TECHNICAL REPORT
R-63
First version: Jun 2020
Last revision: Nov 2020

Characterizing Optimal Mixed Policies: Where to Intervene and What to Observe

Sanghack Lee **Elias Bareinboim**
Causal Artificial Intelligence Laboratory
Columbia University
{sanghacklee,eb}@cs.columbia.edu

Abstract

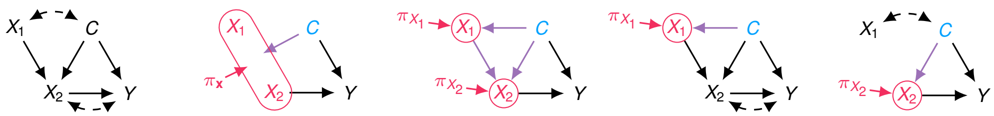
Intelligent agents are continuously faced with the challenge of optimizing a policy based on what they can observe (see) and which actions they can take (do) in the environment where they are deployed. Most policy can be parametrized in terms of these two dimensions, i.e., as a function of what can be seen and done given a certain situation, which we call a *mixed policy*. In this paper, we investigate several properties of the class of mixed policies and provide an efficient and effective characterization, including optimality and non-redundancy. Specifically, we introduce a graphical criterion to identify unnecessary contexts for a set of actions, leading to a natural characterization of non-redundancy of mixed policies. We then derive sufficient conditions under which one strategy can dominate the other with respect to their maximum achievable expected rewards (optimality). This characterization leads to a fundamental understanding of the space of mixed policies and a possible refinement of the agent's strategy so that it converges to the optimum faster and more robustly. One surprising result of the causal characterization is that the agent following a more standard approach—intervening on all intervenable variables and observing all available contexts—may be hurting itself, and will never achieve an optimal performance.

1 Introduction

Agents are deployed in complex and uncertain environments where they are bombarded with high volumes of information and are expected to operate efficiently, safely, and rationally. The discipline of causal inference (CI) offers a compelling set of tools and a language that allows one to reason with the structural invariances present in complex environments [1–5]. Whenever the causal mechanisms of an underlying environment are sufficiently well-understood, the agent can design very precise interventions, bringing a certain desired state of affairs about swiftly and cleanly (e.g., personalized medical treatments, inequality-reducing tax policies). In the field of ML, bandits and reinforcement learning (RL) constitute the *de facto* framework in which agents are designed such that a certain policy is optimized and the corresponding goals can be efficiently achieved [6–8].

There is a growing literature exploring how these two frameworks (RL and CI) are related, and how this understanding can be translated into more efficient decision-making in more challenging and realistic settings. Recently, the more explicit connection between these frameworks has been made by eliciting how causal knowledge—unobserved confounders and the causal relations between actions, contexts, and rewards—can be used to improve decision-making in a variety of settings, including for both interventional [9–11] and counterfactual [12, 13] reasoning (see also [14–17] and [18–21]). Outside more traditional RL, causal inference researchers have embraced the idea of sequential decision making in terms of conditional plans or dynamic treatment regimes, while focusing on, e.g., the identifiability of causal effects from observational data [22–27].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.



(a) $\mathcal{G}$ (b) CB (abstract) (c) $\mathcal{G}_{\mathcal{S}_{\text{CB}}}$ (d) $\mathcal{G}_{\mathcal{S}_d}$ (e) $\mathcal{G}_{\mathcal{S}_e}$

Figure 1: (a) a causal graph, (b) abstract representation of a contextual bandit policy, and (c,d,e) policy-induced graphs. Red circles for the intervened variables and, as a supplement, blue for their non-action contexts and purple for induced edges (i.e., contexts *cause* an action). π nodes are intervention indicators, which will be left implicit throughout the paper.

One of the main tasks in decision-making is to optimize the parameters associated with a specific policy. The scope of each policy is usually fixed, in the sense that the set of actions and contexts are pre-specified, *a priori*. By and large, the literature considers policies with scope that is (1) observational, where the system is allowed to evolve without any intervention; (2) fully experimental, where all the *action* variables are intervened on and all the *context* variables are observed. The former tends to be more common in CI while the latter tends to be more common in RL. A causal understanding of the world gives rise to a rich spectrum of policies with different scopes, allowing agents to choose how to interact with the environment, meaning, which variables to intervene and to observe (as a context). Against this background, we consider exploiting causal relationships for systematic decision making in the context of, so called, *mixed policies*, which consists of a set of decision rules where each rule corresponds to the way an action for an intervenable variable is determined given its contexts.

For concreteness, consider an agent deployed in an environment represented as a *causal graph* $\mathcal{G}$ (Fig. 1a), where C, $\mathbf{X} = \{X_1, X_2\}$, Y represent the context, two action variables, and the reward variable, respectively. Graphically, bidirected edges roughly represent unobserved confounders (UCs, for short) affecting both ends of the arrow. The agent's task is to maximize the reward $\mu_{\boldsymbol{\pi}} \doteq \mathbb{E}_{\boldsymbol{\pi}}[Y]$ under a mixed policy (or simply, policy) $\boldsymbol{\pi} \in \mathbf{\Pi}$, where $\mathbf{\Pi}$ is a mixed policy space. A mixed policy is associated with its *scope*, called *mixed policy scope* (MPS), which specifies the variables the policy are intervening, and the variables taken into account for each intervened variables.

A standard contextual bandit (CB) optimizes a policy $\boldsymbol{\pi}_{\text{CB}}$ (Fig. 1b), a (stochastic) mapping from contexts to actions, which can be equally represented as a pair of decision rules $\boldsymbol{\pi}_{\text{CB}} = \{\pi(x_1|c), \pi(x_2|x_1, c)\}$ (Fig. 1c). Traditionally, the policy is optimized within a restricted space $\mathbf{\Pi}_{\text{CB}}$, characterized by policies following a scope $\mathcal{S}_{\text{CB}} = \{\langle X_1, \{C\}\rangle, \langle X_2, \{X_1, C\}\rangle\}$ that X_1 is determined by C and X_2 is decided based on C and X_1. Unfortunately, the optimal policy $\boldsymbol{\pi}^*_{\text{CB}} \doteq \arg\max_{\boldsymbol{\pi} \in \mathbf{\Pi}_{\text{CB}}} \mu_{\boldsymbol{\pi}}$ can be suboptimal, i.e., $\mu^*_{\mathcal{S}_{\text{CB}}} \doteq \mu_{\boldsymbol{\pi}^*_{\text{CB}}} < \mu^*$ where μ^* is the optimal expected reward. To ground what this means, let every variable be binary and U_1 and U_2, the UCs adjacent to X_1 and X_2, be fair coins and ϵ be a noise over X_1 following $P(\epsilon = 1) = 0.2$. Also, let the unobserved causal mechanisms be specified as $X_1 \leftarrow U_1 \oplus \epsilon$, $C \leftarrow U_1$, $X_2 \leftarrow U_2 \oplus X_1 \oplus C$, and $Y \leftarrow (1 - (X_2 \oplus U_2)) \vee C$, where $\oplus$ is the exclusive-or operator. Since the policy determines X_2 irrelevant to U_2 and the context C is also independent to U_2, we can elicit that $\mu^*_{\mathcal{S}_{\text{CB}}} = 0.75$. In this setting, the best policy is intervening only on X_1 given C, i.e., $\{\pi(x_1|c)\}$ as depicted in Fig. 1d. With $X_1 = C$, the policy suppresses the noise ϵ over X_1 and makes $X_2 = U_2$ so that its optimal expected reward in this environment is 1.0.

In the example of Fig. 1a, if $\{X_1, X_2\}$ are intervenable and $\{C, X_1\}$ can become a context, there are 15 mixed policy scopes. These different modes of interaction can be represented as induced graphs and can be classified based on two desiderata: *non-redundancy* and *optimality*. We explain these desiderata through an illustration (Fig. 2) of the four MPSes $\mathcal{S}_a = \{\}$, $\mathcal{S}_{\text{CB}}$, $\mathcal{S}_d = \{\langle X_1, \{C\}\rangle\}$, and $\mathcal{S}_e = \{\langle X_2, \{C\}\rangle\}$. We annotate their relationships with a superset symbol $\supset$, whether one scope has more actions or contexts than the other, and with a comparison symbol $\geq_\mu$ (or $=_\mu$), whether one's optimal reward is at least as good as (or equal to) the other's in *every* world compatible with a causal graph. The equality forms an equivalence class among scopes with respect to optimal rewards.

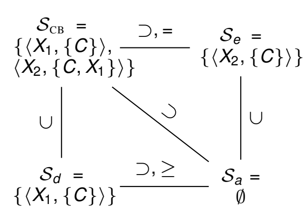


Figure 2: Relationships between the MPSes

Roughly speaking (to be formalized later on), *non-redundancy* refers to the condition of a scope such that removing any of its actions or contexts can negatively affect its maximum performance. In other words, given two scopes $\mathcal{S}$ and $\mathcal{S}'$, if $\mathcal{S} \supsetneq \mathcal{S}'$ and $\mathcal{S} =_{\mu} \mathcal{S}'$, then $\mathcal{S}$ is said to be redundant. For instance, since $\mathcal{S}_c \supset \mathcal{S}_e$ while $\mathcal{S}_c =_{\mu} \mathcal{S}_e$, the CB policy (Fig. 1c) is redundant and the CB agent wastes its resources not only for intervening on X_1 (a redundant action) but also for taking X_1 into account for X_2 (a redundant context). Furthermore, *optimality* of a scope $\mathcal{S}$ represents that there exists no other scope $\mathcal{S}'$ (not in the equivalence class of $\mathcal{S}$) such that $\mathcal{S}' \geq_{\mu} \mathcal{S}$. For example, $\mathcal{S}_d$, when optimized, is at least as good as $\mathcal{S}_a$ (i.e., $\mu^*_{\mathcal{S}_d} \geq \mu^*_{\mathcal{S}_a}$) in every environment, and can outperform it in some environments (i.e., $\mu^*_{\mathcal{S}_d} > \mu^*_{\mathcal{S}_a}$), which demonstrates that $\mathcal{S}_a$ does not meet the optimality criterion. Not every pair of scopes can be comparable: $\mathcal{S}_e$ is not comparable to $\mathcal{S}_a$ nor $\mathcal{S}_d$. After a careful examination, we can indeed able to show that MPSes $\mathcal{S}_c$, $\mathcal{S}_d$, $\mathcal{S}_e$ meet the optimality criterion. Both non-redundancy and optimality are satisfied only by $\mathcal{S}_d$ and $\mathcal{S}_e$ among all 15 scopes. This example demonstrates that an intelligent agent should judiciously intervene on a carefully chosen subset of variables with side information (context) relevant to attaining an optimal reward. More detailed account is given in Appendix A.

Contributions In this work, we investigate mixed policies with respect to their expected rewards. Our contributions are as follows. (i) We developed a graphical criterion that detects the redundancy of contexts relative to a collection of actions taking advantage of properties pertain to optimal mixed policies. (ii) We established sufficient conditions under which one policy scope can outperform another, characterizing the partial order defined over the space of scopes with respect to their maximum expected rewards achievable. We believe these results have practical implications for the design of intelligent agents providing the basis for efficient and effective explorations of the policy space. One fundamental implication of our analysis is that the agent following a standard approach (i.e., intervening and observing whenever possible) may be hurting itself, and, regardless of the number of interactions, will never be able to achieve an optimal performance.

Preliminaries Let us denote a variable by an uppercase letter X, whose value is denoted by its corresponding lowercase letter x. A set of variables will be denoted by a bold uppercase letter $\mathbf{X}$ with its value $\mathbf{x}$. We follow notational conventions from literature on measure theory, algebra of sets, and causal inference. We may use $\dot\cup$, instead of $\cup$, to emphasize the union of two disjoint sets. We use structural causal models (SCMs) [1, Ch. 7] as the semantical framework to represent an underlying environment. An SCM $\mathcal{M}$ is a quadruple $\langle \mathbf{U}, \mathbf{V}, P(\mathbf{U}), \mathbf{F} \rangle$, where $\mathbf{U}$ is a set of exogenous variables determined by factors outside the model following a joint distribution $P(\mathbf{U})$, and $\mathbf{V}$ is a set of endogenous variables whose values are determined following a collection of functions $\mathbf{F} \doteq \{f_i\}_{V_i \in \mathbf{V}}$ such that $V_i \leftarrow f_i(\mathbf{pa}_i, \mathbf{u}_i)$ where $\mathbf{PA}_i \subseteq \mathbf{V} \setminus \{V_i\}$ and $\mathbf{U}_i \subseteq \mathbf{U}$. The observational distribution $P(\mathbf{v})$ is defined as $\sum_{\mathbf{u}} \prod_{V_i \in \mathbf{V}} P(v_i | \mathbf{pa}_i, \mathbf{u}_i) P(\mathbf{u})$. Further, $do(\mathbf{X} = \mathbf{x})$ represents the operation of fixing a set $\mathbf{X}$ to a constant $\mathbf{x}$ regardless of their original mechanisms. Such intervention induces a submodel $\mathcal{M}_{\mathbf{x}}$, which is $\mathcal{M}$ with f_X replaced to x for $X \in \mathbf{X}$. Then, an interventional distribution $P_{\mathbf{x}}(\mathbf{v} \setminus \mathbf{x})$ (or also $P(\mathbf{v} \setminus \mathbf{x} | do(\mathbf{x}))$) follows from $\mathcal{M}_{\mathbf{x}}$, and is such that $P_{\mathbf{x}}(\mathbf{v} \setminus \mathbf{x}) = \sum_{\mathbf{u}} \prod_{V_i \in \mathbf{V} \setminus \mathbf{X}} P(v_i | \mathbf{pa}_i, \mathbf{u}_i) P(\mathbf{u})$.

Graphically, each SCM (model, for short) is associated with a causal diagram $\mathcal{G} = \langle \mathbf{V}, \mathbf{E} \rangle$, where each type of edge represents a different relationship among variables: (i) $X \rightarrow Y$ if X is an argument of f_Y (a direct causal relationship); and (ii) $X \leftrightarrow Y$ if for a maximal subset $\mathbf{W} \subseteq \mathbf{V} \setminus \{X\}$ such that $\mathbf{U}_{\mathbf{W}} \perp\!\!\!\perp \mathbf{U}_X$ and $\mathbf{U}_Y \not\subseteq \mathbf{U}_{\mathbf{W}}$; From the agent's perspective, only the causal graph $\mathcal{G}$ of the environment $\mathcal{M}$ is available, while its reward is validated through $\mathcal{M}$. We operate in the non-parametric setting, where no assumption about the form or shape of the pair $\langle P(\mathbf{U}), \mathbf{F} \rangle$ is made, but for the structural knowledge encoded in $\mathcal{G}$. Whenever not even $\mathcal{G}$ is known, the agent can perform active interventions to learn it; for example, see [28, 29]. We denote by $\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}$ an edge subgraph of $\mathcal{G}$ which removes edges incoming to $\mathbf{X}$ and outgoing from $\mathbf{Z}$. A submodel $\mathcal{M}_{\mathbf{x}}$ can be presented as $\mathcal{G}_{\overline{\mathbf{X}}}$ with $\mathbf{X}$ fixed to $\mathbf{x}$. Hence, causal relationships among other variables are captured in $\mathcal{G} \setminus \mathbf{X}$, which is the subgraph of $\mathcal{G}$ over $\mathbf{V} \setminus \mathbf{X}$. We denote by $\mathcal{G}\langle \mathbf{V}' \rangle$ the *latent projection* of $\mathcal{G}$ onto $\mathbf{V}'$, the causal graph retaining causal relationships among $\mathbf{V}'$ [30]. We adopt familial notation, ch, pa, an, de for children, parents, ancestors, and descendants, respectively, with Ch, Pa, An, De including arguments. Our work utilizes d-separation [31, 32] and do-calculus [33], classic graphical rules to ascertain equalities between distributions (for further details, see Appendix B). Also, the omitted proofs and derivations are provided in Appendix.

2 Mixed Policies: Fundamentals & Basic Results

As discussed in the previous section, a causal understanding of the underlying world helps recognize a broad spectrum of policies with diverse scopes so as for agents to select the mode of interaction. We now formally define the space of mixed policies with the notion of mixed policy scope.

Definition 1 (Mixed Policy Scope (MPS))**.** Let $\mathcal{G}$ be a causal graph, Y be a specific reward variable, $\mathbf{X}^\star \subseteq \mathbf{V}\backslash\{Y\}$ a set of intervenable variables, and $\mathbf{C}^\star \subseteq \mathbf{V}\backslash\{Y\}$ a set of contextualizable variables. A *mixed policy scope* $\mathcal{S}$ is defined as a collection of pairs $\langle X, \mathbf{C}_X\rangle$ such that (i) $X \in \mathbf{X}^\star$, $\mathbf{C}_X \subseteq \mathbf{C}^\star\backslash\{X\}$, and (ii) $\mathcal{G}_\mathcal{S}$ is acyclic, where $\mathcal{G}_\mathcal{S}$ is defined as $\mathcal{G}$ with edges onto X removed and directed edges from $\mathbf{C}_X$ to X added for every $\langle X, \mathbf{C}_X\rangle \in \mathcal{S}$.

For concreteness, given a causal graph $\mathcal{G}$ (Fig. 1a), the observational case is an MPS $\{\}$. An MPS $\mathcal{S}_{\text{CB}} = \{\langle X_1, \{C\}\rangle, \langle X_2, \{X_1, C\}\rangle\}$ induces a graph (Fig. 1c) while $\{\langle X_2, \{C\}\rangle\}$ induces a graph in Fig. 1e. An MPS represents a class of mixed policies that share the same graphical characteristics manifested by $\mathcal{G}_\mathcal{S}$, an induced graph for $\mathcal{M}_{\boldsymbol{\pi}}$.

Definition 2 (Mixed Policy)**.** Given $\langle \mathcal{G}, Y, \mathbf{X}^\star, \mathbf{C}^\star\rangle$ and an SCM $\mathcal{M} \sim \mathcal{G}$ with $\mathfrak{X}_Y \subseteq \mathbb{R}$, a *mixed policy* $\boldsymbol{\pi}$ is a realization of a mixed policy scope $\mathcal{S}$ compatible with the tuple $\boldsymbol{\pi} \doteq \{\pi_{X|\mathbf{C}_X}\}_{\langle X, \mathbf{C}_X\rangle \in \mathcal{S}}$, where $\pi_{X|\mathbf{C}_X} : \mathfrak{X}_X \times \mathfrak{X}_{\mathbf{C}_X} \mapsto [0, 1]$ is a proper probability mapping.

If we consider the MPS $\mathcal{S}_{\text{CB}}$ discussed above, its mixed policy $\boldsymbol{\pi}$ is $\{\pi_{X_1|\{C\}}, \pi_{X_2|\{C,X_1\}}\}$, which is a specific instantiation of the parameters with respect to the corresponding scope. For readability, we may write $\{\pi(x_1|c), \pi(x_2|x_1, c)\}$. Given an underlying SCM $\mathcal{M}$, a mixed policy $\boldsymbol{\pi}$ induces a variant of SCM $\mathcal{M}_{\boldsymbol{\pi}}$ where the function for $X \in \mathbf{X}(\boldsymbol{\pi})$ is replaced by the corresponding $\pi_{X|\mathbf{C}_X}$ (see [34] for a detailed account). We denote by $P_{\boldsymbol{\pi}}$ the joint distribution over the variables from the system under the policy $\boldsymbol{\pi}$. Throughout the paper, $\mathcal{G}$, Y, $\mathbf{C}^\star$, and $\mathbf{X}^\star$ are oftentimes implicit including an underlying SCM $\mathcal{M} \sim \mathcal{G}$ and, thus, $\boldsymbol{\Pi}$, as well.

