$

By Lemma 7, the above event is equivalent to:

$$S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-1])},s^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$

which is the event 16. □

Lemma 9. *Given a MDPUC model $M\langle\gamma, U, X, Y, S, F, P(u)\rangle$, starting from state $S^{(t)}$, for any $\forall s^{(t)}, \ldots, s^{(t+k)} \in S, x^{(t)}, \ldots, x^{(t+k)} \in X, k \in \mathbb{Z}_+$, if any of following statements holds:*

$$S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-1])},s^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)} \quad (17)$$

$$S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t)} = s^{(t)} \quad (18)$$

we must have:

$$S^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} = S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = S^{(t+k+1)}_{x^{([t,t+k])}} \quad (19)$$

$$Y^{(t+k)}_{x^{(t+k)},s^{(t+k)}} = Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}} = Y^{(t+k)}_{x^{([t,t+k])}} \quad (20)$$

Proof. Since by Lemma 7, events 17, 18 equivalent, we can focus on the condition 17.
Step 1: the statement 19 holds. Since the variable $S^{(k+1)}_{x^{(k)},s^{(k)}}$ has all its parents fixed, $S^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} = S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}}$ is obvious by exclusion restrictions. Let us now consider $S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = S^{(t+k+1)}_{x^{([t,t+k])}}$, for any $\forall s^{(t+k+1)} \in S$, let $S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = s^{(t+k+1)}$. Given condition 17, by exclusion restrictions, we have that:

$$\begin{aligned}
&S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = s^{(t+k+1)}, S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}} = s^{(t+k)}, \ldots, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)} \\
&\Rightarrow S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = s^{(t+k+1)}, S^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}} = s^{(t+k)}, \ldots, S^{(t)}_{x^{([t,t+k])}} = s^{(t)}
\end{aligned}$$

The above event is in fact case $k+1$ in Lemma 7. By Lemma 7, we must have:

$$\begin{aligned}
&S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = s^{(t+k+1)}, S^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}} = s^{(t+k)}, \ldots, S^{(t)}_{x^{([t,t+k])}} = s^{(t)} \\
&\Leftrightarrow S^{(t+k+1)}_{x^{([t,t+k])}} = s^{(t+k+1)}, S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, \ldots, S^{(t)} = s^{(t)}
\end{aligned}$$

This implies that:

$$S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = S^{(t+k+1)}_{x^{([t,t+k])}} = s^{(t+k+1)}$$

Step 2: the statement 20 holds. Since the variable $Y^{(t+k)}_{x^{(t+k)},s^{(t+k)}}$ has all its parents fixed, $Y^{(t+k)}_{x^{(t+k)},s^{(t+k)}} = Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}}$ is obvious by exclusion restrictions. Let us now consider $Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}} = Y^{(t+k)}_{x^{([t,t+k])}}$. For any $\forall y^{(t+k)} \in Y$, let $Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}} = y^{(t+k)}$. Given condition 17, by exclusion restrictions, we have:

$$\begin{aligned}
&S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-1])},s^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)} \\
&\Rightarrow S^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k])},s^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t)}_{x^{([t,t+k])}} = s^{(t)}
\end{aligned}$$

Given $S^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}} = s^{(t+k)}$, by composition, we have that:

$$S^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}} = s^{(t+k)} \Rightarrow Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}} = Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}}$$

By applying the above process given conditions $S^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}} = s^{(t+k)}, \ldots, S^{(t)}_{x^{([t,t+k])}} = s^{(t)}$, we must have:

$$Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}} = Y^{(t+k)}_{x^{([t,t+k])}} = y^{(t+k)}$$

which proves the statement 20. □

Lemma 10. *Given a MDPUC model $M\langle\gamma, U, X, Y, S, F, P(u)\rangle$, starting from state $S^{(t)}$, for any $\forall s^{(t)}, \ldots, s^{(t+k)} \in S, x^{(t)}, \ldots, x^{(t+k)} \in X, k \in \mathbb{Z}_+$, if any of following statements holds:*

$$S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-1])},s^{([t,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \tag{21}$$

$$S^{(t+k)}_{x^{([t+1,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t+1,t+k-2])}} = s^{(t+k-1)}, \ldots, S^{(t+1)} = s^{(t+1)}, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \tag{22}$$

the following statements must hold:

$$S^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} = S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = S^{(t+k+1)}_{x^{([t,t+k])}} = S^{(t+k+1)}_{x^{([t+1,t+k])}} \tag{23}$$

$$Y^{(t+k)}_{x^{(t+k)},s^{(t+k)}} = Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}} = Y^{(t+k)}_{x^{([t,t+k])}} = Y^{(t+k)}_{x^{([t+1,t+k])}} \tag{24}$$

Proof. Since by Lemma 8, events 21, 22 are equivalent, we can focus on the condition 21.
Step 1: the statement 23 holds. Since the variable $S^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}$ has all its parents fixed, $S^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} = S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}}$ is obvious by exclusion restrictions. Let us now consider $S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = S^{(t+k+1)}_{x^{([t,t+k])}} = S^{(t+k+1)}_{x^{([t+1,t+k])}}$. For any $\forall s^{(t+k+1)} \in S$, let $S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = s^{(t+k+1)}$. Given condition 21, by exclusion restrictions, we have that:

$$S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = s^{(t+k+1)}, S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}} = s^{(t+k)}, \dots, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$
$$\Rightarrow S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = s^{(t+k+1)}, S^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}} = s^{(t+k)}, \dots, S^{(t)}_{x^{([t,t+k])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$

The above satisfies case $k+1$ in 7 and 8, which leads to:

$$S^{(t+k+1)}_{x^{([t,t+k])}} = s^{(t+k+1)}, S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, \dots, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$
$$S^{(t+k+1)}_{x^{([t+1,t+k])}} = s^{(t+k+1)}, S^{(t+k)}_{x^{([t+1,t+k-1])}} = s^{(t+k)}, \dots, S^{(t+1)} = s^{(t+1)}, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$

This implies that:

$$S^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}} = S^{(t+k+1)}_{x^{([t,t+k])}} = S^{(t+k+1)}_{x^{([t+1,t+k])}} = s^{(t+k+1)}$$

Step 1: the statement 24 holds. Since the variable $Y^{(t+k)}_{x^{(t+k)},s^{(t+k)}}$ has all its parents fixed, $Y^{(t+k)}_{x^{(t+k)},s^{(t+k)}} = Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}}$ is obvious by exclusion restrictions. Let us now consider $Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}} = Y^{(t+k)}_{x^{([t,t+k])}} = Y^{(t+k)}_{x^{([t+1,t+k])}}$. For any $\forall y^{(t+k)} \in Y$, let $Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}} = y^{(t+k)}$. Given condition 21, by exclusion restrictions, we have:

$$S^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-1])},s^{([t,t+k-2])}} = s^{(t+k-1)}, \dots, S^{(t)}_{x^{([t,t+k-1])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$
$$\Rightarrow S^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k])},s^{([t,t+k-2])}} = s^{(t+k-1)}, \dots, S^{(t)}_{x^{([t,t+k])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$