Expected Reward We define the expected reward of a mixed policy. To begin with, we define intervened variables $\mathbf{X}(\mathcal{S}) \doteq \{X \mid \langle X, \mathbf{C}_X\rangle \in \mathcal{S}\}$ and active contexts $\mathbf{C}(\mathcal{S}) \doteq \bigcup_{X\in\mathbf{X}(\mathcal{S})} \mathbf{C}_X$. Similarly, given $\boldsymbol{\pi} \sim \mathcal{S}$ (a mixed policy following the MPS), $\mathbf{X}(\boldsymbol{\pi}) \doteq \mathbf{X}(\mathcal{S})$ and $\mathbf{C}(\boldsymbol{\pi}) \doteq \mathbf{C}(\mathcal{S})$. Let $\mathbf{C}^- = \mathbf{C}(\boldsymbol{\pi})\backslash\mathbf{X}(\boldsymbol{\pi})$ be the *non-action* contexts. Then, the expected reward for $\boldsymbol{\pi}$ can be expressed as, with $\mathbf{x}$ simply denoting the value of $\mathbf{X}(\mathcal{S})$,

$$\mu_{\boldsymbol{\pi}} = \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\mathbf{x}}(y, \mathbf{c}^-) \prod_{X\in\mathbf{X}(\boldsymbol{\pi})} \pi(x|\mathbf{c}_x). \tag{1}$$

The expression separates the atomic interventional probability (first factor), which is inherent to the underlying world and not affected by the policy $\boldsymbol{\pi}$, from the likelihood of a specific intervention given contexts (second factor), which is optimizable and defined by $\boldsymbol{\pi}$. The expected reward can also be written focusing only on a subset of intervened variables. Given $\mathbf{X}' \subseteq \mathbf{X}(\boldsymbol{\pi})$, let $\mathbf{C}' = \bigcup_{X\in\mathbf{X}'} \mathbf{C}_X\backslash\mathbf{X}'$, and $Q' = P_{\boldsymbol{\pi}\backslash\mathbf{X}'}$ where $\boldsymbol{\pi}\backslash\mathbf{X}'$ represents $\boldsymbol{\pi}$ with decision rules over $\mathbf{X}'$ removed. Then, $\mu_{\boldsymbol{\pi}} = \sum_{y,\mathbf{x}',\mathbf{c}'} y Q'_{\mathbf{x}'}(y, \mathbf{c}') \prod_{X\in\mathbf{X}'} \pi(x|\mathbf{c}_x)$. This expression, which hides the details of uninteresting actions and contexts, is the building block to characterize mixed policies.

Optimality and deterministic mixed policy A mixed policy $\boldsymbol{\pi}$ is said to be *optimal* in the given environment if and only if $\mu_{\boldsymbol{\pi}} = \mu^* \doteq \max_{\boldsymbol{\pi}'\in\boldsymbol{\Pi}} \mu_{\boldsymbol{\pi}'}$. Restricting our attention to $\boldsymbol{\Pi}_\mathcal{S} \doteq \{\boldsymbol{\pi} \in \boldsymbol{\Pi} \mid \boldsymbol{\pi} \sim \mathcal{S}\}$, we define $\mu^*_\mathcal{S} \doteq \max_{\boldsymbol{\pi}'\in\boldsymbol{\Pi}_\mathcal{S}} \mu_{\boldsymbol{\pi}'}$, an optimal policy $\boldsymbol{\pi}$ *with respect to* $\mathcal{S}$. We call a mixed policy *deterministic* if, for every $\pi_{X|\mathbf{C}_X} \in \boldsymbol{\pi}$, X is determined by a function of $\mathbf{C}_X$.

Proposition 1. *Given a mixed policy scope, there always exists a deterministic mixed policy, which is optimal with respect to the given scope.*

Not surprisingly at this point, a stochastic policy is no better than the best deterministic policy [35–37]. Still, this result has a particular importance to the treatment provided here due to its implications to the d-separation criterion [38], which will be instrumental and discussed in depth in Sec. 3.1. Another key implication is shown next.

Proposition 2 (Separation of Actions and Contexts)**.** *Given an MPS $\mathcal{S}$, there always exists a deterministic mixed policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ such that $\mathbf{X}(\boldsymbol{\pi})$ and $\mathbf{C}(\boldsymbol{\pi})$ are disjoint and $\mu_{\boldsymbol{\pi}} = \mu^*_\mathcal{S}$.*

A deterministic policy gives rise to the autonomy of each action allowing them to be determined only by *non-action* contexts. For concreteness, consider the example shown in Fig. 3a. A mixed policy (Fig. 3b) includes X_2 listening to X_1, which enables systematic coordination between X_1 and X_2. The proposition implies that X_2 can rather listen to C (which is the context of X_1) directly (Fig. 3d). Further, in Fig. 3c, X_2 utilizes both X_1 and C. However, it is sufficient to make use of only C. By noting that the policy relative to Fig. 3d can achieve optimality, while relying on lesser information than the one relative to Fig. 3c, we investigate how to capture non-redundancy within MPSes.

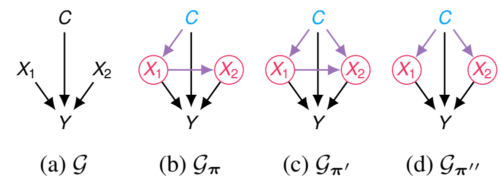


Figure 3: Given a causal graph (a), three induced graphs (b,c,d) for different mixed policies where (d) the separation is demonstrated.

3 Non-Redundant Mixed Policy

Optimizing a mixed policy involves assessments of the effectiveness of its scope so that an agent can avoid intervening or observing on unnecessary actions or contexts. Here, we define and characterize non-redundancy of MPS. We say $\mathcal{S}$ subsumes $\mathcal{S}'$, denoted by $\mathcal{S}' \subseteq \mathcal{S}$, if $\mathbf{X}(\mathcal{S}') \subseteq \mathbf{X}(\mathcal{S})$ and $\mathbf{C}'_X \subseteq \mathbf{C}_X$, for every $\langle X, \mathbf{C}'_X \rangle \in \mathcal{S}'$. Further, we denote by $\boldsymbol{\pi}' \subseteq \boldsymbol{\pi}$, where $\boldsymbol{\pi}' \sim \mathcal{S}'$ and $\boldsymbol{\pi} \sim \mathcal{S}$ if $\pi'(x|\mathbf{c}'_x) = \sum_{\mathbf{c}''_x} \pi(x|\mathbf{c}_x) P_{\boldsymbol{\pi}}(\mathbf{c}''_x|\mathbf{c}'_x)$, for every $X \in \mathbf{X}(\mathcal{S}')$ where $\mathbf{C}''_X = \mathbf{C}_X \setminus \mathbf{C}'_X$.

Definition 3. Given $\langle \mathcal{G}, Y, \mathbf{X}^\star, \mathbf{C}^\star \rangle$, an MPS $\mathcal{S}$ is said to be *non-redundant* if there exists an SCM $\mathcal{M} \sim \mathcal{G}$ and $\boldsymbol{\pi} \sim (\mathcal{S}, \mathcal{M})$ such that $\mu_{\boldsymbol{\pi}} \neq \mu_{\boldsymbol{\pi}'}$ for every $\boldsymbol{\pi}' \subsetneq \boldsymbol{\pi}$.

The constraint on $\boldsymbol{\pi}'$ ensures that the definition of non-redundancy of MPS is focused on the differences in actions or contexts while the behavior (i.e., $\pi(\cdot|\cdot)$) remains the same—$\pi'(x|\mathbf{c}'_x) = Q(x|\mathbf{c}'_x)$ if $\mathbf{C}'_X \neq \mathbf{C}_X$ and $Q(x|\mathbf{c}_x) = \pi'(x|\mathbf{c}_x) = \pi(x|\mathbf{c}_x)$, otherwise. Hence, the constraint provides a basis to characterize non-redundancy of MPS utilizing well-established graphical criteria.

Theorem 1. *Let $\mathcal{S} = \{\langle X, \mathbf{C}_X \rangle\}_{X \in \mathbf{X}}$ be an MPS and let $\mathcal{H} = \mathcal{G}_{\mathcal{S}}$. $\mathcal{S}$ is non-redundant if and only if (i) $\mathbf{X} \subseteq an(Y)_{\mathcal{H}}$ and (ii) $(C \not\perp\!\!\!\perp Y \mid \mathbf{C}_X \setminus \{C\})$ in $\mathcal{H} \setminus \{X\}$, for every $X \in \mathbf{X}$ and $C \in \mathbf{C}_X$.*

The condition (i) can be seen through rule 3 of do-calculus such that the change of the mechanism of X has a consequence on the reward Y.[1] The condition (ii) coincides with rule 2 of do-calculus $Q(y|x, \mathbf{c}_x \setminus \{c\}) = Q_x(y|\mathbf{c}_x \setminus \{c\})$, where $Q = P_{\boldsymbol{\pi}}$. In words, the path from C to Y can be concatenated with $X \leftarrow C$ to form a back-door path from X to Y.[2] Consider the example in Fig. 4 where both X_1 and X_2 are ancestors of Y (condition (i)). Regarding condition (ii), C_1 being adjacent to Y, C_2 having a path to Y through X_2, and C_3 being connected to Y as $C_3 \rightarrow C_2 \rightarrow X_1 \rightarrow Y$ demonstrate that every context is non-redundant. We provide an efficient algorithm for obtaining a unique, maximal, non-redundant MPS (nr-mps, Alg. 2) of a given MPS in Appendix E.

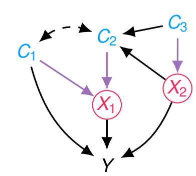


Figure 4: A non-redundant MPS

3.1 Non-Redundancy under Optimality

Non-redundancy of MPS (Def. 3) based on a stringent constraint imposed on $\boldsymbol{\pi}'$ is insufficient to understand, e.g., whether a context of an action would be still relevant even when $\boldsymbol{\pi} \sim \mathcal{S}$ is fully-optimized. Hence, we characterize the non-redundancy of MPS under optimality, which has practical implications to an agent adapting its suboptimal policy. Recall Fig. 3c where X_2 listens to X_1 as context. We showed that the dependence is vanished under the optimality (Fig. 3d). That is, the agent would better avoid learning $\pi(x_2|c, x_1)$ at the beginning, but optimize $\pi(x_2|c)$ instead.

Definition 4 (Non-Redundacy under Optimality (NRO))**.** Given $\langle \mathcal{G}, Y, \mathbf{X}^\star, \mathbf{C}^\star \rangle$, an MPS $\mathcal{S}$ is said to be *non-redundant under optimality* if there exists an SCM $\mathcal{M}$ compatible with $\mathcal{G}$ such that $\mu^*_{\mathcal{S}} > \mu^*_{\mathcal{S}'}$ for every strictly subsumed MPS $\mathcal{S}' \subsetneq \mathcal{S}$, i.e., $\exists_{\mathcal{M} \sim \mathcal{G}} \forall_{\mathcal{S}' \subsetneq \mathcal{S}} (\mu^*_{\mathcal{S}} > \mu^*_{\mathcal{S}'})$.

[1]This condition was leveraged in the atomic interventions case to establish minimality [16, 17]; see also [39].

[2]The relevance of contextual information has been discussed in the influence diagrams literature [40, 20]. More recently, this condition was used in the case of singleton decisions (i.e., $|\mathbf{X}(\mathcal{S})| = 1$), see [41, 42].

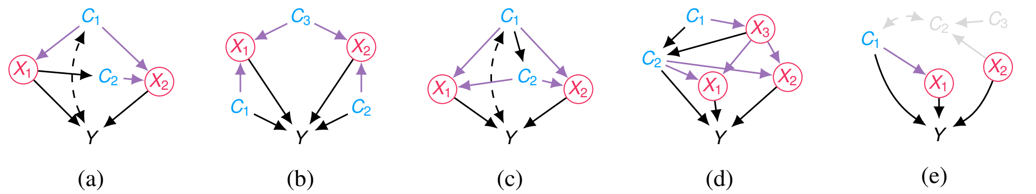


Figure 5: Causal graphs exemplifying redundancies of (a) $C_2 \to X_2$ by deterministic relationships, edges to X_1 and X_2 from (b) C_3, (c) C_2, (d) X_3 due to marginally or conditionally fixable contexts; (e) represents a maximal, non-redundant MPS under an optimal condition for Fig. 4.

We will investigate a criterion more general than Thm. 1—whether, for a *set* of actions $\mathbf{X}' \subseteq \mathbf{X}^\star$, a *set* of contexts $\mathbf{C}' \subsetneq \mathbf{C}_{\mathbf{X}'} \backslash \mathbf{X}'$ are relevant while taking account of deterministic relationships (Prop. 1). One approach is to characterize an opposite condition, i.e., $\mu^*_{\mathcal{S}} = \mu^*_{\mathcal{S}'}$ for $\mathcal{S}' \subsetneq \mathcal{S}$, as follows.

Proposition 3. *Given an MPS $\mathcal{S}$, let $\mathbf{X}' \subseteq \mathbf{X}(\mathcal{S})$ and $\mathbf{C}' \subsetneq \mathbf{C}_{\mathbf{X}'} \backslash \mathbf{X}'$ be actions and non-action contexts of interest, respectively, and let $Q' = P_{\boldsymbol{\pi} \backslash \mathbf{X}'}$. Given a mixed policy $\boldsymbol{\pi} \sim \mathcal{S}$ optimal with respect to $\mathcal{S}$, if there exist decision rules $\{\pi'(x|(\mathbf{x}' \,\dot\cup\, \mathbf{c}') \cap \mathbf{c}_x)\}_{X \in \mathbf{X}'}$ such that*

$$\mu^*_{\mathcal{S}} = \sum_{y, \mathbf{c}', \mathbf{x}'} y Q'_{\mathbf{x}'}(y, \mathbf{c}') \prod_{X \in \mathbf{X}'} \pi'(x|(\mathbf{x}' \,\dot\cup\, \mathbf{c}') \cap \mathbf{c}_x), \tag{2}$$

then, $\mathbf{C}_{\mathbf{X}'} \backslash (\mathbf{C}' \,\dot\cup\, \mathbf{X}')$ are jointly redundant to $\mathbf{X}'$ under optimality, and $\mathcal{S}' \doteq (\mathcal{S} \backslash \mathbf{X}') \cup \{\langle X, \mathbf{C}' \cap \mathbf{C}_X \rangle\}_{X \in \mathbf{X}'}$ satisfies $\mu^*_{\mathcal{S}} = \mu^*_{\mathcal{S}'}$.

Proof. This follows from the definition of non-redundancy under optimality and expected reward. □

To closely investigate a sufficient condition for Prop. 3, we start by discussing the implication of deterministic relationships, which characterizes an optimal policy, on the d-separation criterion. The graphical criterion handles deterministic mechanisms (i.e., *conditional intervention*) by excluding them appearing as common causes, e.g., $\leftarrow X \rightarrow$, in a trail [38]. This corresponds to adding those *implied* variables to the conditionals, in which we explicitly represent with an operation $\lceil \cdot \rceil$ for clarity. Given conditionals $\mathbf{Z}$, the implied variables with respect to $\mathbf{Z}$ is computed as follows. Initially setting $\lceil \mathbf{Z} \rceil \leftarrow \mathbf{Z}$, we update $\lceil \mathbf{Z} \rceil \leftarrow \lceil \mathbf{Z} \rceil \cup \{X \in \mathbf{X}(\mathcal{S}) \mid \mathbf{C}_X \subseteq \lceil \mathbf{Z} \rceil\}$ until it is converged. Then, given $\boldsymbol{\pi} \sim \mathcal{S}$, an optimal policy with respect to $\mathcal{S}$, a conditional independence statement $\mathbf{W} \perp\!\!\!\perp \mathbf{T} \mid \mathbf{Z}$ for $P_{\boldsymbol{\pi}}$ becomes $\mathbf{W} \perp\!\!\!\perp \mathbf{T} \mid \lceil \mathbf{Z} \rceil$ in $\mathcal{G}_{\boldsymbol{\pi}}$. Consider $C \in \mathbf{C}_X$ for some $X \in \mathbf{X}(\boldsymbol{\pi})$. The redundancy of a single context can now be expressed as $(C \perp\!\!\!\perp Y \mid \lceil \mathbf{C}_X \backslash \{C\} \rceil)_{\mathcal{H} \backslash \{X\}}$. For instance, in Fig. 5a, C_2 as a context of X_2 is independent to Y given C_1 in a graph with X_2 removed since $\lceil \{C_1\} \rceil = \{C_1, X_1\}$ and $C_2 \leftarrow X_1 \rightarrow Y$ is not a valid trail anymore. Hence, C_2 is removable from $\mathbf{C}_{X_2}$.

Next, we illustrate contexts that unnecessarily induce correlations among actions without any implications on Y (see Appendix E.1 [43] for the derivations of the examples in Fig. 5). In Fig. 5b, both X_1 and X_2 utilize C_3 as their contexts. where $\mu_{\boldsymbol{\pi}} = \mathbb{E}_{c_3}[\mathbb{E}_{\boldsymbol{\pi}}[y|c_3]]$. Since there exists $c_3^* = \arg\max_{c_3 \in \mathfrak{X}_{C_3}} \mathbb{E}_{\boldsymbol{\pi}}[y|c_3]$, we can derive that $\mu_{\boldsymbol{\pi}} \leq \mathbb{E}_{\boldsymbol{\pi}}[y|c_3^*]$. Given that c_3^* is merely a constant, new decision rules $\pi'(x_i|c_1) \doteq Q(x_i|c_1, c_3^*) = \pi(x_i|c_1, c_3^*)$ for $i \in \{1, 2\}$ yield the same optimal reward. A more sophisticated example is shown in Fig. 5c where a redundant context can be *fixed* conditioned on the remaining contexts. The expected reward is expressed as

$$\mu_{\boldsymbol{\pi}} = \textstyle\sum_{c_1, c_2} Q(c_2|c_1) \big(\sum_{y, \mathbf{x}} y P_{\mathbf{x}}(y, c_1) \pi(x_1|c_1, c_2) \pi(x_2|c_1, c_2) \big) = \sum_{c_1, c_2} Q(c_2|c_1) \mu_{\boldsymbol{\pi}}(c_1, c_2).$$

Let c_2^* be a function taking c_1 such that $c_2^*(c_1) = \arg\max_{c_2} \mu_{\boldsymbol{\pi}}(c_1, c_2)$ for $c_1 \in \mathfrak{X}_{C_1}$. Then,

$$\leq \textstyle\sum_{c_1} \mu_{\boldsymbol{\pi}}(c_1, c_2^*(c_1)) = \sum_{y, c_1, \mathbf{x}} y P_{\mathbf{x}}(y, c_1) \pi(x_1|c_1, c_2^*(c_1)) \pi(x_2|c_1, c_2^*(c_1))).$$

By incorporating c_2^* into $\boldsymbol{\pi}$, we can introduce $\boldsymbol{\pi}'$ such that $\pi(x_1|c_1, c_2^*(c_1)) \pi(x_2|c_1, c_2^*(c_1)) = \pi'(x_1|c_1) \pi'(x_2|c_1)$, satisfying Prop. 3. The variables being fixed are not necessarily conditioned on its parents (or ancestors). An example conditioning on its child is illustrated in (Fig. 5d) where we can elicit, e.g., $\pi(x_1|x_3^*(c_2), c_2) \doteq \pi'(x_1|c_2)$.

Given a general causal graph and an MPS, the aforementioned phenomena can be arbitrarily complex. We present a general criterion to test such redundancies by first proposing a lemma to obtain an intermediate expression. Let $\mathbf{V}_{\prec V}$ denote a subset of $\mathbf{V}$ preceding $V \in \mathbf{V}$ given an order $\prec$ over $\mathbf{V}$.

Lemma 1. *Given an MPS $\mathcal{S}$, which satisfies non-redundancy (Thm. 1), let $\mathbf{X}' \subseteq \mathbf{X}(\mathcal{S})$, actions of interest, $\mathbf{C}' \subsetneq \mathbf{C}_{\mathbf{X}'} \backslash \mathbf{X}'$. non-action contexts of interest. If there exists a subset of exogenous variables $\mathbf{U}'$ in $\mathcal{G}_{\mathcal{S}}$, a subset of endogenous variables $\mathbf{Z}$ in $\mathcal{G}_{\mathcal{S}}$ that disjoints with $\mathbf{C}' \dot\cup \mathbf{X}'$ and subsumes $\mathbf{C}_{\mathbf{X}'} \backslash (\mathbf{C}' \dot\cup \mathbf{X}')$, and an order $\prec$ over $\mathbf{V}' \doteq \mathbf{C}' \dot\cup \mathbf{X}' \dot\cup \mathbf{Z}$ such that*

1. *$(Y \perp\!\!\!\perp \boldsymbol{\pi}_{\mathbf{X}'} \mid \lceil \mathbf{X}' \dot\cup \mathbf{C}' \rceil)_{\mathcal{G}_{\mathcal{S}}}$,*
2. *$(C \perp\!\!\!\perp \boldsymbol{\pi}_{\mathbf{X}'_{\prec C}}, \mathbf{Z}_{\prec C}, \mathbf{U}' \mid \lceil (\mathbf{X}' \dot\cup \mathbf{C}')_{\prec C} \rceil)_{\mathcal{G}_{\mathcal{S}}}$ for every $C \in \mathbf{C}'$, and*
3. *$\mathbf{V}'_{\prec X}$ is disjoint with $de(X)_{\mathcal{G}_{\mathcal{S}}}$ and subsumes $pa(X)_{\mathcal{G}_{\mathcal{S}}}$ for every $X \in \mathbf{X}'$,*

then, the expected reward for $\boldsymbol{\pi}$, a deterministic policy optimal with respect to $\mathcal{S}$, can be written as

$$\mu_{\boldsymbol{\pi}} = \sum_{y,\mathbf{c}',\mathbf{x}'} y Q'_{\mathbf{x}'}(y, \mathbf{c}') \sum_{\mathbf{u}',\mathbf{z}} Q(\mathbf{u}') \prod_{Z \in \mathbf{Z}} Q(z | \mathbf{v}'_{\prec z}, \mathbf{u}') \prod_{X \in \mathbf{X}'} \pi(x | \mathbf{c}_x). \tag{3}$$

Lemma 1 offers a sufficient condition for obtaining the intermediate expression (Eq. (3)) for us to rewrite $\mu_{\boldsymbol{\pi}}$ as proposed in Prop. 3. The order $\prec$ dictates how the chain rule is applied in deriving the expression and what variables will appear as conditional for the probability terms. The first two conditions are relevant to separate $Q'_{\mathbf{x}'}(y, \mathbf{c}')$ from the rest. The third one is to obtain $\pi(x|\mathbf{c}_x)$ from $Q(x|\mathbf{v}'_{\prec x}, \mathbf{u}')$. We revisit Fig. 4 where we will ultimately show that, indeed C_2 and C_3 are *redundant contexts under optimality*. Given $\mathbf{C}' = \{C_1\}$ and $\mathbf{X}' = \{X_1, X_2\}$, consider $\mathbf{Z} = \{C_2, C_3\}$, $\mathbf{U}' = \emptyset$, and order $\prec = \langle C_3, C_1, X_2, C_2, X_1 \rangle$. We can derive the following expression for the expected reward (with subscripts concatenated),

$$\mu^*_{\mathcal{S}} = \textstyle\sum_{y,\mathbf{x},c_1} y Q'_{\mathbf{x}}(y|c_1) \sum_{c_{23}} Q(c_{123}, \mathbf{x}) \tag{4}$$

$$= \textstyle\sum_{y,\mathbf{x},c_1} y Q'_{\mathbf{x}}(y|c_1) \sum_{c_{23}} Q(c_3) Q(c_1|c_3) Q(x_2|c_{13}) Q(c_2|c_{13}, x_2) Q(x_1|c_{123}, x_2) \tag{5}$$

$$= \textstyle\sum_{c_3} Q(c_3) \sum_{y,\mathbf{x},c_1} y Q'_{\mathbf{x}}(y, c_1) \sum_{c_2} Q(c_2|c_{13}, x_2) \pi(x_2|c_3) \pi(x_1|c_{12}). \tag{6}$$

We now provide a sufficient condition that further polishes the intermediate expression from Lemma 1 so as to represent it as the expected reward for a smaller MPS than the original one, fulfilling the condition presented in Prop. 3.

Theorem 2. *Let $\mathbf{U}'$, $\mathbf{Z}$, and $\prec$ satisfy Lemma 1. For $Z \in \mathbf{Z}$, let $\mathbf{V}_Z$ be a minimal subset of $\mathbf{V}'_{\prec Z} \cup \mathbf{U}'$ such that $Q(Z \mid \mathbf{V}_Z) = Q(Z \mid \mathbf{V}'_{\prec Z}, \mathbf{U}')$. We define $\mathsf{fix}(\mathbf{T})$ with respect to $\{\langle Z, \mathbf{V}_Z \rangle\}_{Z \in \mathbf{Z}}$, that is, with $\hat{\mathbf{T}} \doteq \lceil \mathbf{T} \rceil \cup \{Z \in \mathbf{Z} \mid \mathbf{V}_Z \backslash \mathbf{U}' \subseteq \lceil \mathbf{T} \rceil\}$, $\mathsf{fixed}(\mathbf{T})$ is $\mathbf{T}$ if $\mathbf{T} = \hat{\mathbf{T}}$ and $\mathsf{fixed}(\hat{\mathbf{T}})$, otherwise. If $\mathsf{fixed}(\mathbf{C}_X \backslash \mathbf{Z}) \supseteq \mathbf{C}_X$ for $X \in \mathbf{X}'$, then, $\mathcal{S}' \doteq (\mathcal{S} \backslash \mathbf{X}') \cup \{\langle X, \mathbf{C}_X \backslash \mathbf{Z} \rangle\}_{X \in \mathbf{X}'}$ satisfies $\mu^*_{\mathcal{S}'} = \mu^*_{\mathcal{S}}$.*

Thm. 2 provides a condition where Eq. (3) can be transformed to $\mu^*_{\mathcal{S}'}$. To do so, it examines whether terms $Q(z|\mathbf{v}_z)$ can be removed by fixing Z to z^* conditional on $\mathbf{v}_z$ in connection with the context to be removed. That is,

$$\mu^*_{\mathcal{S}} = \overbrace{\sum_{\mathbf{u}'} Q(\mathbf{u}')}^{\text{marginally fixable}} \sum_{y,\mathbf{c}',\mathbf{x}'} y \underbrace{Q'_{\mathbf{x}'}(y, \mathbf{c}')}_{\text{irrelevant to } \mathbf{Z}} \overbrace{\sum_{\mathbf{z}} \underbrace{\prod_{Z \in \mathbf{Z}} Q(z|\mathbf{v}_z)}_{\text{defines dependency}}}^{\text{to fix conditionally}} \prod_{X \in \mathbf{X}'} \underbrace{\pi(x| \overbrace{\mathbf{c}_x \backslash \mathbf{z}}^{\text{given}}, \overbrace{\mathbf{c}_x \cap \mathbf{z}}^{\text{to infer}})}_{\text{to be } \pi'(x|\mathbf{c}_x \backslash \mathbf{z})}, \tag{7}$$

We explain the theorem by deriving further from Eq. (6). C_3 can be fixed to a constant c_3^* so that,

$$\leq \textstyle\sum_{y,\mathbf{x},c_1} y Q'_{\mathbf{x}}(y, c_1) \sum_{c_2} Q(c_2|c_1, c_3^*, x_2) \pi(x_2|c_3^*) \pi(x_1|c_1, c_2). \tag{8}$$

There exists $x_2^* \in \mathfrak{X}_{X_2}$ where we can replace $\pi(x_2|c_3^*)$ with $\pi'(x_2)$ such that $\pi'(x_2^*) = 1$.

$$\leq \textstyle\sum_{y,\mathbf{x},c_1} y Q'_{\mathbf{x}}(y, c_1) \sum_{c_2} Q(c_2|c_1, c_3^*, x_2^*) \pi'(x_2) \pi(x_1|c_1, c_2). \tag{9}$$

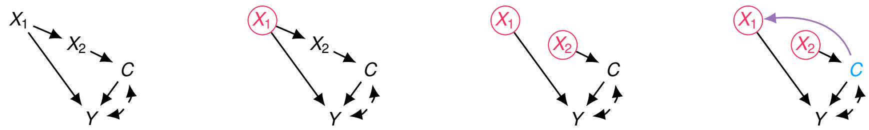
These steps first correspond to checking $\mathsf{fixed}(\emptyset) = \{C_3, X_2\}$ and, then, safely replacing the decision rule for X_2 by eliminating C_3 from its context since $\mathsf{fixed}(\mathbf{C}_{X_2} \backslash \mathbf{Z}) \supseteq \mathbf{C}_{X_2} = \{C_3\}$. Next, the optimal c_2 is determined with respect to c_1, i.e., $Q(c_2|c_1, c_3^*, x_2^*)$, where we can replace $\pi(x_1|c_1, c_2^*(c_1))$ by $\pi'(x_1|c_1)$,

$$= \textstyle\sum_{c_1,c_2} Q(c_2|c_1, c_3^*, x_2^*) \sum_{y,\mathbf{x}} y Q'_{\mathbf{x}}(y, c_1) \pi'(x_2) \pi(x_1|c_1, c_2) \tag{10}$$

$$\leq \textstyle\sum_{y,\mathbf{x},c_1} y Q'_{\mathbf{x}}(y, c_1) \pi'(x_2) \pi(x_1|c_1, c_2^*(c_1)) \tag{11}$$

$$= \textstyle\sum_{y,\mathbf{x},c_1} y Q'_{\mathbf{x}}(y, c_1) \pi'(x_1|c_1) \pi'(x_2) = \mu^*_{\mathcal{S}'}. \tag{12}$$

These steps correspond to checking $\mathsf{fixed}(\mathbf{C}_{X_1} \backslash \mathbf{Z}) = \mathsf{fixed}(\{C_1\}) = \{C_1, C_3, X_2, C_2\} \supseteq \{C_1, C_2\}$ for X_1. Since $\mu^*_{\mathcal{S}'} \leq \mu^*_{\mathcal{S}}$ by the existence of $\boldsymbol{\pi} \in \mathcal{S}$ that can emulate $\boldsymbol{\pi}' \in \mathcal{S}'$, and $\mu^*_{\mathcal{S}'} \geq \mu^*_{\mathcal{S}}$ by the derivation (Eq. (12)), we can conclude that $\mu^*_{\mathcal{S}'} = \mu^*_{\mathcal{S}}$. As a consequence, MPS $\mathcal{S}$ is *not non-redundant under optimality* due to the ineffective contexts $\{C_2, C_3\}$ with respect to $\{X_1, X_2\}$.



(a) $\mathcal{G} = \mathcal{G}_\emptyset$ (b) $\mathcal{G}_{\mathcal{S}'}$ (c) $\mathcal{G}_{\mathcal{S}''}$ (d) $\mathcal{G}_{\mathcal{S}'''}$

Figure 6: A causal graph $\mathcal{G}$ (a) and its induced graphs (a,b,c,d) where the mixed policy scope on the right is better than or equal to the one on the left with respect to their optimal rewards.

4 A Partial Order over Mixed Policies and Possible-Optimality

Equipped with the notion of non-redundancy under optimality (NRO, Def. 4), an agent can more efficiently optimize its policy than relying on generic non-redundancy (Def. 3). Yet, an important question is whether an MPS is worth to explore for an agent to converge to an optimal policy. Consider for an instance, see Figs. 6a to 6d which represent various NRO MPSes. However, even without interacting with an environment, we can claim $\mu \leq \mu^*_{\mathcal{S}'} \leq \mu^*_{\mathcal{S}''} \leq \mu^*_{\mathcal{S}'''}$, that is, the next MPS is better than or equal to (simply *better* or *improved* hereinafter) the one regarding their optimal expected rewards in *any* model: First, $\mu \leq \mu^*_{\mathcal{S}'}$ since there exists an optimal X_1 value, x_1^*; Next, $\mu^*_{\mathcal{S}'} \leq \mu^*_{\mathcal{S}''}$, there exists an optimal X_2 value, and can be determined without conditional on X_1, which is implied; Finally, $\mu^*_{\mathcal{S}''} \leq \mu^*_{\mathcal{S}'''}$ since X_1 can better behave by taking an effective context C into account. Therefore, the agent can only optimize parameters involving $\mathcal{S}'''$ (Fig. 6d) to obain an effective policy. Against this background, we characterize such a partial order over the space of MPSes with respect to their maximum expected rewards achievable: when one MPS is better than the other. To begin a formal discussion, we introduce *possible-optimality* of MPS.

Definition 5 (Possibly-Optimal MPS)**.** Given $\langle \mathcal{G}, \mathbf{X}^\star, \mathbf{C}^\star, Y \rangle$, let $\mathbb{S}$ be a set of NRO MPSes. An MPS $\mathcal{S} \in \mathbb{S}$ is said to be *possibly-optimal* if there exists $\mathcal{M} \sim \mathcal{G}$ such that $\mu^*_{\mathcal{S}} > \max_{\mathcal{S}' \in \mathbb{S} \setminus \{\mathcal{S}\}} \mu^*_{\mathcal{S}'}$.

In the partial order sense, POMPSes are the maximal elements among NRO MPSes. To study the partial order, we present two operations which take an MPS and return an improved MPS: (i) adding observations for existing actions and (ii) adding new interventions. These two operations offer sufficient conditions for identifying non-POMPSes.

Proposition 4. *Given an MPS $\mathcal{S}$ and $X \in \mathbf{X}(\mathcal{S})$, adding $C \in \mathbf{C}^\star \setminus \{X\}$ as a context of X, resulting $\mathcal{S}' = (\mathcal{S} \setminus \{X\}) \cup \{\langle X, \mathbf{C}_X \cup \{C\} \rangle\}$ improves $\mathcal{S}$ if $C \notin de(X)_{\mathcal{G}_\mathcal{S}}$ and $C \perp\!\!\!\perp Y \mid \lceil \mathbf{C}_X \rceil$ in $\mathcal{H} \setminus \{X\}$.*

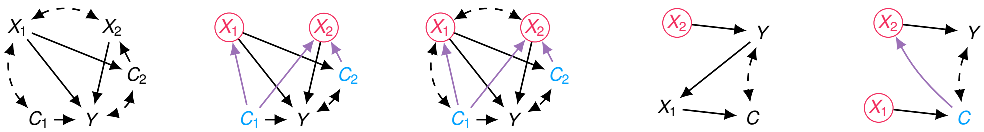
This proposition is straightforward. Note however that the resulting MPS may not be NRO as an added observation can cancel out the relevance of the existing contexts, e.g., Prop. 2 can be viewed as adding observations and removing now irrelevant observations. Further, any set of observations that can be added to a set of actions to improve an MPS can also simply be added sequentially.

Adding new interventions Intervention replaces the natural mechanism for $X \in \mathbf{X}^\star$ with an artificial one $\pi(x|\mathbf{z})$. To guarantee that the alternative one can perform at least as good as the natural one, we should understand what information X originally takes and whether the new contexts $\mathbf{Z}$ carry information tantamount to the original one. If every parent of $X \in \mathbf{X}^\star$ is contextualizable (e.g., no UC), the problem becomes trivial (e.g., Markovian). Otherwise, we examine the existence of a back-door path.[3] Let $Q = P_{\boldsymbol{\pi}}$ and $\mathcal{H} = \mathcal{G}_{\boldsymbol{\pi}}$ for some $\mathcal{S} \sim^{-1} \boldsymbol{\pi}$. Given $X \in \mathbf{X}^\star \setminus \mathbf{X}(\boldsymbol{\pi})$ and $\mathbf{Z} \subseteq \mathbf{C}^\star \setminus \{X\}$, if (i) $(Y \perp\!\!\!\perp X \mid \lceil \mathbf{Z} \rceil)_{\mathcal{H}_{\underline{X}}}$ and (ii) $X \notin an(\mathbf{Z})_{\mathcal{H}}$, then

$$\begin{aligned}\mu_{\boldsymbol{\pi}} = \textstyle\sum_{y,x,\mathbf{z}} yQ(y|x,\mathbf{z})Q(x|\mathbf{z})Q(\mathbf{z}) &\overset{\text{(i)}}{=} \textstyle\sum_{y,x,\mathbf{z}} yQ'_x(y|\mathbf{z})Q(x|\mathbf{z})Q(\mathbf{z}) \\ &\overset{\text{(ii)}}{=} \textstyle\sum_{y,x,\mathbf{z}} yQ'_x(y|\mathbf{z})Q(x|\mathbf{z})Q'_x(\mathbf{z}) \doteq \textstyle\sum_{y,x,\mathbf{z}} yQ'_x(y,\mathbf{z})\pi'(x|\mathbf{z}) \doteq \mu_{\boldsymbol{\pi}'},\end{aligned}$$

for some π'. Since π' can be optimized, $\mu^*_{\mathcal{S}} \leq \mu^*_{\mathcal{S} \cup \{\langle X, \mathbf{Z} \rangle\}}$. However, naively generalizing the criterion to handle a set of interventions is insufficient. Consider Fig. 7a, an observational policy where $\mathbf{X} = \mathbf{X}^\star$ and $\mathbf{C} = \mathbf{C}^\star$. Based on the aforementioned criteria, X_1 and X_2 shall not be intervened simultaneously (by replacing X to $\mathbf{X}$): C_2 cannot be used as $\mathbf{Z}$ since $X_1 \in an(C_2)_{\mathcal{G}}$; $X_2 \leftrightarrow C_2 \rightarrow Y$ is an open back-door path. We propose a solution for adding interventions simultaneously, powered by Thm. 2.

[3] [16, 17] studied 'possibly-optimal' atomic interventions ($\mathbf{C}^\star = \emptyset$) where their conclusions can be essentially reduced to finding actions with no back-door path to Y while varying the strengths of UCs.



(a) $\mathcal{G}_{\mathcal{S}}$ (b) $\mathcal{G}_{\mathcal{S}'}$ (c) $\mathcal{G}_{\mathcal{S}''}$ (d) (e)

Figure 7: (a) a given MPS to construct (c) an improved MPS through (b) an intermediate, invalid MPS. (d,e) demonstrate the use of post-reward interventions to improve a given MPS.

Theorem 3. *Given an MPS $\mathcal{S}$, let $\mathcal{S}' \neq \mathcal{S}$ be an MPS with $\mathbf{X}(\mathcal{S}) \subseteq \mathbf{X}(\mathcal{S}')$ such that $\mathcal{H}''$ the union of induced graphs $\mathcal{G}_{\mathcal{S}} \cup \mathcal{G}_{\mathcal{S}'}$ is acyclic. Let $\mathbf{X}'$ be actions that the MPSes disagree on, i.e., $(\mathbf{X}(\mathcal{S}')\setminus\mathbf{X}(\mathcal{S}))\cup\{X \in \mathbf{X}(\mathcal{S}) \mid \mathbf{C}'_X \neq \mathbf{C}_X\}$, and (invalid) MPS $\mathcal{S}'' \doteq \{\langle X, pa(X)_{\mathcal{H}''} \cup \mathbf{U}_X\rangle\}_{X\in\mathbf{X}'}$. $\mu^*_{\mathcal{S}''} = \mu^*_{\mathcal{S}'}$ can be elicited by Thm. 2, then, $\mu^*_{\mathcal{S}} \leq \mu^*_{\mathcal{S}'}$.*

Given an MPS $\mathcal{S}$, an intermediate MPS is constructed adding new contexts to a subset of $\mathbf{X}^\star$ while assuming that any non-contextualizable variables can be used as contexts. Consider comparing $\mathcal{S} = \emptyset$ (Fig. 7a) and $\mathcal{S}'$ (Fig. 7b) where we employ an intermediate representation $\mathcal{S}''$ (Fig. 7c) to ultimately inspect $\mu^*_{\mathcal{S}} \leq \mu^*_{\mathcal{S}'}$. Thm. 2 is applicable with $\mathbf{U}' = \emptyset$ and $\prec = \langle C_1, \mathbf{U}_{X_1}, \mathbf{U}_{X_2}, X_1, C_2, X_2\rangle$ to demonstrate $\mu^*_{\mathcal{S}''} = \mu^*_{\mathcal{S}'}$. Since $\mu^*_{\mathcal{S}} \leq \mu^*_{\mathcal{S}''}$, we can elicit $\mu^*_{\mathcal{S}} \leq \mu^*_{\mathcal{S}'}$, confirming that $\mathcal{S}$ is not a POMPS. By allowing $\mathbf{X}'$ to intersect with $\mathbf{X}(\mathcal{S})$, the theorem not only adds new inventions but also can replace the contexts of existing interventions.

Refining the space of MPSes Equipped with the characterizations, we can refine the space of MPSes, hence, the space of mixed policies, by filtering out MPSes that are either redundant or dominated by other MPS, eliciting a superset of POMPSes in a given setting. This can be achieved in a brute-force manner by enumerating all MPSes, and examining whether any of Thm. 2, Thm. 3, and Prop. 4 is applicable. One of barriers to design a more principled approach (e.g., dynamic programming [16]) to obtaining POMPSes (or a superset of) is that contexts are interleaved in both terms in Eq. (1) representing a reward mechanism and a policy.

Nevertheless, we investigate simplifying a mixed policy setting while preserving its POMPSes. First, one may think that the descendants of Y can be ignored since neither intervening action variables among them changes the reward nor observing contextualizable variables among them is feasible. Surprisingly, Fig. 7d, where X_1 and C take place *after* the reward is evaluated, remarkably demonstrates the opposite. With X_1 intervened on, C can become a context for X_2 (Fig. 7e) without inducing a cycle. This implies that contexts in the descendants of the reward becomes usable if interventions can break the ancestral relationships. Second, $X \in \mathbf{X}^\star$ that cannot affect $\mathbf{C}^\star$ or Y is not intervene-worthy — if $de(X)_{\mathcal{G}} \cap (\mathbf{C}^\star \cup \{Y\}) = \emptyset$, there exists no MPS that makes $X \in an(\mathbf{C}^\star \cup \{Y\})_{\mathcal{G}}$, and, thus, X can be excluded from $\mathbf{X}^\star$.