Given $S^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}} = s^{(t+k)}$, by composition, we have that:

$$S^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}} = s^{(t+k)} \Rightarrow Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}} = Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}}$$

By applying the above process given conditions $S^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k-1])}} = s^{(t+k)}, \dots, S^{(t)}_{x^{([t,t+k])}} = s^{(t)}$, we must have:

$$Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}} = Y^{(t+k)}_{x^{([t,t+k])}} = y^{(t+k)}$$

Given condition 21, Lemma 7 implies that:

$$S^{(t+k)}_{x^{([t,t+k-1])}} = s^{(t+k)}, S^{(t+k-1)}_{x^{([t,t+k-2])}} = s^{(t+k-1)}, \dots, S^{(t)} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$

Given $S^{(t)}_{x^{([t,t+k])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$, $Y^{(t+k)}_{x^{([t,t+k])}} = y^{(t+k)}$ can be written as:

$$Y^{(t+k)}_{x^{([t,t+k])}} = y^{(t+k)}, S^{(t)}_{x^{([t,t+k])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)}$$
$$\Rightarrow Y^{(t+k)}_{x^{([t,t+k])}} = y^{(t+k)}, S^{(t)}_{x^{([t,t+k])}} = s^{(t)}, X^{(t)}_{x^{([t+1,t+k])},s^{(t)}} = x^{(t)} \quad \text{By exclusion restrictions}$$
$$\Rightarrow Y^{(t+k)}_{x^{([t+1,t+k])},s^{(t)}} = y^{(t+k)}, S^{(t)}_{x^{([t,t+k])}} = s^{(t)}, X^{(t)}_{x^{([t+1,t+k])},s^{(t)}} = x^{(t)} \quad \text{By composition}$$
$$\Rightarrow Y^{(t+k)}_{x^{([t+1,t+k])},s^{(t)}} = y^{(t+k)}, S^{(t)}_{x^{([t+1,t+k])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \quad \text{By exclusion restrictions}$$
$$\Rightarrow Y^{(t+k)}_{x^{([t+1,t+k])}} = y^{(t+k)}, S^{(t)}_{x^{([t,t+k])}} = s^{(t)}, X^{(t)}_{s^{(t)}} = x^{(t)} \quad \text{By composition and exclusion restrictions}$$

Therefore, the following statement must hold:

$$Y^{(t+k)}_{x^{([t,t+k])},s^{([t,t+k])}} = Y^{(t+k)}_{x^{([t,t+k])}} = Y^{(t+k)}_{x^{([t+1,t+k])}} = y^{(t+k)}$$

which is the statement 24. □

Lemma 11. *For a MDPUC model $M = \langle \gamma, U, X, Y, S, F, P(u) \rangle$, starting from state $S^{(t)}$, $\forall s^{(t)}, \ldots, s^{(t+k)} \in S, x^{(t)}, \ldots, x^{(t+k)} \in X, \forall y^{(t+k)} \in Y, k \in \mathbb{Z}_+$, the following statements holds:*

$$P\left(s^{(t+k+1)}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t)}\right) = P\left(s^{(t+k+1)}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}\right) \tag{25}$$

$$P\left(y^{(t+k)}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t)}\right) = P\left(y^{(t+k)}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}\right) \tag{26}$$

Proof. We will only prove for Equation 25, since the proof for Equation 26 follows very similar steps.

By Lemma 6, for any $\forall y^{(t)}, \ldots, y^{(t+k)} \in Y$, we have that:

$$\begin{aligned}
&\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, y^{(t+k)}_{x^{(t+k)},s^{(t+k)}}, x^{(t+k)}_{s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{(t+k-1)},s^{(t+k-1)}}, y^{(t+k-1)}_{x^{(t+k-1)},s^{(t+k-1)}}, x^{(t+k-1)}_{s^{(t+k-1)}}, \ldots, s^{(t+1)}_{x^{(t)},s^{(t)}}, y^{(t)}_{x^{(t)},s^{(t)}}, x^{(t)}_{s^{(t)}}, s^{(t)}\right) \\
&\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{(t+k-1)},s^{(t+k-1)}}, \ldots, s^{(t+2)}_{x^{(t+1)},s^{(t+1)}}, s^{(t+1)}_{x^{(t)},s^{(t)}}, x^{(t)}_{s^{(t)}}, s^{(t)}\right) \quad \text{by decomposition axiom} \\
&\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}}, \ldots, s^{(t+2)}_{x^{([t,t+k-1])},s^{([t,t+1])}}, s^{(t+1)}_{x^{([t,t+k-1])},s^{(t)}}, s^{(t)}_{x^{([t,t+k-1])}}, x^{(t)}_{s^{(t)}}\right) \\
&\qquad \text{by exclusion restrictions} \\
&\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])}}, s^{(t+1)}_{x^{(t)}}, s^{(t)}, x^{(t)}_{s^{(t)}}\right) \quad \text{By Lemma 7} \\
&\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \perp\!\!\!\perp x^{(t)}_{s^{(t)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])}}, s^{(t+1)}_{x^{(t)}}, s^{(t)}\right) \quad \text{by weak union}
\end{aligned} \tag{27}$$

Similarly, we have that:

$$\begin{aligned}
&\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}}, \ldots, s^{(t+2)}_{x^{([t,t+k-1])},s^{([t,t+1])}}, s^{(t+1)}_{x^{([t,t+k-1])},s^{(t)}}, s^{(t)}_{x^{([t,t+k-1])}}, x^{(t)}_{s^{(t)}}\right) \\
&\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}, s^{(t)}, x^{(t)}_{s^{(t)}}\right) \quad \text{By Lemma 8} \\
&\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \perp\!\!\!\perp x^{(t)}_{s^{(t)}}, s^{(t)} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right) \quad \text{by weak union}
\end{aligned} \tag{28}$$

The independence relation 27 implies that

$$\begin{aligned}
&P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])}}, s^{(t+1)}_{x^{(t)}}, s^{(t)}, x^{(t)}_{s^{(t)}}\right) \\
&= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])}}, s^{(t+1)}_{x^{(t)}}, s^{(t)}\right)
\end{aligned} \tag{29}$$

Similarly, the independence relation 28 implies that:

$$\begin{aligned}
&P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}, s^{(t)}, x^{(t)}_{s^{(t)}}\right) \\
&= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right)
\end{aligned} \tag{30}$$

Also, Lemma 7 and 8 imply that:

$$
\begin{aligned}
&P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])}}, s^{(t+1)}_{x^{(t)}}, s^{(t)}, x^{(t)}_{s^{(t)}}\right) \\
&= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])},s^{([t,t+k-2])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])},s^{([t,t+1])}}, s^{(t+1)}_{x^{(t)},s^{(t)}}, s^{(t)}, x^{(t)}_{s^{(t)}}\right) \\
&\qquad \text{By Lemma 7} \\
&= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}, s^{(t)}, x^{(t)}_{s^{(t)}}\right) \quad \text{By Lemma 8}
\end{aligned}
$$