Proposition 5. *Given $\langle \mathcal{G}, Y, \mathbf{X}^\star, \mathbf{C}^\star\rangle$, let $\mathbf{X}' \doteq \{X \in \mathbf{X}^\star \mid de(X)_{\mathcal{G}} \cap (\mathbf{C}^\star \cup \{Y\}) \neq \emptyset\}$, $\mathbf{X}'' \doteq de(Y)_{\mathcal{G}_{\underline{\mathbf{X}'}}} \cap \mathbf{X}'$, and $\mathbf{Z} \doteq de(Y)_{\mathcal{G}_{\overline{\mathbf{X}''}}}$. The POMPSes for $\langle \mathcal{G}, Y, \mathbf{X}^\star, \mathbf{C}^\star\rangle$ are the same as those for $\langle \mathcal{G}\setminus\mathbf{Z}, Y, \mathbf{X}', \mathbf{C}^\star\setminus\mathbf{Z}\setminus\mathbf{X}''\rangle$.*

5 Conclusions

In this paper, we studied the space of mixed policies that emerges through the empowerment of an agent to determine the mode it will interact with the environment — i.e., which variables to intervene on and which contexts it decides to look into. Facing new challenges to optimize this new mode of interaction, which has many additional degrees of freedom, we studied the topological structure induced by the different mixed policies, which could in turn be leveraged to determine partial orders across the policy space w.r.t. the maximum expected rewards achievable. As a practical result, we provided a general characterization of the space of mixed policies with respect to properties that allow the agent to detect inefficient and suboptimal strategies. One of the surprising implications of this characterization provided here is that agents following a more standard approach (i.e., intervening on all intervenable variables and observing all available contexts) may be hurting themselves, and may never be able to achieve an optimal performance regardless of the number of interactions performed.

Acknowledgments and Disclosure of Funding

This research is supported in parts by grants from NSF (IIS-1704352 and IIS-1750807 (CAREER)).

References

[1] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

[2] P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2001.

[3] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

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

[5] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[6] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

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

[8] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[9] Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandits: a causal approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1340–1346. AAAI Press, 2017.

[10] J. Zhang and E. Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems 32*, pages 13401–13411. Curran Associates, Inc., 2019.

[11] Guy Tennenholtz, Shie Mannor, and Uri Shalit. Off-policy evaluation in partially observable environments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press, 2020.

[12] Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2015.

[13] Andrew Forney, Judea Pearl, and Elias Bareinboim. Counterfactual data-fusion for online reinforcement learners. In *International Conference on Machine Learning*, pages 1156–1164, 2017.

[14] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *arXiv preprint arXiv:1606.03203*, 2016.

[15] Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*, pages 3057–3066, 2017.

[16] Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems 31*, pages 2568–2578. Curran Associates, Inc., 2018.

[17] Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 4164–4172. AAAI Press, 2019.

[18] R.A. Howard and J.E. Matheson. Influence diagrams. *Principles and Applications of Decision Analysis*, 1981.

[19] Dennis Nilsson and Steffen L. Lauritzen. Evaluating influence diagrams using limids. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 436–445, 2000.

[20] Steffen L. Lauritzen and Dennis Nilsson. Representing and solving decision problems with limited information. *Management Science*, 47:1235–1251, 2001.

[21] Daphne Koller and Brian Milch. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, 45(1):181–221, 2003.

[22] J.M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[23] J. Pearl and J.M. Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, pages 444–453. Morgan Kaufmann, San Francisco, 1995.

[24] S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

[25] Vanessa Didelez, A. Philip Dawid, and Sara Geneletti. Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 138–146. AUAI press, 2006.

[26] A.P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189, 2002.

[27] Jin Tian. Identifying dynamic sequential plans. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*. AUAI press, 2008.

[28] Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 7021–7031, 2017.

[29] Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In *Advances in Neural Information Processing Systems 32*, pages 14369–14379. Curran Associates, Inc., 2019.

[30] T.S. Verma. Graphical aspects of causal models. Technical Report R-191, UCLA, Computer Science Department, 1993.

[31] D. Geiger and J. Pearl. On the logic of causal models. In *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence*, pages 136–147, St Paul, MN, 1988.

[32] D. Geiger, T.S. Verma, and J. Pearl. $d-$separation: From theorems to algorithms. In *Proceedings, 5th Workshop on Uncertainty in AI*, pages 118–124, Windsor, Ontario, Canada, August 1989.

[33] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[34] Juan David Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 35nd AAAI Conference on Artificial Intelligence*. AAAI Press, 2020.

[35] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

[36] Philip Dawid and Vanessa Didelez. Identifying optimal sequential decisions. In *Proceedings of The Twenty-Fourth Annual Conference on Uncertainty in Artificial Intelligence*. AUAI press, 2008.

[37] Qiang Liu and Alexander Ihler. Belief propagation for structured decision making. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*. AUAI press, 2012.

[38] D. Geiger, T.S. Verma, and J. Pearl. Identifying independence in Bayesian networks. In *Networks*, volume 20, pages 507–534. John Wiley, Sussex, England, 1990.

[39] R.D. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6):871–882, 1986.

[40] Nevin Lianwen Zhang. Probabilistic inference in influence diagrams. In *Proceedings of the Fourteenth conference on Uncertainty in Artificial Intelligence*, pages 514–522, 1998.

[41] Tom Everitt, Pedro A. Ortega, Elizabeth Barnes, and Shane Legg. Understanding agent incentives using causal influence diagrams, part i: single action settings, 2019.

[42] Ryan Carey, Eric Langlois, Tom Everitt, and Shane Legg. The incentives that shape behaviour, 2020. arXiv.

[43] Sanghack Lee and Elias Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. Technical Report R-63, Columbia Causal Artificial Intelligence Lab, Department of Computer Science, Columbia University, 2020.

[44] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

A Discussion – Introductory Example

We provide further elaboration on the example discussed in the introduction (Fig. 1a), which is shown again for convenience in Fig. 8. We recall that X_1 and X_2 are intervenable and X_1 and C are observable variables (i.e., can be used as context), so we write $\mathbf{X}^\star = \{X_1, X_2\}$ and $\mathbf{C}^\star = \{C, X_1\}$, following the corresponding notation. There are 15 distinct ways (i.e., mixed policy scopes) for an agent to interact with the system, which is explicitly shown in Fig. 9. This plot is known as a Hasse diagram (i.e., a diagram with transitive reductions) and represents the relationship between the different policies based on two dimensions:

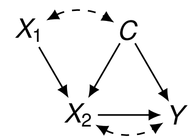


Figure 8: Introduction's causal graph (originally, Fig. 1a).

(a) whether one policy scope can behave always better than or equal to another with respect to maximum achievable expected rewards, which induces a dominance relationship (Fig. 9a);
(b) whether one policy scope has more actions or contexts than another, which is called a subsumption relationship (Fig. 9b).

Regarding the dominance relation, a blue directed edge $\mathcal{S}_\alpha \rightarrow \mathcal{S}_\beta$ corresponds to $\mu^*_{\mathcal{S}_\alpha} \leq \mu^*_{\mathcal{S}_\beta}$ and a gray dotted undirected edge $\mathcal{S}_\alpha - \mathcal{S}_\beta$ represents the equivalence in their maximum achievable expected rewards, (i.e., $\mu^*_\alpha = \mu^*_\beta$); we usually call the nodes connected through these edges an *equivalence class*, given their indistinguishability in terms of achievable rewards.

Regarding the subsumption relation, a red directed edge $\mathcal{S}_\alpha \rightarrow \mathcal{S}_\beta$ represents a subsumption relationship meaning that $\mathcal{S}_\beta$ has more actions or contexts than $\mathcal{S}_\alpha$, so is able to mimic it. The goal is usually to find policies that achieve higher rewards (relative to dimension (a)) and are more parsimonious, or simpler (relative to dimension (b)). For grounding the discussion, we start with dimension (a) and consider μ, the expected reward for the observational policy, and $\mu^*_{\{\langle X_1, \{C\}\rangle\}}$, the maximum expected reward with the policy intervening on X_1 given C. We will show below $\mu \leq \mu^*_{\pi(x_1|c)}$ by using do-calculus,

$$\begin{aligned}
\mu &= \sum\nolimits_y yP(y) && \text{by definition} && (13)\\
&= \sum\nolimits_{y,x_1,c} yP(y, x_1, c) && \text{marginal probability} && (14)\\
&= \sum\nolimits_{y,x_1,c} yP(y|x_1, c)P(x_1|c)P(c) && \text{chain rule} && (15)\\
&= \sum\nolimits_{y,x_1,c} yP_{x_1}(y|c)P(x_1|c)P_{x_1}(c) && \text{Rule 2 of do-calculus} && (16)\\
&= \sum\nolimits_{y,x_1,c} yP_{x_1}(y, c)P(x_1|c) && \text{chain rule} && (17)\\
&= \sum\nolimits_{y,x_1,c} yP_{x_1}(y, c)\pi(x_1|c) && \text{by definition} && (18)\\
&\leq \mu^*_{\{\langle X_1, \{C\}\rangle\}},
\end{aligned}$$

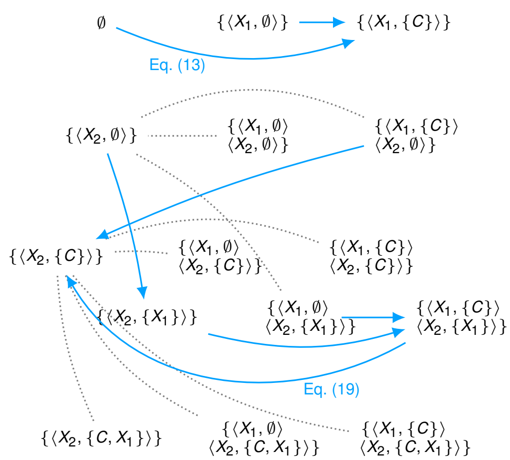
where the last equality comes from the expression for the expected reward (Eq. (1)) with $\pi(x_1|c)$ set to $P(x_1|c)$; the last inequality comes from the fact that the decision rule $\pi(x_1|c)$ can be optimized to yield a higher expected reward.

For further illustration of the dominance relation, we relate two policy scopes $\mathcal{S} = \{\langle X_1, \{C\}\rangle, \langle X_2, \{X_1\}\rangle\}$ and $\mathcal{S}' = \{\langle X_2, \{C\}\rangle\}$ through the following derivation,

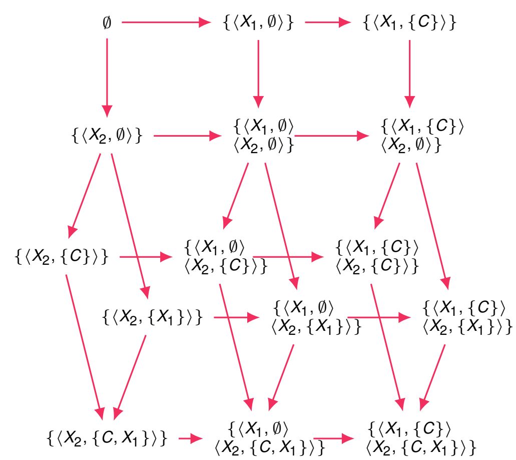
$$\begin{aligned}
\mu^*_{\mathcal{S}} &= \sum\nolimits_{y,\mathbf{x},c} yP_{\mathbf{x}}(y, c)\pi(x_1|c)\pi(x_2|x_1) && \text{by Eq. (1)} && (19)\\
&= \sum\nolimits_{y,\mathbf{x},c} yP_{x_2}(y, c)\pi(x_1|c)\pi(x_2|x_1) && \text{Rule 3 of do-calculus} && (20)\\
&= \sum\nolimits_{y,x_2,c} yP_{x_2}(y, c)\sum\nolimits_{x_1} \pi(x_1|c)\pi(x_2|x_1) && \text{algebra} && (21)
\end{aligned}$$

There exists a probability mapping π' that can listen to C (while preserving the equality),

$$= \sum\nolimits_{y,x_2,c} yP_{x_2}(y, c)\sum\nolimits_{x_1} \pi(x_1|c)\pi'(x_2|x_1, c) \qquad \text{by construction} \qquad (22)$$



(a) dominance



(b) subsumption

Figure 9: All 15 mixed policy scopes and their relationships in terms of two dimensions: (a) maximum achievable expected rewards and (b) policy subsumption. (a) Blue edges $\mathcal{S}_\alpha \rightarrow \mathcal{S}_\beta$ correspond to $\mu^*_{\mathcal{S}_\alpha} \leq \mu^*_{\mathcal{S}_\beta}$ with gray dotted edges the equivalence of their maximum achievable expected rewards, (b) red solid edges $\mathcal{S}_\alpha \rightarrow \mathcal{S}_\beta$ imply $\mathcal{S}_\alpha \subset \mathcal{S}_\beta$. Policy scopes are located based on interventions on X_2 (vertical) and on X_1 (horizontal). Also, their positions are preserved to facilitate corresponding comparisons.

$$= \sum_c \sum_{x_1} \pi(x_1|c) \sum_{y,x_2} y P_{x_2}(y,c) \pi'(x_2|x_1,c) \quad \text{algebra} \qquad (23)$$

There exists a value x_1 for each c that can maximize the expression. Let x_1^* be a function mapping from c to such value of X_1. Then,

$$\leq \sum_{y,x_2,c} y P_{x_2}(y,c) \pi'(x_2|x_1^*(c),c) \quad \text{by definition} \qquad (24)$$

Since $x_1^*(c)$ is determined by c, there exists $\pi(x_2|c)$ such that

$$= \sum_{y,x_2,c} y P_{x_2}(y,c) \pi(x_2|c) \quad \text{by construction} \qquad (25)$$

$$\leq \mu^*_{\mathcal{S}'} \quad \text{by Eq. (1).} \qquad (26)$$

Therefore, $\mu^*_{\{\langle X_1,\{C\}\rangle,\langle X_2,\{X_1\}\rangle\}} \leq \mu^*_{\{\langle X_2,\{C\}\rangle\}}$. It is not immediately obvious how we can formally derive such inequalities between two optimal expected rewards for arbitrary environments. Throughout the paper, we build graphical and algorithmic criteria that tell whether one policy can dominate another.

After all, $\{\langle X_1,\{C\}\rangle\}$ (top-right in Fig. 9a) dominates its neighbors and can attain optimality. Sometimes, a set of policies forming an equivalence class can achieve optimality, i.e.,

$$\{\{\langle X_2,\{C\}\rangle\},\{\langle X_1,\emptyset\rangle,\langle X_2,\{C\}\rangle\},\{\langle X_1,\{C\}\rangle,\langle X_2,\{C\}\rangle\},\{\langle X_2,\{C,X_1\}\rangle\},$$
$$\{\langle X_1,\emptyset\rangle,\langle X_2,\{C,X_1\}\rangle\},\{\langle X_1,\{C\}\rangle,\langle X_2,\{C,X_1\}\rangle\}\}.$$

Now, we turn our attention to the subsumption relation as shown in Fig. 9b. We first note that the construction of this diagram is based on the scope of the given policy (Def. 1), as defined in the paper, namely, the set of actions (before conditioning bar) and the corresponding context (after the conditioning bar). The construction is graph-insensitive, but will play a key role when combined with the analysis of dominance. Specifically, if a policy scope does not have an incoming red edge from other policy scope in its equivalence class, the policy scope is *non-redundant*. To better understand how the dominance and subsumption dimensions are related, we superimpose both relations in Fig. 10. The scopes forming an equivalence class are clustered and highlighted within a gray rectangle, and marked its boundary with black for optimality. For each equivalence class, there exists one non-redundant scope, highlighted in yellow. For instance, in the aforementioned equivalence class (the right most equivalence class in Fig. 10), $\{\langle X_2,\{C\}\rangle\}$ is subsumed by other

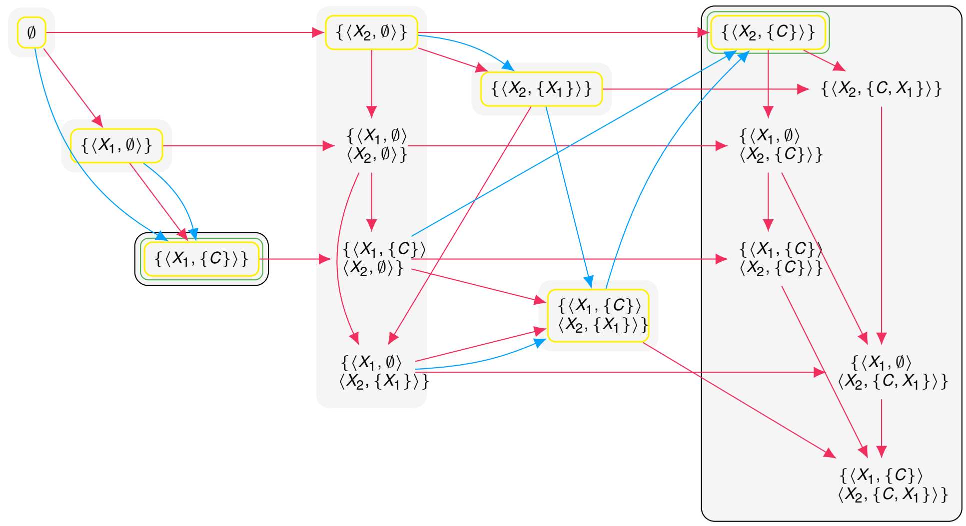


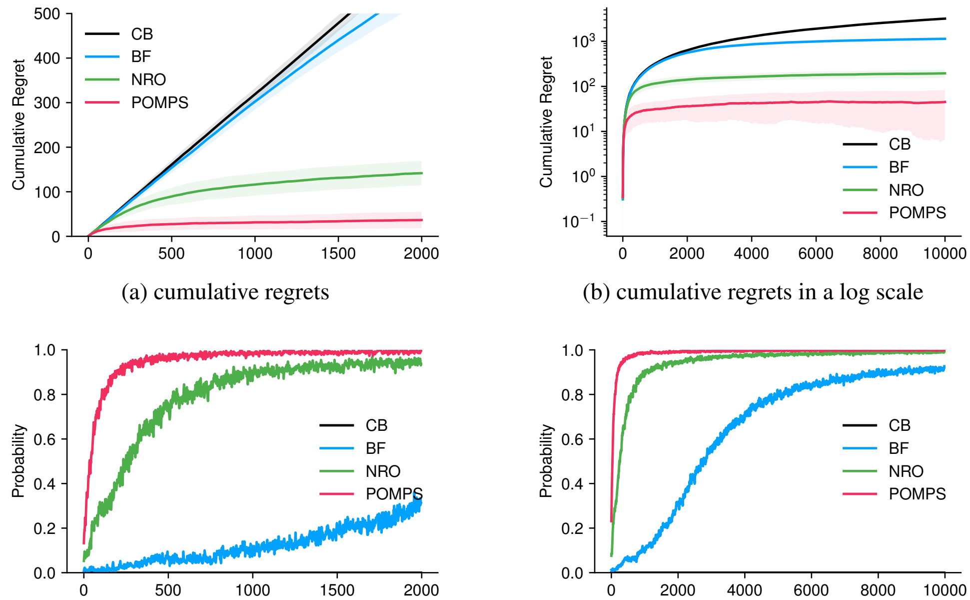
Figure 10: The superimposition of the dominance and subsumption relations of Fig. 1a, which were individually shown in Figs. 9a and 9b. The equivalence classes are highlighted in gray-shaded rectangles, policies (EC) that can achieve optimality are in black boundaries, non-redundant policies are in yellow. Mixed policy scopes satisfying both non-redundancy and optimality are in green.

scopes while maintaining the same optimal reward. In fact, this implies that $\{\langle X_2, \{C\}\rangle\}$ will be preferred over its counterparts in the equivalence class given its capability of achieving the optimality while being the most parsimonious within its class. Comparing whether one policy scope subsumes the other outside the equivalence class makes less sense since they are not comparable. For example, $\{\langle X_2, \{C\}\rangle\}$ subsumes $\{\langle X_2, \emptyset\rangle\}$ and $\emptyset$ through the red arrows, but belongs to a different equivalence class, so they are non-comparable, one is not preferred over the other. In this example, we can see through Fig. 10 that there are 7 non-redundant mixed policy scopes (yellow): $\emptyset$, $\{\langle X_1, \emptyset\rangle\}$, $\{\langle X_1, \{C\}\rangle\}$, $\{\langle X_2, \emptyset\rangle\}$, $\{\langle X_2, \{X_1\}\rangle\}$, $\{\langle X_1, \{C\}\rangle, \langle X_2, \{X_1\}\rangle\}$, and $\{\langle X_2, \{C\}\rangle\}$. Putting this information together, we can see in Fig. 10 that only $\{\langle X_1, \{C\}\rangle\}$ and $\{\langle X_2, \{C\}\rangle\}$ satisfy both optimality (black boundaries) and non-redundancy (yellow), which are then marked in green.

Once the intelligent agent has access to causal information (e.g., in the form of the causal graph), it can explore the underlying environment with scopes that can achieve optimal reward efficiently. We now describe different approaches the agent can take. A standard approach would be taking all actions (e.g., in this case, X_1 and X_2) and observing all available contexts (X_1 and C), which leads to the scope $\{\langle X_1, \{C\}\rangle, \langle X_2, \{X_1, C\}\rangle\}$. Another approach could be brute-force, where all the 15 different scopes are experimented. A more efficient route would be to avoid redundant policy scopes, where the agent plays only the 7 non-redundant scopes. Knowing that 5 out of the 7 scopes are *no better than* the other 2, the most efficient approach would be to assess those, i.e., $\{\langle X_1, \{C\}\rangle\}$ and $\{\langle X_2, \{C\}\rangle\}$. We name these approaches as CB, BF, NRO, and POMPS, where the exact meaning of NRO (Non-Redundant under Optimality) and POMPS (Possibly-Optimal Mixed Policy Scope) will become clearer through out the paper.

We empirically validate that the use of refined policies leads to a better performance measured. In particular, we use the cumulative regret (the lower the better), i.e., $T\mu^* - \sum_{t=1}^{T} Y_t$, where T is the number of time steps (i.e., interactions) and Y_t is a random variable for the reward at time t. Further, we demonstrate that the CB approach cannot achieve the optimal reward in a certain environment, incurring a linear cumulative regret. The basic experimental setup is, for each time step, the agent assesses policies using samples from posterior reward distributions (i.e., Thompson sampling) based on its interaction history, and executes the chosen policy.[4] We exemplified next an environment (structural causal model) compatible with the example discussed above, which will validate the

[4]Here, we mean a policy by fully-specified decision rules. For instance, there are four discrete policies corresponding to a strategy $\pi(x_1|c)$ with binary C and X_1.



(a) cumulative regrets

(b) cumulative regrets in a log scale

(c) probability selecting optimal policies

(d) probability selecting optimal policies

Figure 11: Performance comparison for different approaches (CB, BF, NRO, POMPS). (a,b) Each line represents cumulative regrets averaged over 100 repetitions (the lower the better), and its shade represents standard deviation. (c,d) Probability the agent selects the best policies (with CB 0% and the lines are smoothed with moving average). The figures in the left side highlight the first 2,000 time steps and ones in the right side the whole 10,000 steps.

rewards, and is unknown by the agent:

$$\mathcal{M} = \begin{cases} C \leftarrow U_C \oplus U_1 \\ X_1 \leftarrow U_1 \oplus U_{X_1} \\ X_2 \leftarrow X_1 \oplus C \oplus U_2 \oplus U_{X_2} \\ Y \leftarrow X_2 \oplus C \oplus U_2 \oplus U_Y, \end{cases}$$

where the unobserved confounders U_1 (between C and X_1) and U_2 (between X_2 and Y) are fair coins and each of $U \in \{U_{X_1}, U_{X_2}, U_C, U_Y\}$ is binary and follows $P(U = 1) = 0.1$.

The corresponding simulation is shown in Fig. 11 reporting two types of plots based on (a,b) cumulative regrets and (c,d) the probability selecting the optimal policy. It is evident from the specification that the CB agent cannot optimize its policy to achieve the optimality, regardless of the number of interactions with the environment. This is demonstrated as a linear cumulative regret (Figs. 11a and 11b) and 0% probability selecting an optimal policy in Figs. 11c and 11d. The BF approach is almost equally inefficient up to around 2000 steps, but is still able to find the optimal since it includes the possible optimal policies, POMPSes. As expected, the performance improves with the use of smaller number of policies. After all, the CB approach does not guarantee the optimality, while BF, NRO, POMPS are always guaranteed to converge. Further, the use of only non-redundant policies by NRO and POMPS helps the agent to converge to the optimal policy faster.

B Preliminaries

We use in the paper classic causal inference results such as do-calculus, which we summarize here.

D-separation We start with the definition of d-separation [44], without a particular consideration of deterministic relationships.

Definition 6 (d-separation)**.** Two sets of vertices $\mathbf{X}, \mathbf{Y}$ are said to be d-separated by a another set $\mathbf{Z}$ in a directed acyclic graph $\mathcal{G}$, denoted by $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$, if every path $\mathbf{P}$ from vertices in $\mathbf{X}$ to vertices in $\mathbf{Y}$ are blocked where blockage occurs when one of the following holds:

1. $\mathbf{P}$ contains at least one arrow-emitting node that is in $\mathbf{Z}$, or
2. $\mathbf{P}$ contains at least one collider that is outside $\mathbf{Z}$ and has no descendant in $\mathbf{Z}$.

Do-calculus Do-calculus [33] is an essential machinery to reason about the equivalence of conditional interventional probabilities induced by any model conforming to a given causal graph. Do-calculus consists of three rules where each rule ascertains that an equality between two probability distributions holds if a certain graphical test (separation) holds. The three rules are

R1 (Insertion/deletion of observations):	$P_{\mathbf{x}}(\mathbf{y}\mid\mathbf{z},\mathbf{w}) = P_{\mathbf{x}}(\mathbf{y}\mid\mathbf{w})$	if $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}, \mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}}}$
R2 (Action/observation exchange):	$P_{\mathbf{x}}(\mathbf{y}\mid\mathbf{z},\mathbf{w}) = P_{\mathbf{x},\mathbf{z}}(\mathbf{y}\mid\mathbf{w})$	if $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}, \mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}$
R3 (Insertion/deletion of actions):	$P_{\mathbf{x}}(\mathbf{y}\mid\mathbf{w}) = P_{\mathbf{x},\mathbf{z}}(\mathbf{y}\mid\mathbf{w})$	if $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}, \mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}\mathbf{Z}(\mathbf{W})}}$

where $\mathbf{Z}(\mathbf{W})$ is a subset of $\mathbf{Z}$ that is not ancestor of $\mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}}}$. For convenience of some of the proofs, let $\mathcal{H} = \mathcal{G}\backslash\mathbf{X}$ and $Q = P_{\mathbf{x}}$, and consider the following (rewritten) what rules:

R1 (Insertion/deletion of observations):	$Q(\mathbf{y}\mid\mathbf{z},\mathbf{w}) = Q(\mathbf{y}\mid\mathbf{w})$	if $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}$ in $\mathcal{H}$
R2 (Action/observation exchange):	$Q(\mathbf{y}\mid\mathbf{z},\mathbf{w}) = Q_{\mathbf{z}}(\mathbf{y}\mid\mathbf{w})$	if $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}$ in $\mathcal{H}_{\underline{\mathbf{Z}}}$
R3 (Insertion/deletion of actions):	$Q(\mathbf{y}\mid\mathbf{w}) = Q_{\mathbf{z}}(\mathbf{y}\mid\mathbf{w})$	if $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}$ in $\mathcal{H}_{\overline{\mathbf{Z}(\mathbf{W})}}$.

This representation and will help us to highlight the differences between distributions while abstracting away unnecessary details.

C Mixed Policies

Derivation of Expected Reward A derivation for Eq. (1) is shown below with abbreviations: MP, the definition of marginal probability; CR, the chain rule; R#: rule # of do-calculus; and def: by definition. Let $\prec$ be a topological order compatible with $\mathcal{G}_{\boldsymbol{\pi}}$, and let $\mathbf{X}_{\prec C}$ be $\{X \in \mathbf{X} \mid X \prec C\}$. We may write capital subscripts of a value as lowercase, e.g., $\mathbf{c}_x$ instead of $\mathbf{c}_X$ for a value for $\mathbf{C}_X$. With $\mathbf{X} = \mathbf{X}(\boldsymbol{\pi})$, $\mathbf{C} = \mathbf{C}(\boldsymbol{\pi})$, and $\mathbf{C}^- = \mathbf{C}\backslash\mathbf{X}$, we start by writing the expected reward of $\boldsymbol{\pi}$,

$$
\begin{aligned}
\mu_{\boldsymbol{\pi}} &\doteq \mathbb{E}_{\boldsymbol{\pi}}[Y] && \text{def} \quad (27)\\
&= \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\boldsymbol{\pi}}(y, \mathbf{x}, \mathbf{c}^-) && \text{MP} \quad (28)\\
&= \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\boldsymbol{\pi}}(y\mid\mathbf{x}, \mathbf{c}^-) P_{\boldsymbol{\pi}}(\mathbf{x}, \mathbf{c}^-) && \text{CR} \quad (29)\\
&= \sum_{y,\mathbf{x},\mathbf{c}^-} y P(y\mid\mathbf{x}, \mathbf{c}^-) P_{\boldsymbol{\pi}}(\mathbf{x}, \mathbf{c}^-) && \text{R1} \quad (30)\\
&= \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\mathbf{x}}(y\mid\mathbf{c}^-) P_{\boldsymbol{\pi}}(\mathbf{x}, \mathbf{c}^-) && \text{R2} \quad (31)\\
&= \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\mathbf{x}}(y\mid\mathbf{c}^-) \prod_{X\in\mathbf{X}} P_{\boldsymbol{\pi}}(x\mid\mathbf{x}_{\prec x}, \mathbf{c}^-_{\prec x}) \prod_{C\in\mathbf{C}^-} P_{\boldsymbol{\pi}}(c\mid\mathbf{x}_{\prec c}, \mathbf{c}^-_{\prec c}) && \text{CR} \quad (32)\\
&= \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\mathbf{x}}(y\mid\mathbf{c}^-) \prod_{X\in\mathbf{X}} P_{\boldsymbol{\pi}}(x\mid\mathbf{c}_x) \prod_{C\in\mathbf{C}^-} P_{\boldsymbol{\pi}}(c\mid\mathbf{x}_{\prec c}, \mathbf{c}^-_{\prec c}) && \text{R1} \quad (33)\\
&= \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\mathbf{x}}(y\mid\mathbf{c}^-) \prod_{X\in\mathbf{X}} \pi(x\mid\mathbf{c}_x) \prod_{C\in\mathbf{C}^-} P_{\boldsymbol{\pi}}(c\mid\mathbf{x}_{\prec c}, \mathbf{c}^-_{\prec c}) && \text{def} \quad (34)\\
&= \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\mathbf{x}}(y\mid\mathbf{c}^-) \prod_{X\in\mathbf{X}} \pi(x\mid\mathbf{c}_x) \prod_{C\in\mathbf{C}^-} P(c\mid\mathbf{x}_{\prec c}, \mathbf{c}^-_{\prec c}) && \text{R1} \quad (35)\\
&= \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\mathbf{x}}(y\mid\mathbf{c}^-) \prod_{X\in\mathbf{X}} \pi(x\mid\mathbf{c}_x) \prod_{C\in\mathbf{C}^-} P_{\mathbf{x}_{\prec c}}(c\mid\mathbf{c}^-_{\prec c}) && \text{R2} \quad (36)\\
&= \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\mathbf{x}}(y\mid\mathbf{c}^-) \prod_{X\in\mathbf{X}} \pi(x\mid\mathbf{c}_x) \prod_{C\in\mathbf{C}^-} P_{\mathbf{x}}(c\mid\mathbf{c}^-_{\prec c}) && \text{R3} \quad (37)
\end{aligned}
$$

Algorithm 1 Separation of actions and contexts of an MPS

1: **function** sep-mps($\mathcal{S}, \mathcal{G}$)
  input: a mixed policy scope $\mathcal{S}$, a causal graph $\mathcal{G}$
  output: an updated, action-context separated mixed policy scope $\mathcal{S}$
2:   **for** $X \in$ topological-order($\mathbf{X}(\mathcal{S}); \mathcal{G}_{\mathcal{S}}$) **do**
3:     Replace $\mathbf{C}_X$ in $\mathcal{S}$ by $(\bigcup_{X' \in \mathbf{C}_X \cap \mathbf{X}^\star} \mathbf{C}_{X'}) \cup (\mathbf{C}_X \backslash \mathbf{X}^\star)$.
4:   **return** $\mathcal{S}$

$$= \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\mathbf{x}}(y|\mathbf{c}^-) \prod_{X\in\mathbf{X}} \pi(x|\mathbf{c}_x) P_{\mathbf{x}}(\mathbf{c}^-) \quad \text{CR} \tag{38}$$

$$= \sum_{y,\mathbf{x},\mathbf{c}^-} y P_{\mathbf{x}}(y, \mathbf{c}^-) \prod_{X\in\mathbf{X}} \pi(x|\mathbf{c}_x), \quad \text{CR} \tag{39}$$

Note that Rule 1 of do-calculus applied to the regime nodes (Eq. (30)) is identical to Rule 2 applied to $\mathbf{X}$ in Eq. (31). The derivation for a subset of $\mathbf{X}$ and its contexts follows directly by treating uninteresting decision rules as natural mechanisms.

Note on Multi-Agent Systems Although the treatment given to mixed policies is framed with respect to a single agent, its implications to a multi-agent setting is apparent – each action variable can be considered as an agent where the absence of directed edges among them exhibits their autonomy. Further, from the multi-agent point of view, the current definition of mixed policy assumes that each agent has the same ability to sense contextual variables $\mathbf{C}^\star$. More realistic multi-agent settings will allow for different sensing capabilities for agents. The results presented in this paper can be effortlessly generalized to this case, where each agent (or action) is associated with its own set of contextualizable variables. Another almost immediate extension is for multi-reward settings, e.g., where one attempts to optimize Y_1 and Y_2. Depending on the task, one may focus on a specific reward, or one can create a new aggregate reward $Y = Y_1 + Y_2$ to perform a task over the setting.

D Optimality and Deterministic Mixed Policy

Proposition 1. *Given a mixed policy scope, there always exists a deterministic mixed policy, which is optimal with respect to the given scope.*

Proof. Consider an arbitrary optimal policy $\boldsymbol{\pi} \sim \mathcal{S}$ given an MPS $\mathcal{S}$. Let $\mathbf{X} = \mathbf{X}(\boldsymbol{\pi})$ and $\mathbf{C}^- = \mathbf{C}(\boldsymbol{\pi})\backslash\mathbf{X}$. Given a topological order among $\mathbf{X}$ defined over $\mathcal{G}_{\mathcal{S}}$ such that $X_i \prec X_j$ if $i < j$, let $Q' = P_{\boldsymbol{\pi}\backslash\{X_1\}}$ where $\boldsymbol{\pi}\backslash\mathbf{X}'$ denotes a policy $\boldsymbol{\pi}$ with actions $\mathbf{X}' \subseteq \mathbf{X}$ removed. Then,

$$\mu_{\boldsymbol{\pi}} = \sum_{y,x_1,\mathbf{c}_1} y Q'_{x_1}(y, \mathbf{c}_1)\pi(x_1|\mathbf{c}_1) = \sum_{\mathbf{c}_1,x_1} \pi(x_1|\mathbf{c}_1) \sum_y y Q'_{x_1}(y, \mathbf{c}_1).$$

If $\pi_{X_1|\mathbf{C}_1}$ is not deterministic with respect to $\mathbf{c}_1$ where $P_{\boldsymbol{\pi}}(\mathbf{c}_1) = Q'(\mathbf{c}_1) > 0$, there must be at least two values of x'_1 and x''_1 such that

$$\pi(x'_1|\mathbf{c}_1) \sum_y y Q'_{x'_1}(y, \mathbf{c}_1) = \pi(x''_1|\mathbf{c}_1) \sum_y y Q'_{x''_1}(y, \mathbf{c}_1).$$

Otherwise, if one value is larger than the other, this contradicts the optimality since $\pi_{X_1|\mathbf{C}_1}$ can select the value that yields a larger value than the other. In case of $Q'(\mathbf{c}_1) = 0$, the choice of x_1 becomes irrelevant. Hence, we can modify the strategy on a single action to be deterministic for a specific context. This argument can be sequentially applied to the rest of intervened variables. As a result, one can elicit a deterministic optimal mixed policy from a given optimal mixed policy. Therefore, there exists a deterministic mixed policy, which is optimal with respect to the given MPS. □

Proposition 2 (Separation of Actions and Contexts)**.** *Given an MPS $\mathcal{S}$, there always exists a deterministic mixed policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ such that $\mathbf{X}(\boldsymbol{\pi})$ and $\mathbf{C}(\boldsymbol{\pi})$ are disjoint and $\mu_{\boldsymbol{\pi}} = \mu^*_{\mathcal{S}}$.*

Proof. Let a mixed policy $\boldsymbol{\rho} \sim \mathcal{S}$ be optimal with respect to $\mathcal{S}$. First, there exists an optimal deterministic mixed-policy $\boldsymbol{\rho}'$ equivalent to $\boldsymbol{\rho}$ with respect to the expected reward (Prop. 1). Since the graph $\mathcal{G}_{\boldsymbol{\rho}'}$ is acyclic, there exists a topological order among $\mathbf{X}$. Consider $X \in \mathbf{X}$ such that $\mathbf{C}_{X'} \cap \mathbf{X} = \emptyset$ for every $X' \in \mathbf{C}_X \cap \mathbf{X}$. We can create a new function π_X based on ρ'_X and $\rho'_{X'}$:

$$\pi_X((\mathbf{c}_X \backslash \{x'\}) \dot{\cup} \mathbf{c}_{X'}) \doteq \rho'_X(\mathbf{C}_X \backslash \{X'\} = \mathbf{c}_x \backslash \{x'\}, X' = \rho'_{X'}(\mathbf{c}_{X'})).$$

This can be iteratively applied following the topological order among $\mathbf{X}$ to obtain a new deterministic policy $\boldsymbol{\pi}$ such that $\mathbf{X}(\boldsymbol{\pi})$ and $\mathbf{C}(\boldsymbol{\pi})$ are disjoint without changing the expected reward (see Alg. 1). □

E Non-Redundant Mixed Policy

Theorem 1. *Let $\mathcal{S} = \{\langle X, \mathbf{C}_X \rangle\}_{X \in \mathbf{X}}$ be an MPS and let $\mathcal{H} = \mathcal{G}_{\mathcal{S}}$. $\mathcal{S}$ is non-redundant if and only if (i) $\mathbf{X} \subseteq an(Y)_{\mathcal{H}}$ and (ii) $(C \not\perp\!\!\!\perp Y \mid \mathbf{C}_X \backslash \{C\})$ in $\mathcal{H} \backslash \{X\}$, for every $X \in \mathbf{X}$ and $C \in \mathbf{C}_X$.*

Proof. **(Only if) (i)** Let $X \in \mathbf{X} \backslash an(Y)_{\mathcal{H}}$, $Q' = P_{\boldsymbol{\pi} \backslash \{X\}}$ and $\mathcal{H}' = \mathcal{G}_{\boldsymbol{\pi} \backslash \{X\}}$. First, $X \in \mathbf{X} \backslash an(Y)_{\mathcal{H}'}$ since intervening on X does not change the descendants of X. Then, $Q'_x(y|\mathbf{c}_x) = Q'(y|\mathbf{c}_x)$ since $(X \perp\!\!\!\perp Y \mid \mathbf{C}_X)$ in $\mathcal{H}'_{\overline{X(\mathbf{C}_X)}} = \mathcal{H}'_{\overline{X}}$ (i.e., Rule 3 of do-calculus). Further, $Q'_x(\mathbf{c}_x) = Q'(\mathbf{c}_x)$ since, again, Rule 3 that $X \perp\!\!\!\perp \mathbf{C}_X$ in $\mathcal{H}'_{\overline{X}}$ as no $\mathbf{C}_X$ is a descendant of X and nothing is given. Then,

$$\begin{aligned}
\mu_{\boldsymbol{\pi}} &= \sum_{y,x,\mathbf{c}_x} y Q'_x(y, \mathbf{c}_x) \pi(x|\mathbf{c}_x) && \text{def} \\
&= \sum_{y,x,\mathbf{c}_x} y Q'_x(y|\mathbf{c}_x) Q'_x(\mathbf{c}_x) \pi(x|\mathbf{c}_x) && \text{CR} \\
&= \sum_{y,x,\mathbf{c}_x} y Q'(y|\mathbf{c}_x) Q'(\mathbf{c}_x) \pi(x|\mathbf{c}_x) && \text{violation of (i)} \\
&= \sum_{y,x,\mathbf{c}_x} y Q'(y, \mathbf{c}_x) \pi(x|\mathbf{c}_x) && \text{CR} \\
&= \sum_{y,\mathbf{c}_x} y Q'(y, \mathbf{c}_x) \sum_x \pi(x|\mathbf{c}_x) && \text{algebra} \\
&= \sum_y y Q'(y) && \text{MP} \\
&= \mu_{\boldsymbol{\pi} \backslash \{X\}} && \text{def.}
\end{aligned}$$