Together with equation 29 and 30, we have:

$$
\begin{aligned}
&\underbrace{P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])}}, s^{(t+1)}_{x^{(t)}}, s^{(t)}\right)}_{\text{Term 1}} \\
&= \underbrace{P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right)}_{\text{Term 2}}
\end{aligned} \tag{31}
$$

By Lemma 9, given $s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])}}, s^{(t+1)}_{x^{(t)}}, s^{(t)}$, Term 1 equals to:

$$
\begin{aligned}
&P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])}}, s^{(t+1)}_{x^{(t)}}, s^{(t)}\right) \\
&= P\left(s^{(t+k+1)}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])}}, s^{(t+1)}_{x^{(t)}}, s^{(t)}\right)
\end{aligned}
$$

Term 2 can be written as:

$$
\begin{aligned}
&P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right) \\
&= \sum_{s'(t)\in S} \sum_{x^{'(t)}_{s'(t)}\in X} P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}, x^{'(t)}_{s'(t)}, s'(t)\right) \\
&\cdot P\left(x^{'(t)}_{s'(t)}, s'(t) \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right)
\end{aligned} \tag{32}
$$

By Lemma 10, given $s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}, x^{'(t)}_{s'(t)}, s'(t)$, we have:

$$
\begin{aligned}
&P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}, x^{'(t)}_{s'(t)}, s'(t)\right) \\
&= P\left(s^{(t+k+1)}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}, x^{'(t)}_{s'(t)}, s'(t)\right) \\
&= \frac{P\left(s^{(t+k+1)}_{x^{([t+1,t+k])}}, x^{'(t)}_{s'(t)}, s'(t) \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right)}{P\left(x^{'(t)}_{s'(t)}, s'(t) \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right)}
\end{aligned} \tag{33}
$$

Replace $P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}, x^{'(t)}_{s'(t)}, s'(t)\right)$ in equa-

tion 32 with 33:

$$
\begin{aligned}
&P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right) \\
&= \sum_{s'^{(t)} \in S} \sum_{x'^{(t)}_{s'^{(t)}} \in X} \frac{P\left(s^{(t+k+1)}_{x^{([t+1,t+k])}}, x'^{(t)}_{s'^{(t)}}, s'^{(t)} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right)}{P\left(x'^{(t)}_{s'^{(t)}}, s'^{(t)} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right)} \\
&\cdot P\left(x'^{(t)}_{s'^{(t)}}, s'^{(t)} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right) \\
&= \sum_{s'^{(t)} \in S} \sum_{x'^{(t)}_{s'^{(t)}} \in X} P\left(s^{(t+k+1)}_{x^{([t+1,t+k])}}, x'^{(t)}_{s'^{(t)}}, s'^{(t)} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right) \\
&= P\left(s^{(t+k+1)}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}\right)
\end{aligned}
$$

Together with equation 31, we have that:

$$
P(s^{(t+k+1)}_{x^{([t,t+k])}} | s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t)}) = P(s^{(t+k+1)}_{x^{([t+1,t+k])}} | s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)})
$$

□

Lemma 1 (Proof). $P\left(Y^{t+k}_{x^{(t)},x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right)$ can be written as:

$$
\begin{aligned}
&P\left(Y^{t+k}_{x^{(t)},x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right) \\
&= \sum_{s \in S^{k-2+1}} P\left(y^{t+k}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, s^{(t)}\right)\Big|_{x^{([t+1,t+k])}=\pi} \\
&\cdot P\left(s^{(t+k)}_{x^{([t,t+k-1])}} \mid s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, s^{(t)}\right) \\
&\cdot P\left(s^{(t+k-1)}_{x^{([t,t+k-2])}} \mid s^{(t+k-2)}_{x^{([t,t+k-3])}}, \ldots, s^{(t+1)}_{x^{(t)}}, s^{(t)}\right) \\
&\vdots \\
&P\left(s^{t+2}_{x^{([t,t+1])}} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right)
\end{aligned}
$$

where s is defined as a sequence $s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{t+2}_{x^{([t,t+1])}}$. By Lemma 11, we have that:

$$
\begin{aligned}
&P\left(y^{t+k}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, s^{(t)}\right) \\
&= P\left(y^{t+k}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}\right)
\end{aligned}
$$

By Lemma 11, for any $\forall k \in \mathbb{Z}_+$:

$$
\begin{aligned}
&P\left(s^{(t+k)}_{x^{([t,t+k-1])}} \mid s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, s^{(t)}\right) \\
&= P\left(s^{(t+k)}_{x^{([t+1,t+k-1])}} \mid s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}\right)
\end{aligned}
$$

We then have that:

$$
\begin{aligned}
&P\left(Y^{t+k}_{x^{(t)},x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right) \\
&= \sum_{s \in S^{k-2+1}} P\left(y^{t+k}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}\right)\Bigg|_{x^{([t+1,t+k])}=\pi} \\
&\cdot P\left(s^{(t+k)}_{x^{([t+1,t+k-1])}} \mid s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}\right) \\
&\cdot P\left(s^{(t+k-1)}_{x^{([t+1,t+k-2])}} \mid s^{(t+k-2)}_{x^{([t+1,t+k-3])}}, \ldots, s^{(t+1)}\right) \\
&\vdots \\
&P\left(s^{t+2}_{x^{(t+1)}} \mid s^{(t+1)}\right) \\
&= P\left(Y^{t+k}_{x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}\right)
\end{aligned}
$$

which proves the statement. □

Theorem 1 (Proof). We first expand $V^{\pi}(s^{(t)})$ as

$$
\begin{aligned}
V^{\pi}(s^{(t)}) &= \mathbb{E}\left[Y^{(t)}_{x^{(t)}=\pi} + \sum_{k=1}^{\infty} \gamma^{t+k} Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}=\pi} \mid s^{(t)}\right] + \gamma \sum_{s^{(t+1)} \in S} P(s^{(t+1)}_{x^{(t)}=\pi} \mid s^{(t)}) \sum_{k=1}^{\infty} \gamma^{t+k} \mathbb{E}\left[Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t+1)}_{x^{(t)}=\pi}, s^{(t)}\right] \quad (34)
\end{aligned}
$$

We can re-write the first factor of Eq. 34 as

$$
\begin{aligned}
\mathbb{E}\left[Y^{(t)}_{x^{(t)}=\pi} \mid s^{(t)}\right] &= \sum_{y^{(t)} \in Y} y^{(t)} P\left(y^{(t)}_{x^{(t)}=\pi} \mid s^{(t)}\right) \\
&= \sum_{y^{(t)} \in Y} y^{(t)} P\left(y^{(t)}_{x^{(t)}} \mid s^{(t)}\right)\Bigg|_{x^{(t)}=\pi} \quad (35) \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}\right]\Bigg|_{x^{(t)}=\pi} \quad (36)
\end{aligned}
$$

Eq. 35 follows from Lemma 3, since $x^{(t)} = \pi(s^{(t)})$ and $s^{(t)}$ is fixed. For similar reason, the second factor of Eq. 34 can be written as

$$
P\left(s^{(t+1)}_{x^{(t)}=\pi} \mid s^{(t)}\right) = P\left(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}\right)\Bigg|_{x^{(t)}=\pi} \quad (37)
$$

Further, the third factor of Eq. 34 can be written as

$$
\begin{aligned}
&\mathbb{E}\left[Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right] = \\
&= \sum_{y^{(t+k)} \in Y} y^{(t+k)} P\left(Y^{(t+k)}_{x^{([t,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}=\pi}, s^{(t)}\right) \\
&= \sum_{y^{(t+k)} \in Y} y^{(t+k)} P\left(Y^{(t+k)}_{x^{(t)},x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right)\Bigg|_{x^{(t)}=\pi}
\end{aligned}
$$

The last step follows from Lemma 3. By Lemma 1, we then have

$$
\begin{aligned}
&\mathbb{E}\left[Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t+1)}_{x^{(t)}=\pi}, s^{(t)}\right] \\
&= \sum_{y^{(t+k)}\in Y} y^{(t+k)} P\left(Y^{t+k}_{x^{(t)},x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right)\Big|_{x^{(t)}=\pi} \\
&= \sum_{y^{(t+k)}\in Y} y^{(t+k)} P\left(Y^{t+k}_{x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}\right)\Big|_{x^{(t)}=\pi} \quad \text{By Lemma 1} \\
&= \sum_{y^{(t+k)}\in Y} y^{(t+k)} P\left(Y^{t+k}_{x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}\right) \quad Y^{(t)} \text{ is now independent of } x^{(t)}. \\
&= \mathbb{E}\left[Y^{(t+k)}_{x^{[t+1,t+k]}=\pi} \mid s^{(t+1)}\right] \qquad (38)
\end{aligned}
$$

We then have substituting Eqs. 36, 37 and 38 back into Eq. 34,

$$
\begin{aligned}
V^{\pi}(s^{(t)}) &= \mathbb{E}\left[Y^{(t)}_{x^{(t)}=\pi} \mid s^{(t)}\right] + \gamma \sum_{s^{(t+1)}\in S} P\left(s^{(t+1)}_{x^{(t)}=\pi} \mid s^{(t)}\right) \sum_{k=1}^{\infty} \gamma^{t+k} \mathbb{E}\left[Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}\right]\Big|_{x^{(t)}=\pi} + \gamma \sum_{s^{(t+1)}\in S} P\left(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}\right)\Big|_{x^{(t)}=\pi} \sum_{k=1}^{\infty} \gamma^{t+k} \mathbb{E}\left[Y^{(t+k)}_{x^{([t+1,t+k])}=\pi} \mid s^{(t+1)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}\right]\Big|_{x^{(t)}=\pi} + \gamma \sum_{s^{(t+1)}\in S} P\left(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}\right)\Big|_{x^{(t)}=\pi} \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{t+k} Y^{(t+k)}_{x^{([t+1,t+k])}=\pi} \mid s^{(t+1)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}\right]\Big|_{x^{(t)}=\pi} + \gamma \sum_{s^{(t+1)}\in S} P\left(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}\right)\Big|_{x^{(t)}=\pi} V^{\pi}(s^{(t+1)})
\end{aligned}
$$

Similarly, we can expand $Q^{\pi}(s^{(t)}, x^{(t)})$ as

$$
\begin{aligned}
Q^{\pi}(s^{(t)}, x^{(t)}) &= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}\right] + \gamma \sum_{s^{(t+1)}\in S} P(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}) \sum_{k=1}^{\infty} \gamma^{t+k} \mathbb{E}\left[Y^{(t+k)}_{x^{(t)},x^{([t+1,t+k])}=\pi} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}\right] + \gamma \sum_{s^{(t+1)}\in S} P\left(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}\right) \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{t+k} Y^{(t+k)}_{x^{([t+1,t+k])}=\pi} \mid s^{(t+1)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}\right] + \gamma \sum_{s^{(t+1)}\in S} P\left(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}\right) V^{\pi}(s^{(t+1)})
\end{aligned}
$$

□

Theorem 2 (Proof). For any policy $f \in F_{exp}$ which is function from S to X, f is also a function from $S \times X$ to X. We must have $f \in F_{ctf}$. This means that $F_{exp} \subseteq F_{ctf}$ and $\pi^*_{exp} \in F_{ctf}$.

Since for any state $s^{(t)}$ and intuition $x'^{(t)}$, $\pi^*_{ctf} = \arg\max_{\pi \in F_{ctf}} V^{\pi}(s^{(t)}, x'^{(t)})$, we have

$V^{\pi^*_{exp}}(s^{(t)}, x'^{(t)}) \leq V^{\pi^*_{ctf}}(s^{(t)}, x'^{(t)})$. For $V^{\pi^*_{exp}}(s^{(t)})$, we have

$$
\begin{aligned}
V^{\pi^*_{exp}}(s[t]) &= \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k Y^{(t+k)}_{x^{([t,t+k])}=\pi^*_{exp}} \mid s^{(t)}\Big] \\
&= \sum_{x'^{(t)} \in X} P(x'^{(t)}|s^{(t)}) \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k Y^{(t+k)}_{x^{([t,t+k])}=\pi^*_{exp}} \mid s^{(t)}, x'^{(t)}\Big] \\
&= \sum_{x'^{(t)} \in X} P(x'^{(t)}|s^{(t)}) V^{\pi^*_{exp}}(s^{(t)}, x'^{(t)}) \\
&\leq \sum_{x'[t] \in X} P(x'^{(t)}|s^{(t)}) V^{\pi^*_{ctf}}(s^{(t)}, x'^{(t)}) \\
&= V^{\pi^*_{ctf}}(s^{(t)})
\end{aligned}
$$

We will show that the equality does not always hold by constructing a counterexample. Consider a simple MDPUC with $S_t \in \{0\}$, $X_t \in \{0,1\}$ and $U_t \in \{0,1\}$. The reward Y_t is defined by function $f(X_t, U_t) = x_t \oplus u_t$. At each round, the agent always receives an intuition $x_t = \neg u_t$. U_t is a uniform random variable, which means $P(U_t = 1) = P(U_t = 0) = \frac{1}{2}$. The transition function is defined as $S_{t+1} = X_t$. Let discount factor $\gamma = 0.5$.