(ii) Let $Q = P_{\boldsymbol{\pi}}$ and $Q' = P_{\boldsymbol{\pi} \backslash \{X\}}$ for some $\boldsymbol{\pi} \sim \mathcal{S}$ and $\mathbf{C}_X^- = \mathbf{C}_X \backslash \{C\}$ where $C \in \mathbf{C}_X$ which violates (ii). Let $\mathcal{H}' = \mathcal{G}_{\mathcal{S} \backslash \{X\}}$. Note that the test in $\mathcal{H} \backslash \{X\}$ is identical to the test in $\mathcal{H}' \backslash \{X\}$ as the only differences in $\mathcal{H}$ and $\mathcal{H}'$ are the parents of X. Then,

$$\begin{aligned}
\mu_{\boldsymbol{\pi}} &= \sum_{y,x,\mathbf{c}_x} y Q'_x(y|\mathbf{c}_x) Q'_x(\mathbf{c}_x) \pi(x|\mathbf{c}_x) && \text{def, CR} \\
&= \sum_{y,x,\mathbf{c}_x} y Q'_x(y|\mathbf{c}_x^-) Q'_x(\mathbf{c}_x) \pi(x|\mathbf{c}_x) && \text{violation of (ii)} \\
&= \sum_{y,x,\mathbf{c}_x} y Q'_x(y|\mathbf{c}_x^-) Q'_x(c|\mathbf{c}_x^-) Q'_x(\mathbf{c}_x^-) \pi(x|\mathbf{c}_x) && \text{CR} \\
&= \sum_{y,x,\mathbf{c}_x^-} y Q'_x(y, \mathbf{c}_x^-) \sum_c Q'_x(c|\mathbf{c}_x^-) \pi(x|\mathbf{c}_x) && \text{CR, algebra} \\
&= \sum_{y,x,\mathbf{c}_x^-} y Q'_x(y, \mathbf{c}_x^-) \sum_c Q_x(c|\mathbf{c}_x^-) \pi(x|\mathbf{c}_x) && \text{R1} \\
&= \sum_{y,x,\mathbf{c}_x^-} y Q'_x(y, \mathbf{c}_x^-) \sum_c P_{\boldsymbol{\pi}}(c|\mathbf{c}_x^-) \pi(x|\mathbf{c}_x) && \text{R3}
\end{aligned}$$

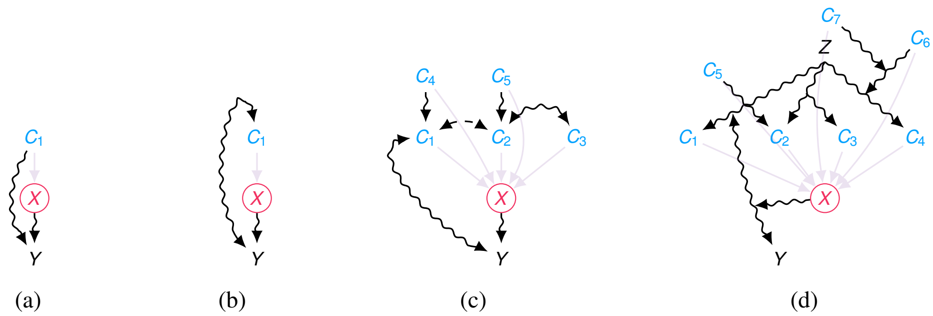


Figure 12: Abstract representation of different minimal edge subgraphs highlighting only a single group of context variables (omitting directed edges from contexts to action). (a, b) a singleton context variable can either have a directed or bidirected path towards Y. (c) a subset of group $\{C_1, C_2, C_3\}$ are connected via bidirected paths with $\{C_4, C_5\}$ having directed paths onto them. (d) bidirected paths between a subset of group $\{C_1, C_2, C_3, C_4\}$ are shared; directed paths from C_7 and C_6 to C_4 are shared; the bidirected path between C_1 and Y intersects bidirected paths between C_1 and $\{C_2, C_3, C_4\}$; finally, the directed path from X to Y is also shared.

$$= \sum_{y,x,\mathbf{c}_x^-} y Q'_x(y, \mathbf{c}_x^-)\pi''(x|\mathbf{c}_x^-) \qquad \text{see below}$$
$$= \mu_{\boldsymbol{\pi}''},$$

where $\pi''(x|\mathbf{c}_x^-) \doteq \sum_c P_{\boldsymbol{\pi}}(c|\mathbf{c}_x^-)\pi(x|\mathbf{c}_x^-, c)$. Since $\boldsymbol{\pi}$ properly subsumes $\boldsymbol{\pi}''$, $\boldsymbol{\pi}$ is redundant.

(If) We show, for an arbitrary MPS $\mathcal{S}' \subsetneq \mathcal{S}$, that we can construct an SCM $\mathcal{M} \sim \mathcal{G}$ such that a mixed policy $\boldsymbol{\pi} \sim \mathcal{S}$ satisfies $\mu_{\boldsymbol{\pi}} \neq \mu_{\boldsymbol{\pi}'}$ for $\boldsymbol{\pi}' \subsetneq \boldsymbol{\pi}$. Let $\mathcal{S}'$ and $\mathcal{S}$ differ on $\langle X, \mathbf{C}_X \rangle$ (either only a subset of $\mathbf{C}_X$ or X itself). We consider a minimal edge subgraph $\mathcal{H}$ of $\mathcal{G}_{\mathcal{S}}$ such that the above conditions are satisfied for X and $\mathbf{C}_X$. The graph is characterized as groups of context variables $\{\mathbf{C}_X^i\}_i$ where, for each group $\mathbf{C}_X^i$, there exists a subset of context variables $\mathbf{C}_X^{i\,\prime}$ connected via bidirected paths and each of the rest $\mathbf{C}_X^i \setminus \mathbf{C}_X^{i\,\prime}$ has a directed path towards the subset $\mathbf{C}_X^{i\,\prime}$. More precisely speaking, there exists no $\mathbf{C}_X$ appearing in the (bi)directed paths as non-ends, while the paths in general can intersect. If the group is a singleton $\mathbf{C}_X^i = \{C\}$, either a bidirected or directed path towards Y, not passing X, exists. Otherwise, a bidirected path exists between $C \in \mathbf{C}_X^{i\,\prime}$ and Y (see Fig. 12 for examples where squiggly lines represent paths and induced edges are hidden).

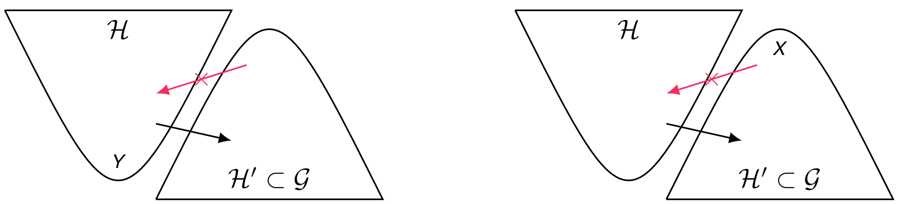
We now construct an SCM demonstrating non-redundancy. As illustrated in Fig. 12d, above mentioned paths can intersect with each other. We consider each bidirected and directed path maintains its own 'channel' where the path in $\mathcal{H}$ can be understood as a cable of multiple bits where non-end variables pass bits to the downstream. This principle is also applied to different groups since the paths connecting each group to Y can be shared. Let every parentless variable (including UCs and $\mathbf{C}_X^i \setminus \mathbf{C}_X^{i\,\prime}$) behaves as a fair coin or a vector of independent fair coins if it involves in multiple paths (e.g., Z in Fig. 12d has 6 bits for every pair among $\{C_1, C_2, C_3, C_4\}$). We design the function for every $C \in \mathbf{C}_X^{i\,\prime}$ and X to be the bit-parity of its parents (i.e., all channels incoming to C) and the mechanism for Y is similarly designed except that it takes its complement. Information of every fair coin is counted twice at X and canceled out (i.e., bit parity) except the one involved between the group and Y (e.g., one between C_1 and Y in every example in Fig. 12) which will be canceled out at Y. Then, the expected reward for $\boldsymbol{\pi}$ becomes 1.0 as every bit-parity is counted twice and complemented at Y. But any $\boldsymbol{\pi}'$, whose X does not listen to $\mathbf{C}_X$ as a whole, makes its expected reward 0.5. $\square$

Proposition 6. *Given an MPS $\mathcal{S}$, every intermediate MPS* nr-mps *is valid MPS.*

Proof. Let $\mathcal{S}$ be an intermediate MPS at a point of time in the execution of nr-mps. Since the algorithm removes a subset of $\mathbf{X}(\mathcal{S})$ and that of contexts $\mathbf{C}_X$ for some $X \in \mathbf{X}(\mathcal{S})$, the condition (i) of the definition of MPS is always satisfied. Hence, we focus on the condition (ii), the acyclicity of an intermediate MPS. The given valid MPS is changed through Line 2, 5, and 8. Removing a context at

Algorithm 2 Non-Redundant Mixed Policy Scope $\mathcal{S}$

1: **function** NR-MPS$(\mathcal{G}, Y, \mathcal{S})$
  input: a mixed policy scope $\mathcal{S}$, a causal graph $\mathcal{G}$
  output: an updated, non-redundant mixed policy scope $\mathcal{S}$
2:   $\mathcal{S} \leftarrow \mathcal{S} \setminus (\mathbf{X} \setminus an(Y)_{\mathcal{G}_\mathcal{S}})$.
3:   **for** $X \in$ reverse-order$(\mathbf{X}(\mathcal{S}); \mathcal{G}_\mathcal{S})$ **do**
4:     **if** $X \notin an(Y)_{\mathcal{G}_\mathcal{S}}$ **then**:
5:       $\mathcal{S} \leftarrow \mathcal{S} \setminus \{X\}$ and **continue**.
6:     **for** $C \in \mathbf{C}_X$ **do**
7:       **if** $(C \perp\!\!\!\perp Y \mid \mathbf{C}_X \setminus \{C\})$ in $\mathcal{G}_\mathcal{S} \setminus \{X\}$ **then**
8:         $\mathcal{S} \leftarrow (\mathcal{S} \setminus \{X\}) \cup \{\langle X, \mathbf{C}_X \setminus \{C\} \rangle\}$.
9:   **return** $\mathcal{S}$.



(a) Schematic for Line 2   (b) Schematic for Line 5

Figure 13: Schematic diagrams for induced graphs after (a) Line 2 and (b) Line 5 where each induced graph is partitioned into two parts by (non-)ancestors of Y and (non-)descendants of X, respectively.

Line 8 from $\mathcal{S}$ results in an induced graph $\mathcal{G}_{\mathcal{S}'}$ which is an edge subgraph of $\mathcal{G}_\mathcal{S}$ where $\mathcal{S}'$ is an MPS after Line 8. Hence, Line 8 does not create a cycle from acyclic $\mathcal{G}_\mathcal{S}$. Through Lines 2 and 8, a subset of or an element of $\mathbf{X}(\mathcal{S})$ is removed. Not only are its induced edges (i.e., $C \to X$ for $C \in \mathbf{C}_X$) removed, but also its original parents in $\mathcal{G}$ are restored. Since the removal of incoming edges onto X does not create a cyclic path, we examine directed edges from $pa(X)_\mathcal{G}$ to X with respect to the acyclicity of the updated induced graph after Line 2 and 8.

(Line 2) Let $\mathcal{S}$ and $\mathcal{S}'$ be MPSes before and after removing $\mathbf{X}' \subseteq an(Y)_{\mathcal{G}_\mathcal{S}}$ at Line 2. If $\mathbf{X}' = \emptyset$, then done. Otherwise, let $\mathcal{H}$ be $\mathcal{G}_\mathcal{S}[An(Y)_{\mathcal{G}_\mathcal{S}}]$ the induced graph restricted to the ancestors of $\mathbf{Y}$, which is the same after the removal, $\mathcal{G}_{\mathcal{S}'}[An(Y)_{\mathcal{G}_{\mathcal{S}'}}]$. Since $\mathcal{G}_\mathcal{S}$ is acyclic, its subgraph $\mathcal{H}$ is acyclic, as well. Let $\mathcal{H}'$ be $\mathcal{G}_{\mathcal{S}'} \setminus \mathbf{V}(\mathcal{H})$, the vertex-induced subgraph of $\mathcal{G}_{\mathcal{S}'}$ by excluding the ancestors of $\mathbf{Y}$. By definition of $\mathcal{S}'$, there is no $X \in \mathbf{X}(\mathcal{S}')$ in $\mathcal{H}'$ and every directed edge in $\mathcal{H}'$ is those in $\mathcal{G}$. That is $\mathcal{H}'$ itself is acyclic. Hence, we only need to check whether there can be a cycle formed across $\mathcal{H}$ and $\mathcal{H}'$. By definition of $\mathcal{S}'$, there cannot be a directed edge from a vertex in $\mathcal{H}'$ to other vertex in $\mathcal{H}'$. Therefore, it is impossible for a cycle to exist in $\mathcal{G}_{\mathcal{S}'}$.

(Line 8) Let X be the action to be removed. Let $\mathcal{H}' = \mathcal{G}_{\mathcal{S}'}[De(X)_{\mathcal{G}_{\mathcal{S}'}}]$ and $\mathcal{H} = \mathcal{G}_{\mathcal{S}'} \setminus \mathbf{V}(\mathcal{H}')$. $\mathcal{H}$ is acyclic since it is the same as $\mathcal{G}_\mathcal{S} \setminus \mathbf{V}(\mathcal{H}')$, the subgraph of acyclic $\mathcal{G}_\mathcal{S}$. By construction, there exists no $X' \in \mathbf{X}(\mathcal{S}')$ other than X in $\mathcal{H}'$ and $\mathcal{H}'$ is a vertex-induced subgraph of $\mathcal{G}$. Hence, $\mathcal{H}'$ is acyclic, too. By definition, there exists no directed edge from $\mathbf{V}(\mathcal{H}')$ to $\mathbf{V}(\mathcal{H})$ in $\mathcal{G}_{\mathcal{S}'}$. Consequently, $\mathcal{G}_{\mathcal{S}'}$ is acyclic. □

Given Prop. 6, we focus on the uniqueness and maximal non-redundancy of a returned MPS from nr-mps.

Theorem 4. *Given an MPS $\mathcal{S}$,* nr-mps *returns a unique, maximal non-redundant MPS of $\mathcal{S}$.*

Proof. The algorithm refines a given mixed policy scope (MPS) $\mathcal{S}$ by iterating over $\mathbf{X}(\mathcal{S})$ and $\mathbf{C}(\mathcal{S})$ *once*. We examine whether the dependency that holds at the testing time is still valid with respect to the returned MPS, which we will be denoted by, $\mathcal{S}^\perp$. Let $\mathcal{S}'$ be an intermediate MPS when an arbitrary X is under examination (Line 4). $X \in an(Y)$ in $\mathcal{G}_{\mathcal{S}'}$ is preserved in $\mathcal{G}_{\mathcal{S}^\perp}$ because the later tests are irrelevant as they are all non-successors of X while the ancestrality is only relevant to its successors.

Next consider examining $C \in \mathbf{C}_X$ for some X where now $\mathcal{S}'$ is the MPS at Line 7. Consider a trail (d-connection path) ρ between C and Y in $\mathcal{G}_{\mathcal{S}'}$. We restrict our attention to a collider-minimal, shortest path. Every collider W in the path has a directed path towards Y through $W \leadsto C_i \to X \leadsto Y$ where $C_i \in \mathbf{C}_X$ by the testing criteria and $W \leadsto C_i$ can be of zero length. Let $\mathcal{G}_\rho \subseteq \mathcal{G}_{\mathcal{S}'}$ be the path graph together with directed paths between colliders and conditionals. Let $\mathcal{S}''$ be the MPS at the end of testing every $C_j \in \mathbf{C}_X$. We show that $(C \not\perp\!\!\!\perp Y \mid \mathbf{C}''_X)$ in $\mathcal{G}_{\mathcal{S}''} \backslash \{X\}$ holds true (with $\mathbf{C}''_X$ as in $\mathcal{S}''$). Let $\mathbf{C}_i$ be the subset of $\mathbf{C}_X$ that associates with the colliders in the path. Since every $C_i \in \mathbf{C}_i$ will have a back-door path to Y (by concatenating a directed path between C_i and W and the subpath of ρ between W and Y) given $\mathbf{C}_i \backslash \{C_i\}$, $\mathbf{C}_i$ is the subset of $\mathbf{C}''_X$. Hence, the result follows (for an illustrative example, please see Fig. 14).

Now we investigate whether the path ρ between C and Y is still valid in $\mathcal{G}_{\mathcal{S}^\perp}$ given $\mathbf{C}^\perp_X$. Specifically, we would like to ensure that the edges in the path are intact throughout the changes made by the algorithm. The removal of $X' \in \mathbf{X}(\mathcal{S})_{\not\succ X}$ may affect the parents of X' and the removal of $C' \in \mathbf{C}_{X'}$ affects $C' \to X'$. That is, for both cases, we investigate whether $C' \to X'$ in ρ is intact at the end of the algorithm. We first state two claims.

Claim 1. *X' has a directed path to Y in $\mathcal{G}_\rho$.*

Claim 2. *$C' \not\perp\!\!\!\perp Y$ in $\mathcal{G}_\rho \backslash \{X'\}$ demonstrates the existence of a collider-free d-connection ϕ, which is disjoint with $\mathbf{C}_{X'} \backslash \{C'\}$.*

The directed path between X' and Y will be valid in $\mathcal{G}_{\mathcal{S}^\perp}$ if every $C'' \to X''$ appeared in the path is intact, which delegates its validity to the bottom-most $X'' \in \mathbf{X}(\mathcal{S})_{\not\succ X}$. $C' \in \mathbf{C}_{X'}$ will be dependent to Y given $\mathbf{C}_{X'} \backslash \{C''\}$ as no $\mathbf{C}_{X'} \backslash \{C'\}$ exists in $\mathcal{G}_\rho$. Hence, by tracing back the validity of each policy-induced edge in $\mathcal{G}_\rho$, we conclude that every policy-induced edge is intact, and, thus, $\mathcal{G}_\rho \subseteq \mathcal{G}_{\mathcal{S}^\perp}$ and $C' \in \mathbf{C}^\perp_X$. $\square$

Now we prove the two claims in the above proof.