With this MDP instance, for all state based policies $\pi \in F_{exp}$, starting from any state $s_t \in \{0\}$, the expected return is always $0.5/(1-0.5) = 1$. While for intuition based policy, let $x'_t = \pi^*_{ctf}(s_t, x_t) = \neg x_t$. The expected return for π^*_{ctf} starting from any state $s_t \in \{0,1\}$ equals to $1/(1-0.5) = 2$. Therefore, the optimal value function for state based policy and intuition based policy are not always the same. The optimal intuition based policy is also optimal in terms of state based value function. □

Lemma 12. *For a MDPUC model $M = \langle \gamma, U, X, Y, S, F, P(u) \rangle$, starting from state $S^{(t)}$, for any $\forall s^{(t+k)} \in S, x^{([t+i,t+j])} \in X^{j-i+1}, i, j, k \in \mathbb{Z}, k > j$ and $k \geq 1$, the following statement must hold:*

$$
S^{(t+k)}_{x^{([t+i,t+j])}} = s^{(t+k)} \Rightarrow X^{(t+k)}_{s^{(t+k)}} = X^{(t+k)}_{x^{([t+i,t+j])}}
$$

Proof. This can be easily proved by exclusion restrictions and composition. For any $\forall x^{(t+k)} \in X$, we have that:

$$
\begin{aligned}
& S^{(t+k)}_{x^{([t+i,t+j])}} = s^{(t+k)}, X^{(t+k)}_{s^{(t+k)}} = x^{(t+k)} \\
& \Rightarrow S^{(t+k)}_{x^{([t+i,t+j])}} = s^{(t+k)}, X^{(t+k)}_{x^{([t+i,t+j])}, s^{(t+k)}} = x^{(t+k)} \quad \text{By exclusion restrictions} \\
& \Rightarrow S^{(t+k)}_{x^{([t+i,t+j])}} = s^{(t+k)}, X^{(t+k)}_{x^{([t+i,t+j])}} = x^{(t+k)} \quad \text{By composition}
\end{aligned}
$$

The last step implies that:

$$
X^{(t+k)}_{s^{(t+k)}} = X^{(t+k)}_{x^{([t+i,t+j])}} = x^{(t+k)}
$$

□

Lemma 13. *For a MDPUC model $M = \langle \gamma, U, X, Y, S, F, P(u) \rangle$, starting from state $S^{(t)}$, for any $\forall y^{(t+k)} \in Y, \forall s^{(t)}, \ldots, s^{(t+k)} \in S, x^{(t)}, x'^{(t)}, \ldots, x^{(t+k)}, x'^{(t+k)} \in X, k \in \mathbb{Z}_+$, the*

following statements holds:

$$P\left(s^{(t+k+1)}_{x^{([t,t+k])}}, x^{'(t+k+1)}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t)}, x^{'(t)}\right)$$
$$= P\left(s^{(t+k+1)}_{x^{([t+1,t+k])}}, x^{'(t+k+1)}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x^{'(t+1)}\right) \tag{39}$$

$$P\left(y^{(t+k)}_{x^{([t,t+k])}}, x^{'(t+k)}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t)}, x^{'(t)}\right)$$
$$= P\left(y^{(t+k)}_{x^{([t+1,t+k])}}, x^{'(t+k)}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x^{'(t+1)}\right) \tag{40}$$

Proof. We will only prove for Equation 39, since the proof for Equation 40 follows very similar steps.

By Lemma 6, for any $\forall y^{(t)}, \ldots, y^{(t+k)} \in Y$, we have that:

$$\left(s^{(t+k+2)}_{x^{(t+k+1)},s^{(t+k+1)}}, y^{(t+k+1)}_{x^{(t+k+1)},s^{(t+k+1)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \perp\!\!\!\perp s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, y^{(t+k)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k)}_{s^{(t+k)}}, \ldots, s^{(t+1)}_{x^{(t)},s^{(t)}}, y^{(t)}_{x^{(t)},s^{(t)}}, x^{'(t)}_{s^{(t)}}, s^{(t)}\right)$$
$$\Rightarrow \left(s^{(t+k+2)}_{x^{(t+k+1)},s^{(t+k+1)}}, y^{(t+k+1)}_{x^{(t+k+1)},s^{(t+k+1)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \perp\!\!\!\perp s^{(t+k)}_{x^{(t+k-1)},s^{(t+k-1)}}, y^{(t+k-1)}_{x^{(t+k-1)},s^{(t+k-1)}}, x^{'(t+k-1)}_{s^{(t+k-1)}} \ldots, y^{(t)}_{x^{(t)},s^{(t)}}, x^{'(t)}_{s^{(t)}}, s^{(t)}\right.$$
$$\left. \mid s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, y^{(t+k)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k)}_{s^{(t+k)}}\right) \quad \text{By weak union} \tag{41}$$

and

$$\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, y^{(t+k)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k)}_{s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{(t+k-1)},s^{(t+k-1)}}, y^{(t+k-1)}_{x^{(t+k-1)},s^{(t+k-1)}}, x^{'(t+k-1)}_{s^{(t+k-1)}} \ldots, y^{(t)}_{x^{(t)},s^{(t)}}, x^{'(t)}_{s^{(t)}}, s^{(t)}\right) \tag{42}$$

With independence relations 41 and 42, by contraction and decomposition graphoid axioms, we have that:

$$\left(s^{(t+k+2)}_{x^{(t+k+1)},s^{(t+k+1)}}, y^{(t+k+1)}_{x^{(t+k+1)},s^{(t+k+1)}}, x^{'(t+k+1)}_{s^{(t+k+1)}}, s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, y^{(t+k)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k)}_{s^{(t+k)}} \perp\!\!\!\perp\right.$$
$$\left. s^{(t+k)}_{x^{(t+k-1)},s^{(t+k-1)}}, y^{(t+k-1)}_{x^{(t+k-1)},s^{(t+k-1)}}, x^{'(t+k-1)}_{s^{(t+k-1)}}, \ldots, s^{(t+1)}_{x^{(t)},s^{(t)}}, y^{(t)}_{x^{(t)},s^{(t)}}, s^{(t)}, x^{'(t)}_{s^{(t)}}\right)$$
$$\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}}, x^{'(t+k)}_{s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{(t+k-1)},s^{(t+k-1)}}, x^{'(t+k-1)}_{s^{(t+k-1)}}, \ldots, s^{(t+1)}_{x^{(t)},s^{(t)}}, s^{(t)}, x^{'(t)}_{s^{(t)}}\right) \quad \text{By decomposition}$$
$$\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}}, x^{'(t+k)}_{s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-1])},s^{([t,t+k-2])}}, x^{'(t+k-1)}_{s^{(t+k-1)}}, \ldots, s^{(t)}_{x^{([t,t+k-1])}}, x^{'(t)}_{s^{(t)}}\right)$$

by exclusion restrictions

$$\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}}, x^{'(t+k)}_{s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{s^{(t+k-1)}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{s^{(t+1)}}, s^{(t)}, x^{'(t)}_{s^{(t)}}\right) \quad \text{By Lemma 7}$$
$$\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}}, x^{'(t+k)}_{s^{(t+k)}} \perp\!\!\!\perp s^{(t)}, x^{'(t)}_{s^{(t)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{s^{(t+k-1)}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{s^{(t+1)}}\right)$$

by weak union

$$\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \perp\!\!\!\perp s^{(t)}, x^{'(t)}_{s^{(t)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{s^{(t+k)}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{s^{(t+k-1)}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{s^{(t+1)}}\right)$$

by weak union

$$\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \perp\!\!\!\perp s^{(t)}, x^{'(t)}_{s^{(t)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{x^{(t)}}\right) \tag{43}$$

by Lemma 12

Let $x^{'(t)} = x^{(t)}$, we have that:

$$\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}}, x^{'(t+k)}_{s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-1])},s^{([t,t+k-2])}}, x^{'(t+k-1)}_{s^{(t+k-1)}}, \ldots, s^{(t)}_{x^{([t,t+k-1])}}, x^{(t)}_{s^{(t)}}\right)$$

$$\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}}, x^{'(t+k)}_{s^{(t+k)}} \perp\!\!\!\perp s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{s^{(t+k-1)}}, \ldots, s^{(t+1)}, x^{'(t+1)}_{s^{(t+1)}}, s^{(t)}, x^{(t)}_{s^{(t)}}\right)$$

By Lemma 8

$$\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}}, x^{'(t+k)}_{s^{(t+k)}} \perp\!\!\!\perp s^{(t)}, x^{(t)}_{s^{(t)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{s^{(t+k-1)}}, \ldots, s^{(t+1)}, x^{'(t+1)}_{s^{(t+1)}}\right)$$

by weak union

$$\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \perp\!\!\!\perp s^{(t)}, x^{(t)}_{s^{(t)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{s^{(t+k)}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{s^{(t+k-1)}}, \ldots, s^{(t+1)}, x^{'(t+1)}_{s^{(t+1)}}\right)$$

by weak union

$$\Rightarrow \left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \perp\!\!\!\perp s^{(t)}, x^{(t)}_{s^{(t)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x^{'(t+1)}\right) \tag{44}$$

by Lemma 12

By the independence relation 43, we have that:

$$P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{x^{(t)}}, s^{(t)}, x^{'(t)}\right)$$

$$= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{x^{(t)}}, s^{(t)}, x^{'(t)}_{s^{(t)}}\right)$$

By composition

$$= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{x^{(t)}}\right) \tag{45}$$

Similarly, by the independence relation 44, we have that:

$$P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x^{'(t+1)}, s^{(t)}, x^{(t)}\right)$$

$$= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x^{'(t+1)}, s^{(t)}, x^{(t)}_{s^{(t)}}\right)$$

By composition

$$= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x^{'(t+1)}\right) \tag{46}$$

Based on equation 45 and 46, let $x^{'(t)} = x^{(t)}$, we have:

$$P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x^{'(t+1)}\right)$$

$$= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x^{'(t+1)}, s^{(t)}, x^{(t)}\right)$$

$$= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{s^{(t+k)}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{s^{(t+k)}}, \ldots, s^{(t+1)}, x^{'(t+1)}_{s^{(t+1)}}, s^{(t)}, x^{(t)}_{s^{(t)}}\right)$$

By Lemma 12

$$= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t,t+k-1])},s^{([t,t+k-1])}}, x^{'(t+k)}_{s^{(t+k)}}, \ldots, s^{(t+1)}_{x^{([t,t+k-1])},s^{(t)}}, x^{'(t+1)}_{s^{(t+1)}}, s^{(t)}_{x^{([t,t+k-1])}}, x^{(t)}_{s^{(t)}}\right)$$

By Lemma 8

$$= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{s^{(t+k)}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{s^{(t+k)}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{s^{(t+1)}}, s^{(t)}, x^{(t)}_{s^{(t)}} \right)$$

By Lemma 7

$$= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{x^{(t)}}, s^{(t)}, x^{(t)} \right)$$

By Lemma 12

$$= P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{x^{(t)}} \right)$$

Together with equation 45 and 46, this implies that, for any $\forall x^{'(t)} \in X$:

$$\underbrace{P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{x^{(t)}}, s^{(t)}, x^{'(t)} \right)}_{\text{Term 1}}$$

$$= \underbrace{P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x^{'(t+1)} \right)}_{\text{Term 2}} \tag{47}$$

By Lemma 9, given $s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])}}, s^{(t+1)}_{x^{(t)}}, s^{(t)}$, Term 1 equals to:

$$P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{x^{(t)}}, s^{(t)}, x^{'(t)} \right)$$

$$= P\left(s^{(t+k+1)}_{x^{([t,t+k])},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{x^{(t)}}, s^{(t)}, x^{'(t)} \right)$$

$$= P\left(s^{(t+k+1)}_{x^{([t,t+k])},s^{([t,t+k])}}, x^{'(t+k+1)}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x^{'(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x^{'(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+1)}_{x^{(t)}}, x^{'(t+1)}_{x^{(t)}}, s^{(t)}, x^{'(t)} \right)$$

By Lemma 12

Term 2 can be written as:

$$P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x^{'(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x^{'(t+1)} \right)$$

$$= \sum_{s^{'(t)} \in S} \sum_{x^{'(t)}_{s^{'(t)}} \in X} P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x^{'(t+1)}, s^{'(t)}, x^{'(t)}_{s^{'(t)}} \right)$$

$$\cdot P\left(s^{'(t)}, x^{'(t)}_{s^{'(t)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x^{'(t+1)} \right) \tag{48}$$

By Lemma 10, given $s^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+2)}_{x^{(t+1)}}, s^{(t+1)}, x^{'(t)}_{s^{'(t)}}, s^{'(t)}$, we have:

$$P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x^{'(t+1)}, s^{'(t)}, x^{'(t)}_{s^{'(t)}} \right)$$

$$= P\left(s^{(t+k+1)}_{x^{([t+1,t+k])},s^{(t+k)}}, x^{'(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x^{'(t+1)}, s^{'(t)}, x^{'(t)}_{s^{'(t)}} \right)$$

$$= P\left(s^{(t+k+1)}_{x^{([t+1,t+k])},x^{([t+1,t+k])}}, x^{'(t+k+1)}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x^{'(t+1)}, s^{'(t)}, x^{'(t)}_{s^{'(t)}} \right) \quad \text{By Lemma 12}$$

$$= \frac{P\left(s^{(t+k+1)}_{x^{([t+1,t+k])},x^{([t+1,t+k])}}, x^{'(t+k+1)}_{x^{([t+1,t+k])}}, s^{'(t)}, x^{'(t)}_{s^{'(t)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x^{'(t+1)} \right)}{P\left(s^{'(t)}, x^{'(t)}_{s^{'(t)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x^{'(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x^{'(t+1)} \right)} \tag{49}$$

Replace $P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x'^{(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x'^{(t+1)}, s'^{(t)}, x'^{(t)}_{s'^{(t)}}\right)$ in equation 48 with 49:

$$
\begin{aligned}
&P\left(s^{(t+k+1)}_{x^{(t+k)},s^{(t+k)}}, x'^{(t+k+1)}_{s^{(t+k+1)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x'^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right) \\
&= \sum_{s'^{(t)} \in S} \sum_{x'^{(t)}_{s'^{(t)}} \in X} \frac{P\left(s^{(t+k+1)}_{x^{([t+1,t+k])}}, x'^{(t+k+1)}_{x^{([t+1,t+k])}}, s'^{(t)}, x'^{(t)}_{s'^{(t)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right)}{P\left(s'^{(t)}, x'^{(t)}_{s'^{(t)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right)} \\
&\cdot P\left(s'^{(t)}, x'^{(t)}_{s'^{(t)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x'^{(t+1)}, s'^{(t)}, x'^{(t)}_{s'^{(t)}}\right) \\
&= \sum_{s'^{(t)} \in S} \sum_{x'^{(t)}_{s'^{(t)}} \in X} P\left(s^{(t+k+1)}_{x^{([t+1,t+k])}}, x'^{(t+k+1)}_{x^{([t+1,t+k])}}, s'^{(t)}, x'^{(t)}_{s'^{(t)}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right) \\
&= P\left(s^{(t+k+1)}_{x^{([t+1,t+k])}}, x'^{(t+k+1)}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right)
\end{aligned}
$$