Claim 1. *X' has a directed path to Y in $\mathcal{G}_\rho$.*

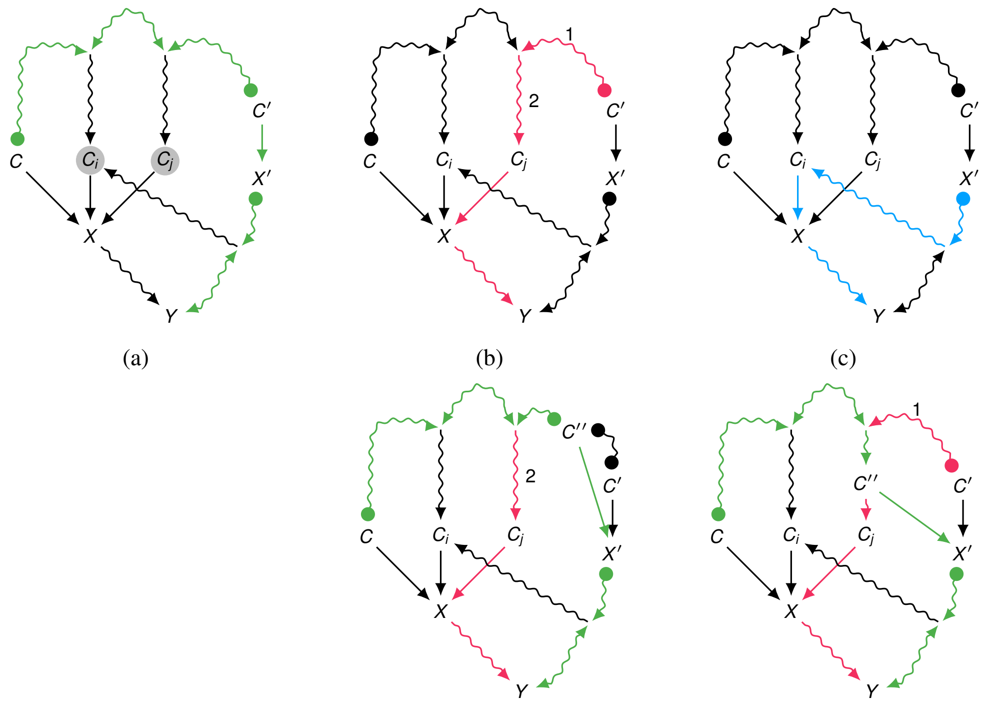
Proof. Let $\bullet$ be an unspecified edge mark representing either arrow or tail and a squiggly edge represents a path. An abstract representation for the path ρ can be one of the following two forms $C \overset{\times}{\bullet\!\sim\!\bullet} C' {\to} X' \overset{\times}{\bullet\!\leadsto} Y$ or $C \overset{\times}{\bullet\!\sim\!\bullet} X' {\leftarrow} C' \overset{\times}{\bullet\!\leadsto} Y$ with $\times$ represents that the path may have colliders. X' has a directed path to (i) C, (ii) Y, or (iii) some $C_a \in \mathbf{C}_X \backslash \{C\}$ via a collider W (which can be X' itself) in the path. Then, a directed path can be of the form: (i) $X' \leadsto C \to X \leadsto Y$, (ii) $X' \leadsto Y$, or (iii) $X' \leadsto W \leadsto C_a \to X \leadsto Y$. $\square$

Claim 2. *$C' \not\perp\!\!\!\perp Y$ in $\mathcal{G}_\rho \backslash \{X'\}$ demonstrates the existence of a collider-free d-connection ϕ, which is disjoint with $\mathbf{C}_{X'} \backslash \{C'\}$.*

Proof. Similar to the proof for the directed path between X' and Y, our abstract representation informs us that we can consider two subpaths (a) $C \overset{\times}{\bullet\!\sim\!\bullet} C'$ or (b) $C' \overset{\times}{\bullet\!\leadsto} Y$ where both avoids passing through X'.

For (a), if the subpath does not contain a collider, (a_1) $C' \bullet\!\sim\!\bullet C \to X \leadsto Y$ is a valid trail signaling $C' \not\perp\!\!\!\perp Y$ in $\mathcal{G}_\rho \backslash \{X'\}$. Otherwise, there exists a collider W in the subpath, which uses a shortcut to Y through $C_a \in \mathbf{C}_X \backslash \{X\}$: (a_2) $C' \bullet\!\leadsto W \leadsto C_a \to X \leadsto Y$.

In case of (a_1), $C'' \in \mathbf{C}_{X'} \backslash \{C'\}$ cannot appear in between C' and C since otherwise it violates the fact that ρ being the shortest—one can create ρ' by replacing $C \ldots C'' \ldots C' \to X$ by $C \ldots C'' \to X$. In case of (a_2), similarly, C'' cannot reside along $C' \bullet\!\leadsto W$. In addition, we prove that C'' does not exist in-between $W \leadsto C_a$ since otherwise there exists ρ' which does not require W in the path ρ as a collider making use of $C'' \to X'$, contradicting the collider-minimality of ρ. In either cases (a_1) and (a_2), $C \to X \leadsto Y$ or $C_a \to X \leadsto Y$ are C''-free otherwise it contradicts the topological order between X' and X.



(a) (b) (c)

(d) violating shortestness (e) violating minimality

Figure 14: An example illustrating the preservation of a d-connecting path in Alg. 2. If $C'' \in \mathbf{C}_{X'}$ in 1 (b), $C'' \to X$ can make the green path shorter, contradicting shortest path. If in 2 (b), the 2nd collider from C is not required contradicting minimality. Since X' is not a successor of X, no C'' can appear between the path between X and Y. Colors here do not follow our convention made in the paper.

In case of (b), either there exists (b_1) a path towards Y, $C' \bullet\!\!\rightsquigarrow Y$ without a collider, or (b_2) through a collider as seen in the case of (a_2), where the same proof is applicable. Given (b_1), again, if C'' exists in the path, we can shorten ρ by connecting $X' \leftarrow C''$ while replacing $X' \leftarrow C' \bullet\!\!\rightsquigarrow C''$, which violates ρ being the shortest.

Hence, the result follows. $\square$

E.1 Non-Redundancy under Optimality

Derivations for the Redundancy of Examples We demonstrate the redundancy for Figs. 5b to 5d. We may employ '$\leq$' to highlight that fixing operation can improve the expected reward although, given the optimality of the left hand side, it becomes '$=$'. We present a derivation for Fig. 5b showing that C_3 is non-informative.

$$\mu_{\boldsymbol{\pi}} = \sum_{y,\mathbf{x},\mathbf{c}} yQ(y|\mathbf{x},\mathbf{c})Q(\mathbf{x},\mathbf{c}) \qquad \text{MP,CR} \quad (40)$$

$$= \sum_{y,\mathbf{x},\mathbf{c}} yQ(y|\mathbf{x},c_1,c_2)Q(\mathbf{x},\mathbf{c}) \qquad \text{R1} \quad (41)$$

$$= \sum_{y,\mathbf{x},\mathbf{c}} yQ'_{\mathbf{x}}(y|c_1,c_2)Q(\mathbf{x},\mathbf{c}) \qquad \text{R1,R2} \quad (42)$$

$$= \sum_{y,\mathbf{x},\mathbf{c}} yQ'_{\mathbf{x}}(y|c_1,c_2)Q(x_1|c_3,c_1)Q(x_2|c_3,c_2)Q(\mathbf{c}) \qquad \text{CR,R1} \quad (43)$$

$$= \sum_{y,\mathbf{x},c_1,c_2} yQ'_{\mathbf{x}}(y|c_1,c_2)Q(c_1,c_2)\sum_{c_3} Q(x_1|c_3,c_1)Q(x_2|c_3,c_2)Q(c_3) \qquad \text{MP,CR} \quad (44)$$

$$= \sum_{c_3} Q(c_3) \sum_{y,\mathbf{x},c_1,c_2} yQ'_{\mathbf{x}}(y|c_1,c_2)Q(c_1,c_2)Q(x_1|c_3,c_1)Q(x_2|c_3,c_2) \quad \text{algebra} \quad (45)$$

$$\leq \sum_{y,\mathbf{x},c_1,c_2} yQ'_{\mathbf{x}}(y|c_1,c_2)Q(c_1,c_2)Q(x_1|c_3^*,c_1)Q(x_2|c_3^*,c_2) \quad \text{def} \quad (46)$$

$$= \sum_{y,\mathbf{x},c_1,c_2} yQ'_{\mathbf{x}}(y|c_1,c_2)Q(c_1,c_2)\pi'(x_1|c_1)\pi'(x_2|c_2) \quad \text{def} \quad (47)$$

$$= \sum_{y,\mathbf{x},c_1,c_2} yQ'_{\mathbf{x}}(y|c_1,c_2)Q'_{\mathbf{x}}(c_1,c_2)\pi'(x_1|c_1)\pi'(x_2|c_2) \quad \text{R1,R3} \quad (48)$$

$$= \sum_{y,\mathbf{x},c_1,c_2} yQ'_{\mathbf{x}}(y,c_1,c_2)\pi'(x_1|c_1)\pi'(x_2|c_2) \quad \text{CR} \quad (49)$$

where $c_3^* \in \mathfrak{X}_{C_3}$ is the value maximizing the inner sum.

The derivation for Fig. 5c is given that C_2 is non-informative.

$$\mu_{\boldsymbol{\pi}} = \sum_{y,\mathbf{x},\mathbf{c}} yQ(y|\mathbf{x},\mathbf{c})Q(\mathbf{x},\mathbf{c}) \quad \text{def} \quad (50)$$

$$= \sum_{y,\mathbf{x},\mathbf{c}} yQ(y|\mathbf{x},c_1)Q(\mathbf{x},\mathbf{c}) \quad \text{R1} \quad (51)$$

$$= \sum_{y,\mathbf{x},\mathbf{c}} yQ'_{\mathbf{x}}(y|c_1)Q(\mathbf{x},\mathbf{c}) \quad \text{R1,R2} \quad (52)$$

$$= \sum_{y,\mathbf{x},\mathbf{c}} yQ'_{\mathbf{x}}(y|c_1)Q(c_1)Q(c_2|c_1)Q(x_1|\mathbf{c})Q(x_2|\mathbf{c}) \quad \text{CR} \quad (53)$$

$$= \sum_{y,\mathbf{x},\mathbf{c}} yQ'_{\mathbf{x}}(y|c_1)Q'_{\mathbf{x}}(c_1)Q(c_2|c_1)Q(x_1|\mathbf{c})Q(x_2|\mathbf{c}) \quad \text{R1,R3} \quad (54)$$

$$= \sum_{y,\mathbf{x},\mathbf{c}} yQ'_{\mathbf{x}}(y,c_1)Q(c_2|c_1)Q(x_1|\mathbf{c})Q(x_2|\mathbf{c}) \quad \text{CR} \quad (55)$$

$$= \sum_{y,\mathbf{x},c_1} yQ'_{\mathbf{x}}(y,c_1) \sum_{c_2} Q(c_2|c_1)Q(x_1|\mathbf{c})Q(x_2|\mathbf{c}) \quad \text{algebra} \quad (56)$$

$$= \sum_{c_1} \sum_{c_2} Q(c_2|c_1) \sum_{y,\mathbf{x}} yQ'_{\mathbf{x}}(y,c_1)Q(x_1|\mathbf{c})Q(x_2|\mathbf{c}) \quad \text{algebra} \quad (57)$$

$$\leq \sum_{c_1} \sum_{y,\mathbf{x}} yQ'_{\mathbf{x}}(y,c_1)Q(x_1|c_1,c_2^*(c_1))Q(x_2|c_1,c_2^*(c_1)) \quad \text{def} \quad (58)$$

$$= \sum_{y,\mathbf{x},c_1} yQ'_{\mathbf{x}}(y,c_1)\pi'(x_1|c_1)\pi'(x_2|c_1) \quad \text{def} \quad (59)$$

The derivation for Fig. 5d is as follows. Let $Q' = P_{\boldsymbol{\pi}\setminus\{X_1,X_2\}}$.

$$\mu_{\boldsymbol{\pi}} = \sum_{y,\mathbf{x}_{12},c_2} yQ(y|\mathbf{x}_{12},c_2)Q(\mathbf{x}_{12},c_2) \quad \text{MP,CR} \quad (60)$$

$$= \sum_{y,\mathbf{x}_{12},c_2} yQ'_{\mathbf{x}_{12}}(y|c_2)Q(\mathbf{x}_{12},c_2) \quad \text{R1,R2} \quad (61)$$

$$= \sum_{y,\mathbf{x}_{12},c_2} yQ'_{\mathbf{x}_{12}}(y|c_2) \sum_{x_3} Q(\mathbf{x},c_2) \quad \text{MP} \quad (62)$$

$$= \sum_{y,\mathbf{x}_{12},c_2} yQ'_{\mathbf{x}_{12}}(y|c_2) \sum_{x_3} Q(c_2)Q(x_3|c_2)Q(x_1|x_3,c_2)Q(x_2|x_3,c_2,x_1) \quad \text{CR} \quad (63)$$

$$= \sum_{y,\mathbf{x}_{12},c_2} yQ'_{\mathbf{x}_{12}}(y|c_2)Q'_{\mathbf{x}_{12}}(c_2) \sum_{x_3} Q(x_3|c_2)Q(x_1|x_3,c_2)Q(x_2|x_3,c_2) \quad \text{R2,R3} \quad (64)$$

$$= \sum_{y,\mathbf{x}_{12},c_2} yQ'_{\mathbf{x}_{12}}(y,c_2) \sum_{x_3} Q(x_3|c_2)Q(x_1|x_3,c_2)Q(x_2|x_3,c_2) \quad \text{CR} \quad (65)$$

$$= \sum_{c_2} \sum_{x_3} Q(x_3|c_2) \sum_{y,\mathbf{x}_{12}} yQ'_{\mathbf{x}_{12}}(y, c_2) Q(x_1|x_3, c_2) Q(x_2|x_3, c_2) \quad \text{algebra} \quad (66)$$

$$\leq \sum_{y,\mathbf{x}_{12},c_2} yQ'_{\mathbf{x}_{12}}(y, c_2) Q(x_1|x_3^*(c_2), c_2) Q(x_2|x_3^*(c_2), c_2) \quad \text{def} \quad (67)$$

$$= \sum_{y,\mathbf{x}_{12},c_2} yQ'_{\mathbf{x}_{12}}(y, c_2) \pi'(x_1|c_2) \pi'(x_2|c_2) \quad \text{def} \quad (68)$$

Proofs and Additional Characterizations

Lemma 1. *Given an MPS $\mathcal{S}$, which satisfies non-redundancy (Thm. 1), let $\mathbf{X}' \subseteq \mathbf{X}(\mathcal{S})$, actions of interest, $\mathbf{C}' \subsetneq \mathbf{C}_{\mathbf{X}'} \backslash \mathbf{X}'$. non-action contexts of interest. If there exists a subset of exogenous variables $\mathbf{U}'$ in $\mathcal{G}_{\mathcal{S}}$, a subset of endogenous variables $\mathbf{Z}$ in $\mathcal{G}_{\mathcal{S}}$ that disjoints with $\mathbf{C}' \dot\cup \mathbf{X}'$ and subsumes $\mathbf{C}_{\mathbf{X}'} \backslash (\mathbf{C}' \dot\cup \mathbf{X}')$, and an order $\prec$ over $\mathbf{V}' \doteq \mathbf{C}' \dot\cup \mathbf{X}' \dot\cup \mathbf{Z}$ such that*

1. *$(Y \perp\!\!\!\perp \boldsymbol{\pi}_{\mathbf{X}'} \mid \lceil \mathbf{X}' \dot\cup \mathbf{C}' \rceil)_{\mathcal{G}_{\mathcal{S}}}$,*
2. *$(C \perp\!\!\!\perp \boldsymbol{\pi}_{\mathbf{X}'_{\prec C}}, \mathbf{Z}_{\prec C}, \mathbf{U}' \mid \lceil (\mathbf{X}' \dot\cup \mathbf{C}')_{\prec C} \rceil)_{\mathcal{G}_{\mathcal{S}}}$ for every $C \in \mathbf{C}'$, and*
3. *$\mathbf{V}'_{\prec X}$ is disjoint with $de(X)_{\mathcal{G}_{\mathcal{S}}}$ and subsumes $pa(X)_{\mathcal{G}_{\mathcal{S}}}$ for every $X \in \mathbf{X}'$,*

then, the expected reward for $\boldsymbol{\pi}$, a deterministic policy optimal with respect to $\mathcal{S}$, can be written as

$$\mu_{\boldsymbol{\pi}} = \sum_{y,\mathbf{c}',\mathbf{x}'} yQ'_{\mathbf{x}'}(y, \mathbf{c}') \sum_{\mathbf{u}',\mathbf{z}} Q(\mathbf{u}') \prod_{Z \in \mathbf{Z}} Q(z|\mathbf{v}'_{\prec z}, \mathbf{u}') \prod_{X \in \mathbf{X}'} \pi(x|\mathbf{c}_x). \quad (3)$$

Proof. We derive the equality using the definitions, axioms of probability, and the given conditions.

$$\mu_{\boldsymbol{\pi}} = \sum_{y} yQ(y) \quad \text{def} \quad (69)$$

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} yQ(y, \mathbf{c}', \mathbf{x}') \quad \text{MP} \quad (70)$$

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} yQ(y|\mathbf{c}', \mathbf{x}') Q(\mathbf{c}', \mathbf{x}') \quad \text{CR} \quad (71)$$

The first condition can be viewed as both (i) rule 1 of do-calculus with respect to the regime nodes $\boldsymbol{\pi}_{\mathbf{X}'}$ and (ii) rule 2 of do-calculus with respect to $\mathbf{X}'$. Then,

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} yQ'_{\mathbf{x}'}(y|\mathbf{c}') Q(\mathbf{c}', \mathbf{x}') \quad \text{C1} \quad (72)$$

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} yQ'_{\mathbf{x}'}(y|\mathbf{c}') \sum_{\mathbf{z},\mathbf{u}'} Q(\mathbf{c}', \mathbf{x}', \mathbf{z}, \mathbf{u}') \quad \text{MP} \quad (73)$$

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} yQ'_{\mathbf{x}'}(y|\mathbf{c}') \sum_{\mathbf{z},\mathbf{u}'} \prod_{C \in \mathbf{C}'} Q(c|\mathbf{v}'_{\prec c}, \mathbf{u}') Q(\mathbf{u}') \prod_{V \in \mathbf{X}' \cup \mathbf{Z}} Q(v|\mathbf{v}'_{\prec v}, \mathbf{u}') \quad \text{CR} \quad (74)$$

The second condition corresponds to (i) rule 1 for $\mathbf{Z}_{\prec C}$ and $\mathbf{U}'$ and (ii) rule 2 for $\mathbf{X}'_{\prec C}$,

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} yQ'_{\mathbf{x}'}(y|\mathbf{c}') \prod_{C \in \mathbf{C}'} Q'(c|\mathbf{x}'_{\prec c}, \mathbf{c}'_{\prec c}) \sum_{\mathbf{z},\mathbf{u}'} Q(\mathbf{u}') \prod_{V \in \mathbf{X}' \cup \mathbf{Z}} Q(v|\mathbf{v}'_{\prec v}, \mathbf{u}') \quad \text{C2} \quad (75)$$

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} yQ'_{\mathbf{x}'}(y|\mathbf{c}') \prod_{C \in \mathbf{C}'} Q'_{\mathbf{x}'_{\prec c}}(c|\mathbf{c}'_{\prec c}) \sum_{\mathbf{z},\mathbf{u}'} Q(\mathbf{u}') \prod_{V \in \mathbf{X}' \cup \mathbf{Z}} Q(v|\mathbf{v}'_{\prec v}, \mathbf{u}') \quad \text{C2} \quad (76)$$

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} yQ'_{\mathbf{x}'}(y|\mathbf{c}') \prod_{C \in \mathbf{C}'} Q'_{\mathbf{x}'}(c|\mathbf{c}'_{\prec c}) \sum_{\mathbf{z},\mathbf{u}'} Q(\mathbf{u}') \prod_{V \in \mathbf{X}' \cup \mathbf{Z}} Q(v|\mathbf{v}'_{\prec v}, \mathbf{u}') \quad \text{R3} \quad (77)$$

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} yQ'_{\mathbf{x}'}(y|\mathbf{c}') Q'_{\mathbf{x}'}(\mathbf{c}') \sum_{\mathbf{z},\mathbf{u}'} Q(\mathbf{u}') \prod_{V \in \mathbf{X}' \cup \mathbf{Z}} Q(v|\mathbf{v}'_{\prec v}, \mathbf{u}') \quad \text{CR} \quad (78)$$

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} yQ'_{\mathbf{x}'}(y, \mathbf{c}') \sum_{\mathbf{z},\mathbf{u}'} Q(\mathbf{u}') \prod_{Z \in \mathbf{Z}} Q(z|\mathbf{v}'_{\prec z}, \mathbf{u}') \prod_{X \in \mathbf{X}'} Q(x|\mathbf{v}'_{\prec x}, \mathbf{u}') \quad \text{CR} \quad (79)$$

The third condition ensures that the conditionals for a term $Q(x|\cdot)$ can be refined only to the parents of X, which is $\mathbf{C}_X$,

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} y Q'_{\mathbf{x}'}(y,\mathbf{c}') \sum_{\mathbf{z},\mathbf{u}'} Q(\mathbf{u}') \prod_{Z\in\mathbf{Z}} Q(z|\mathbf{v}'_{\prec z},\mathbf{u}') \prod_{X\in\mathbf{X}'} Q(x|\mathbf{c}_x) \qquad \text{C3,R1} \qquad (80)$$

□

Theorem 2. *Let $\mathbf{U}'$, $\mathbf{Z}$, and $\prec$ satisfy Lemma 1. For $Z\in\mathbf{Z}$, let $\mathbf{V}_Z$ be a minimal subset of $\mathbf{V}'_{\prec Z}\cup\mathbf{U}'$ such that $Q(Z\mid\mathbf{V}_Z)=Q(Z\mid\mathbf{V}'_{\prec Z},\mathbf{U}')$. We define $\mathsf{fix}(\mathbf{T})$ with respect to $\{\langle Z,\mathbf{V}_Z\rangle\}_{Z\in\mathbf{Z}}$, that is, with $\hat{\mathbf{T}}\doteq\lceil\mathbf{T}\rceil\cup\{Z\in\mathbf{Z}\mid\mathbf{V}_Z\backslash\mathbf{U}'\subseteq\lceil\mathbf{T}\rceil\}$, $\mathsf{fixed}(\mathbf{T})$ is $\mathbf{T}$ if $\mathbf{T}=\hat{\mathbf{T}}$ and $\mathsf{fixed}(\hat{\mathbf{T}})$, otherwise. If $\mathsf{fixed}(\mathbf{C}_X\backslash\mathbf{Z})\supseteq\mathbf{C}_X$ for $X\in\mathbf{X}'$, then, $\mathcal{S}'\doteq(\mathcal{S}\backslash\mathbf{X}')\cup\{\langle X,\mathbf{C}_X\backslash\mathbf{Z}\rangle\}_{X\in\mathbf{X}'}$ satisfies $\mu^*_{\mathcal{S}'}=\mu^*_{\mathcal{S}}$.*