Together with equation 47, we have that:

$$
\begin{aligned}
&P\left(s^{(t+k+1)}_{x^{([t,t+k])}}, x'^{(t+k+1)}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x'^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x'^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t)}, x'^{(t)}\right) \\
&= P\left(s^{(t+k+1)}_{x^{([t+1,t+k])}}, x'^{(t+k+1)}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}}, s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x'^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right)
\end{aligned}
$$

□

Lemma 2 (Proof). $P\left(Y^{t+k}_{x^{(t)},x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, x'^{(t+1)}_{x^{(t)}}, s^{(t)}, x'^{(t)}\right)$ can be written as:

$$
\begin{aligned}
&P\left(Y^{t+k}_{x^{(t)},x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, x'^{(t+1)}_{x^{(t)}}, s^{(t)}, x'^{(t)}\right) \\
&= \sum_{s \in S^{k-2+1}} \sum_{x' \in X^{k-2+1}} P\left(y^{(t+k)}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x'^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x'^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t)}, x'^{(t)}\right)\Big|_{x^{([t+1,t+k])}=\pi} \\
&\cdot P\left(s^{(t+k)}_{x^{([t,t+k-1])}}, x'^{(t+k)}_{x^{([t,t+k-1])}} \mid s^{(t+k-1)}_{x^{([t,t+k-2])}}, x'^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t)}, x'^{(t)}\right) \\
&\cdot P\left(s^{(t+k-1)}_{x^{([t,t+k-2])}}, x'^{(t+k-1)}_{x^{([t,t+k-2])}} \mid s^{(t+k-2)}_{x^{([t,t+k-3])}}, x'^{(t+k-2)}_{x^{([t,t+k-3])}}, \ldots, s^{(t)}, x'^{(t)}\right) \\
&\vdots \\
&P\left(s^{(t+2)}_{x^{([t,t+1])}}, x'^{(t+2)}_{x^{([t,t+1])}} \mid s^{(t+1)}_{x^{(t)}}, x'^{(t+1)}_{x^{(t)}}, s^{(t)}, x'^{(t)}\right)
\end{aligned}
$$

where s is defined as a sequence $s^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t+2)}_{x^{([t,t+1])}}$ and x' a sequence

$x'^{(t+k)}_{x^{([t,t+k-1])}}, x'^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, x'^{(t+2)}_{x^{([t,t+1])}}$. By Lemma 13, we have that:

$$
\begin{aligned}
&P\left(y^{(t+k)}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x'^{(t+k)}_{x^{([t,t+k-1])}}, s^{(t+k-1)}_{x^{([t,t+k-2])}}, x'^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t)}, x'^{(t)}\right) \\
&\sum_{x'^{(t+k)}_{x^{([t,t+k])}} \in X} P\left(y^{(t+k)}_{x^{([t,t+k])}}, x'^{(t+k)}_{x^{([t,t+k])}} \mid s^{(t+k)}_{x^{([t,t+k-1])}}, x'^{(t+k)}_{x^{([t,t+k-1])}}, \ldots, s^{(t)}, x'^{(t)}\right) \\
&\sum_{x'^{(t+k)}_{x^{([t,t+k])}} \in X} P\left(y^{(t+k)}_{x^{([t+1,t+k])}}, x'^{(t+k)}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right) \\
&= P\left(y^{t+k}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right)
\end{aligned}
$$

By Lemma 13, for any $\forall k \in \mathbb{Z}_+$:

$$
\begin{aligned}
&P\left(s^{(t+k)}_{x^{([t,t+k-1])}}, x'^{(t+k)}_{x^{([t,t+k-1])}} \mid s^{(t+k-1)}_{x^{([t,t+k-2])}}, x'^{(t+k-1)}_{x^{([t,t+k-2])}}, \ldots, s^{(t)}, x'^{(t)}\right) \\
&= P\left(s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}} \mid s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x'^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right)
\end{aligned}
$$

We then have that:

$$
\begin{aligned}
&P\left(Y^{t+k}_{x^{(t)}, x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, x'^{(t+1)}_{x^{(t)}}, s^{(t)}, x'^{(t)}\right) \\
&= \sum_{s \in S^{k-2+1}} \sum_{x' \in X^{k-2+1}} P\left(y^{t+k}_{x^{([t+1,t+k])}} \mid s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right)\Bigg|_{x^{([t+1,t+k])}=\pi} \\
&\cdot P\left(s^{(t+k)}_{x^{([t+1,t+k-1])}}, x'^{(t+k)}_{x^{([t+1,t+k-1])}} \mid s^{(t+k-1)}_{x^{([t+1,t+k-2])}}, x'^{(t+k-1)}_{x^{([t+1,t+k-2])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right) \\
&\cdot P\left(s^{(t+k-1)}_{x^{([t,t+k-2])}}, x'^{(t+k-1)}_{x^{([t,t+k-2])}} \mid s^{(t+k-2)}_{x^{([t+1,t+k-3])}}, x'^{(t+k-2)}_{x^{([t+1,t+k-3])}}, \ldots, s^{(t+1)}, x'^{(t+1)}\right) \\
&\vdots \\
&P\left(s^{(t+2)}_{x^{([t,t+1])}}, x'^{(t+2)}_{x^{([t,t+1])}} \mid s^{(t+1)}, x'^{(t+1)}\right) \\
&= P\left(Y^{t+k}_{x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}, x'^{(t+1)}\right)
\end{aligned}
$$

which proves the statement. □

Theorem 3 (Proof). We first expand $V^{\pi}(s^{(t)}, x'^{(t)})$ as

$$
\begin{aligned}
&V^{\pi}(s^{(t)}, x'^{(t)}) = \mathbb{E}\left[Y^{(t)}_{x^{(t)}=\pi} + \sum_{k=1}^{\infty} \gamma^{t+k} Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t)}, x'^{(t)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}=\pi} \mid s^{(t)}, x'^{(t)}\right] + \gamma \sum_{s^{(t+1)} \in S} \sum_{x^{t+1} \in X} P(s^{(t+1)}_{x^{(t)}=\pi}, x^{(t+1)}_{x^{(t)}=\pi} \mid s^{(t)}, x'^{(t)}) \\
&\cdot \sum_{k=1}^{\infty} \gamma^{t+k} \mathbb{E}\left[Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t+1)}_{x^{(t)}=\pi}, x^{(t+1)}_{x^{(t)}=\pi}, s^{(t)}, x'^{(t)}\right] \qquad (50)
\end{aligned}
$$

$\mathbb{E}\left[Y^{(t)}_{x^{(t)}=\pi} \mid s^{(t)}, x'^{(t)}\right]$ equals to

$$\begin{aligned}
&\mathbb{E}\left[Y^{(t)}_{x^{(t)}=\pi} \mid s^{(t)}, x'^{(t)}\right] \\
&= \sum_{y^{(t)} \in Y} y^{(t)} P\left(y^{(t)}_{x^{(t)}=\pi} \mid s^{(t)}, x'^{(t)}\right) \\
&= \sum_{y^{(t)} \in Y} y^{(t)} P\left(y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right)\Big|_{x^{(t)}=\pi} \qquad (51) \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right]\Big|_{x^{(t)}=\pi} \qquad (52)
\end{aligned}$$

Eqs. 51 follows from Lemma 3, since $x^{(t)} = \pi(s^{(t)}, x'^{(t)})$ and $s^{(t)}, x'^{(t)}$ are fixed. Similarly, for $P\left(s^{(t+1)}_{x^{(t)}=\pi}, x^{(t+1)}_{x^{(t)}=\pi} \mid s^{(t)}, x'^{(t)}\right)$, we have

$$P\left(s^{(t+1)}_{x^{(t)}=\pi}, x^{(t+1)}_{x^{(t)}=\pi} \mid s^{(t)}, x'^{(t)}\right) = P\left(s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right)\Big|_{x^{(t)}=\pi} \qquad (53)$$

$\mathbb{E}\left[Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t+1)}_{x^{(t)}=\pi}, x^{(t+1)}_{x^{(t)}=\pi}, s^{(t)}, x'^{(t)}\right]$ is equivalent to

$$\begin{aligned}
&\mathbb{E}\left[Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t+1)}_{x^{(t)}=\pi}, x^{(t+1)}_{x^{(t)}=\pi}, s^{(t)}, x'^{(t)}\right] \\
&= \sum_{y^{(t+k)} \in Y} y^{(t+k)} P\left(Y^{(t+k)}_{x^{([t,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}=\pi}, x^{(t+1)}_{x^{(t)}=\pi}, s^{(t)}, x'^{(t)}\right) \\
&= \sum_{y^{(t+k)} \in Y} y^{(t+k)} P\left(Y^{(t+k)}_{x^{(t)}, x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}}\right)\Big|_{x[t]=\pi}
\end{aligned}$$

By Lemma 2, we have that:

$$\begin{aligned}
&\mathbb{E}\left[Y^{(t+k)}_{x^{([t,t+k])}=\pi} \mid s^{(t+1)}_{x^{(t)}=\pi}, x^{(t+1)}_{x^{(t)}=\pi}, s^{(t)}, x'^{(t)}\right] \\
&= \sum_{y^{(t+k)} \in Y} y^{(t+k)} P\left(Y^{t+k}_{x^{(t)}, x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}}, s^{(t)}, x'^{(t)}\right)\Big|_{x^{(t)}=\pi} \\
&= \sum_{y^{(t+k)} \in Y} y^{(t+k)} P\left(Y^{t+k}_{x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}, x^{'(t+1)}\right) \\
&= \mathbb{E}\left[Y^{(t+k)}_{x^{([t+1,t+k])}=\pi} \mid s^{(t+1)}, x^{'(t+1)}\right] \qquad (54)
\end{aligned}$$

Now, substituting Eqs. 52, 53 and 54 back in Eq. 50, we have

$$
\begin{aligned}
&V^{\pi}(s^{(t)}) = \mathbb{E}\left[Y^{(t)}_{x^{(t)}=\pi} \mid s^{(t)}, x'^{(t)}\right] + \gamma \sum_{s^{(t+1)}\in S}\sum_{x^{t+1}\in X} P(s^{(t+1)}_{x^{(t)}=\pi}, x^{(t+1)}_{x^{(t)}=\pi} \mid s^{(t)}, x'^{(t)}) \\
&\cdot \sum_{k=1}^{\infty}\gamma^{t+k}\mathbb{E}\left[Y^{(t+k)}_{x^{[t,t+k]}=\pi} \mid s^{(t+1)}_{x^{(t)}=\pi}, x^{(t+1)}_{x^{(t)}=\pi}, s^{(t)}, x'^{(t)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right]\Big|_{x^{(t)}=\pi} + \gamma \sum_{s^{(t+1)}\in S}\sum_{x^{t+1}\in X} P\left(s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right)\Big|_{x^{(t)}=\pi} \\
&\cdot \sum_{k=1}^{\infty}\gamma^{t+k}\mathbb{E}\left[Y^{(t+k)}_{x^{([t+1,t+k])}=\pi} \mid s^{(t+1)}, x^{(t+1)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right]\Big|_{x^{(t)}=\pi} + \gamma \sum_{s^{(t+1)}\in S}\sum_{x^{t+1}\in X} P\left(s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right)\Big|_{x^{(t)}=\pi} \\
&\cdot \mathbb{E}\left[\sum_{k=1}^{\infty}\gamma^{t+k}Y^{(t+k)}_{x^{([t+1,t+k])}=\pi} \mid s^{(t+1)}, x^{(t+1)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right]\Big|_{x^{(t)}=\pi} + \gamma \sum_{s^{(t+1)}\in S}\sum_{x^{t+1}\in X} P\left(s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right)\Big|_{x^{(t)}=\pi} V^{\pi}(s^{(t+1)}, x^{(t+1)})
\end{aligned}
$$

Similarly, we can expand $Q^{\pi}(s^{(t)}, x'^{(t)})$ as

$$
\begin{aligned}
&Q^{\pi}(s^{(t)}, x'^{(t)}) = \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right] + \gamma \sum_{s^{(t+1)}\in S}\sum_{x^{t+1}\in X} P(s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}) \\
&\cdot \sum_{k=1}^{\infty}\gamma^{t+k}\mathbb{E}\left[Y^{(t+k)}_{x^{(t)},x^{([t+1,t+k])}=\pi} \mid s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}}, s^{(t)}, x'^{(t)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right] + \gamma \sum_{s^{(t+1)}\in S}\sum_{x^{t+1}\in X} P\left(s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right) \\
&\cdot \mathbb{E}\left[\sum_{k=1}^{\infty}\gamma^{t+k}Y^{(t+k)}_{x^{([t+1,t+k])}=\pi} \mid s^{(t+1)}, x^{(t+1)}\right] \\
&= \mathbb{E}\left[Y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right] + \gamma \sum_{s^{(t+1)}\in S}\sum_{x^{t+1}\in X} P\left(s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right)V^{\pi}(s^{(t+1)}, x^{(t+1)})
\end{aligned}
$$

□

References

Heckman, J.J. Randomization and social policy evaluation. In Manski, C. and Garfinkle, I. (eds.), *Evaluations: Welfare and Training Programs*, pp. 201–230. Harvard University Press, Cambridge, MA, 1992.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

Tian, J. and Pearl, J. On the identification of causal effects. Technical Report R-290-L, Department of Computer Science, University of California, Los Angeles, CA, 2003. http://www.cs.iastate.edu/∼jtian/r290-L.pdf.