Proof. Let $Q=P_{\boldsymbol{\pi}}$ and $Q'=P_{\boldsymbol{\pi}\backslash\mathbf{X}'}$.

$$\begin{aligned}
\mu_{\boldsymbol{\pi}} &= \sum_{y,\mathbf{c}',\mathbf{x}'} y Q'_{\mathbf{x}'}(y,\mathbf{c}') \sum_{\mathbf{z},\mathbf{u}'} Q(\mathbf{u}') \prod_{Z\in\mathbf{Z}} Q(z|\mathbf{v}'_{\prec z},\mathbf{u}') \prod_{X\in\mathbf{X}'} Q(x|\mathbf{c}_x) && \text{Lemma 1}\\
&= \sum_{y,\mathbf{c}',\mathbf{x}'} y Q'_{\mathbf{x}'}(y,\mathbf{c}') \sum_{\mathbf{z},\mathbf{u}'} Q(\mathbf{u}') \prod_{Z\in\mathbf{Z}} Q(z|\mathbf{v}_z) \prod_{X\in\mathbf{X}'} Q(x|\mathbf{c}_x\backslash\mathbf{z},\mathbf{c}_x\cap\mathbf{z}) && \text{def}\\
&= \sum_{\mathbf{u}'} Q(\mathbf{u}') \Big(\sum_{y,\mathbf{c}',\mathbf{x}'} y Q'_{\mathbf{x}'}(y,\mathbf{c}') \sum_{\mathbf{z}} \prod_{Z\in\mathbf{Z}} Q(z|\mathbf{v}_z) \prod_{X\in\mathbf{X}'} Q(x|\mathbf{c}_x\backslash\mathbf{z},\mathbf{c}_x\cap\mathbf{z}) \Big) && \text{algebra}
\end{aligned}$$

Choose $\mathbf{u}'^*$ maximizing the term in the parentheses (note that $\mathbf{V}_Z$ may intersect $\mathbf{U}'$),

$$\le \sum_{y,\mathbf{c}',\mathbf{x}'} y Q'_{\mathbf{x}'}(y,\mathbf{c}') \sum_{\mathbf{z}} \prod_{Z\in\mathbf{Z}} Q(z|\mathbf{v}_z\backslash\mathbf{U}',(\mathbf{V}_Z\cap\mathbf{u}'^*)) \prod_{X\in\mathbf{X}'} Q(x|\mathbf{c}_x\backslash\mathbf{z},\mathbf{c}_x\cap\mathbf{z}) \qquad \text{def}$$

Take $Z_1\in\mathbf{Z}$, the first element among $\mathbf{Z}$ with respect to $\prec$.

$$= \sum_{y,\mathbf{c}',\mathbf{x}'} y Q'_{\mathbf{x}'}(y,\mathbf{c}') \sum_{z_1} Q(z_1|\mathbf{v}_{z_1}\backslash\mathbf{U}',(\mathbf{V}_{Z_1}\cap\mathbf{u}'^*)) \sum_{\mathbf{z}\backslash\{z_1\}} \prod_{Z\in\mathbf{Z}\backslash\{Z_1\}} \cdots$$

Choose z_1^* for each free variable $\mathbf{v}_{z_1}\backslash\mathbf{U}'$,

$$\begin{aligned}
\le \sum_{y,\mathbf{c}',\mathbf{x}'} y Q'_{\mathbf{x}'}(y,\mathbf{c}') \sum_{\mathbf{z}\backslash\{z_1\}} \prod_{Z\in\mathbf{Z}\backslash\{Z_1\}} & Q(z|\mathbf{v}_z\backslash(\mathbf{U}'\cup\{Z_1\}),(\mathbf{V}_Z\cap\{\mathbf{u}'^*,z_1^*(\mathbf{v}_{z_1}\backslash\mathbf{U}',\mathbf{v}_{z_1}\cap\mathbf{u}'^*)\}))\\
&\prod_{X\in\mathbf{X}'} Q(x|\mathbf{c}_x\backslash\mathbf{Z},\mathbf{c}_x\cap(\mathbf{z}\backslash\{Z_1\},z_1^*(\mathbf{v}_{z_1}\backslash\mathbf{U}',\mathbf{v}_{z_1}\cap\mathbf{u}'^*)))
\end{aligned}$$

Consider repeating this procedure for $\mathbf{Z}=\{Z_1,\ldots,Z_m\}$ (following the order $\prec$). We will define z_i^* for each free variables and fixed variables within $\mathbf{v}_{z_i}$. For readability, we define functions for z_i^* recursively as

$$\begin{aligned}
z_1^*(\cdot) &= z_1^*(\mathbf{v}_{z_1}\backslash\mathbf{U}',\mathbf{v}_{z_1}\cap\mathbf{u}'^*) && \text{base case}\\
z_i^*(\cdot) &= z_i^*(\mathbf{v}_{z_i}\backslash(\mathbf{U}'\cup\mathbf{Z}_{<i}),\mathbf{v}_{z_i}\cap(\mathbf{u}'^*\cup\{z_j^*(\cdot)\}_{j=1}^{i-1})) && \text{for } i>1
\end{aligned}$$

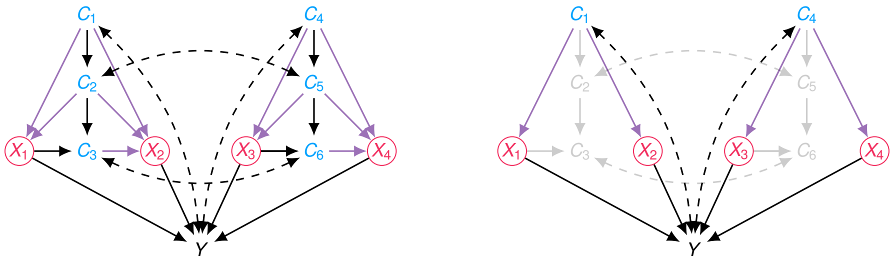
Then,

$$\mu_{\boldsymbol{\pi}} \le \sum_{y,\mathbf{c}',\mathbf{x}'} y Q'_{\mathbf{x}'}(y,\mathbf{c}') \prod_{X\in\mathbf{X}'} Q(x|\mathbf{c}_x\backslash\mathbf{Z},\mathbf{c}_x\cap\{z_i^*(\cdot)\}_{i=1}^m).$$

We want to ensure that $\mathbf{c}_x\cap\{z_i^*(\cdot)\}_{i=1}^m$ is a function of $\mathbf{c}_x\backslash\mathbf{Z}$. By the construction, we can examine the dependence structure defined by the conditionals specified in terms $Q(Z|\cdot)$. Further, we can utilize the deterministic mechanisms of $\mathbf{X}(\mathcal{S})$. If the values for $\mathbf{C}_X\cap\mathbf{Z}$ can be determined from the value of $\mathbf{C}_X\backslash\mathbf{Z}$, then, for some π',

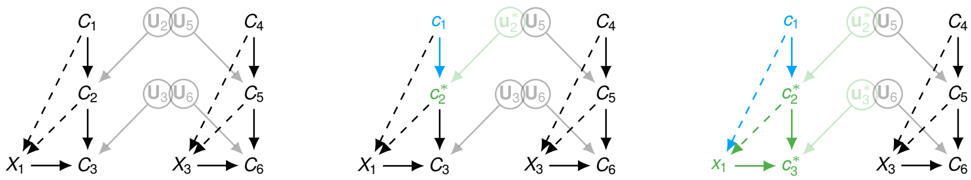
$$\mu_{\boldsymbol{\pi}} \le \sum_{y,\mathbf{c}',\mathbf{x}'} y Q'_{\mathbf{x}'}(y,\mathbf{c}') \prod_{X\in\mathbf{X}'} \pi'(x|\mathbf{c}_x\backslash\mathbf{Z}).$$

This completes the proof. □



(a) a redundant MPS (b) a non-redundant MPS

Figure 15: A more involved example for redundancies in a mixed policy scope



(a) (b) X_1 (c) X_2

Figure 16: (a) a dependency specified $Q(Z|\cdot)$ terms with additional 'implying' relationships (directed dashed edges onto X_1 and X_3) and exogenous variables; (b) X_1 requires to fix C_2 given C_1 (blue for given) where an unknown $\mathbf{u}_2^*$ can be marginally fixed and $c_2^*(c_1, \mathbf{u}_2^*)$ can be inferred (green for inferred); (c) X_2 requires both C_2 and C_3 to be fixed where they can be sequentially inferred where X_1 is implied by given c_1 and inferred $c_2^*(c_1, \mathbf{u}_2^*)$. Results for X_3 and X_4 are similar to X_1 and X_2.

An example Fig. 15 is given accompanied with a derivation below. Consider an order of $\prec = \langle C_1, C_4, C_2, C_5, X_1, X_3, C_3, C_6, X_2, X_4 \rangle$.

$$
\begin{aligned}
&\mu_{\boldsymbol{\pi}} \\
&= \sum_{y,\mathbf{x},\mathbf{c}_{14}} y Q(y|\mathbf{x}, \mathbf{c}_{14}) Q(\mathbf{x}, \mathbf{c}_{14}) \\
&= \sum_{y,\mathbf{x},\mathbf{c}_{14}} y Q'_{\mathbf{x}}(y|\mathbf{c}_{14}) Q(\mathbf{x}, \mathbf{c}_{14}) \\
&= \sum_{y,\mathbf{x},\mathbf{c}_{14}} y Q'_{\mathbf{x}}(y|\mathbf{c}_{14}) \sum_{\mathbf{c}_{2356},\mathbf{u}'} Q(\mathbf{c}_{2356}, \mathbf{x}, \mathbf{c}_{14}, \mathbf{u}') \\
&= \sum_{y,\mathbf{x},\mathbf{c}_{14}} y Q'_{\mathbf{x}}(y|\mathbf{c}_{14}) \sum_{\mathbf{c}_{2356},\mathbf{u}'} Q(\mathbf{u}') Q(c_{14}) Q(c_2|c_1, \mathbf{u}_2) Q(c_5|c_4, \mathbf{u}_5) Q(x_1|\mathbf{c}_{12}) \\
&\qquad Q(x_3|\mathbf{c}_{45}) Q(c_3|c_2, x_1, \mathbf{u}_3) Q(c_6|c_5, x_3, \mathbf{u}_6) Q(x_2|\mathbf{c}_{123}) Q(x_4|\mathbf{c}_{456}) && \text{CR,R1} \\
&= \sum_{y,\mathbf{x},\mathbf{c}_{14}} y Q'_{\mathbf{x}}(y, \mathbf{c}_{14}) \sum_{\mathbf{c}_{2356},\mathbf{u}'} Q(\mathbf{u}') Q(c_2|c_1, \mathbf{u}_2) Q(c_5|c_4, \mathbf{u}_5) Q(x_1|\mathbf{c}_{12}) \\
&\qquad Q(x_3|\mathbf{c}_{45}) Q(c_3|c_2, x_1, \mathbf{u}_3) Q(c_6|c_5, x_3, \mathbf{u}_6) Q(x_2|\mathbf{c}_{123}) Q(x_4|\mathbf{c}_{456}) && \text{R3,CR} \\
&= \sum_{\mathbf{c}_{2356},\mathbf{u}'} \underbrace{Q(\mathbf{u}') Q(c_2|c_1, \mathbf{u}_2) Q(c_5|c_4, \mathbf{u}_5) Q(c_3|c_2, x_1, \mathbf{u}_3) Q(c_6|c_5, x_3, \mathbf{u}_6)}_{\text{defines dependency}} \\
&\qquad \sum_{y,\mathbf{x},\mathbf{c}_{14}} y Q'_{\mathbf{x}}(y, \mathbf{c}_{14}) Q(x_1|c_1, \underbrace{c_2}_{\text{to fix}}) Q(x_2|c_1, \underbrace{\mathbf{c}_{23}}_{\text{to fix}}) Q(x_3|c_4, \underbrace{c_5}_{\text{to fix}}) Q(x_4|c_4, \underbrace{\mathbf{c}_{56}}_{\text{to fix}}) && \text{R1}
\end{aligned}
$$

Please see Fig. 16 how values can be properly fixed. Then,

$$
\leq \sum_{y,\mathbf{x},\mathbf{c}_{14}} y Q'_{\mathbf{x}}(y, \mathbf{c}_{14}) \pi'(x_1|c_1) \pi'(x_2|c_1) \pi'(x_3|c_4) \pi'(x_4|c_4)
$$

Remark 1. $\mathcal{S}' = \{\langle X_1, \{C_1\}\rangle, \langle X_2, \{C_1\}\rangle, \langle X_3, \{C_4\}\rangle, \langle X_4, \{C_4\}\rangle\}$ is non-redundant under optimality.

Proof. Let UCs between Y and C_1 and C_4 be U_1 and U_4 where each one is two-bit fair coins. Let $\mathbf{X}$ be binary variables. Let C_1 and C_4 copy U_1 and U_4, and Y take one minus the bit parity of four bits of $\mathbf{X}$ and the four bits of $\{U_1, U_4\}$. Hence, only when X_1, X_2, X_3, and X_4 pass the matching information, the expected reward becomes 1. Otherwise, the expected reward falls down to 0.5. $\square$

Even when C_2, C_3, C_5, C_6 are all confounded in Fig. 16, we can similarly elicit the same result as UCs affecting those removables are marginally fixable.

F A Partial Order over Mixed Policies and Possible-Optimality

Theorem 3. *Given an MPS $\mathcal{S}$, let $\mathcal{S}' \neq \mathcal{S}$ be an MPS with $\mathbf{X}(\mathcal{S}) \subseteq \mathbf{X}(\mathcal{S}')$ such that $\mathcal{H}''$ the union of induced graphs $\mathcal{G}_{\mathcal{S}} \cup \mathcal{G}_{\mathcal{S}'}$ is acyclic. Let $\mathbf{X}'$ be actions that the MPSes disagree on, i.e., $(\mathbf{X}(\mathcal{S}')\setminus\mathbf{X}(\mathcal{S}))\cup\{X \in \mathbf{X}(\mathcal{S}) \mid \mathbf{C}'_X \neq \mathbf{C}_X\}$, and (invalid) MPS $\mathcal{S}'' \doteq \{\langle X, pa(X)_{\mathcal{H}''} \cup \mathbf{U}_X\rangle\}_{X\in\mathbf{X}'}$. $\mu^*_{\mathcal{S}''} = \mu^*_{\mathcal{S}'}$ can be elicited by Thm. 2, then, $\mu^*_{\mathcal{S}} \leq \mu^*_{\mathcal{S}'}$.*

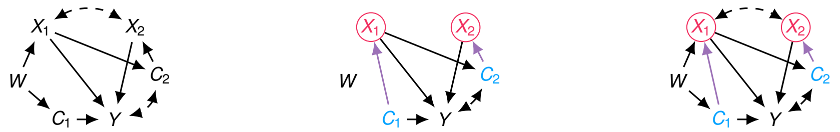
Proof. The functions for endogenous variables as in the definition of SCM are compatible with those decision rules employed in the definition of mixed policy. Further, treating non-contextualizable variables as if they are contextualizable ($\mathbf{C}^\star$) has no effect on the derivation. Hence, the superimposition of the two induced graphs resulting in an directed acyclic graph corresponds to a temporary, invalid MPS $\mathcal{S}''$ induced graph. Since more or the same contexts are used for $\mathcal{S}''$ than both $\mathcal{S}$ and $\mathcal{S}'$, $\mu^*_{\mathcal{S}} \leq \mu^*_{\mathcal{S}''}$ and $\mu^*_{\mathcal{S}'} \leq \mu^*_{\mathcal{S}''}$. If we elicit $\mu^*_{\mathcal{S}''} \leq \mu^*_{\mathcal{S}'}$, we can conclude that $\mu^*_{\mathcal{S}} \leq \mu^*_{\mathcal{S}''} = \mu^*_{\mathcal{S}'}$. $\square$

We present an example for $\mu^*_{\mathcal{S}} \leq \mu^*_{\mathcal{S}'}$ where Fig. 7a and Fig. 7b present $\mathcal{G}_{\mathcal{S}}$ and $\mathcal{G}_{\mathcal{S}'}$, respectively.

$$\begin{aligned}
\mu_{\mathcal{S}''} &= \sum_{y,\mathbf{c},\mathbf{x},\mathbf{u}_\mathbf{x}} yP_\mathbf{x}(y|\mathbf{c})Q(c_1)Q(\mathbf{u}_{x_1}|c_1)Q(\mathbf{u}_{x_2}|c_1,\mathbf{u}_{x_1})Q(x_1|c_1,\mathbf{u}_\mathbf{x})Q(c_2|x_1,c_1,\mathbf{u}_\mathbf{x})Q(x_2|x_1,\mathbf{c},\mathbf{u}_\mathbf{x})\\
&= \sum_{y,\mathbf{c},\mathbf{x},\mathbf{u}_\mathbf{x}} yP_\mathbf{x}(y|\mathbf{c})Q(c_1)Q(\mathbf{u}_{x_1}|c_1)Q(\mathbf{u}_{x_2}|c_1,\mathbf{u}_{x_1})Q(x_1|c_1,\mathbf{u}_{x_1})Q(c_2|x_1,c_1)Q(x_2|\mathbf{c},\mathbf{u}_{x_2})\\
&= \sum_{y,\mathbf{c},\mathbf{x},\mathbf{u}_\mathbf{x}} yP_\mathbf{x}(y|\mathbf{c})Q_\mathbf{x}(c_1)Q(\mathbf{u}_{x_1}|c_1)Q(\mathbf{u}_{x_2}|c_1,\mathbf{u}_{x_1})Q(x_1|c_1,\mathbf{u}_{x_1})Q_{x_1}(c_2|c_1)Q(x_2|\mathbf{c},\mathbf{u}_{x_2})\\
&= \sum_{y,\mathbf{c},\mathbf{x},\mathbf{u}_\mathbf{x}} yP_\mathbf{x}(y|\mathbf{c})Q_\mathbf{x}(c_1)Q(\mathbf{u}_{x_1}|c_1)Q(\mathbf{u}_{x_2}|c_1,\mathbf{u}_{x_1})Q(x_1|c_1,\mathbf{u}_{x_1})Q_\mathbf{x}(c_2|c_1)Q(x_2|\mathbf{c},\mathbf{u}_{x_2})\\
&= \sum_{y,\mathbf{c},\mathbf{x},\mathbf{u}_\mathbf{x}} yP_\mathbf{x}(y,\mathbf{c})Q(\mathbf{u}_{x_1}|c_1)Q(\mathbf{u}_{x_2}|c_1,\mathbf{u}_{x_1})Q(x_1|c_1,\mathbf{u}_{x_1})Q(x_2|\mathbf{c},\mathbf{u}_{x_2})\\
&\leq \sum_{y,\mathbf{c},\mathbf{x},\mathbf{u}_{x_2}} yP_\mathbf{x}(y,\mathbf{c})Q(\mathbf{u}_{x_2}|c_1,\mathbf{u}^*_{x_1}(c_1))Q(x_1|c_1,\mathbf{u}^*_{x_1}(c_1))Q(x_2|\mathbf{c},\mathbf{u}_{x_2})\\
&= \sum_{y,\mathbf{c},\mathbf{x},\mathbf{u}_{x_2}} yP_\mathbf{x}(y,\mathbf{c})Q(\mathbf{u}_{x_2}|c_1,\mathbf{u}^*_{x_1}(c_1))\pi'(x_1|c_1)Q(x_2|\mathbf{c},\mathbf{u}_{x_2})\\
&\leq \sum_{y,\mathbf{c},\mathbf{x}} yP_\mathbf{x}(y,\mathbf{c})\pi'(x_1|c_1)Q(x_2|\mathbf{c},\mathbf{u}^*_{x_2}(c_1))\\
&= \sum_{y,\mathbf{c},\mathbf{x}} yP_\mathbf{x}(y,\mathbf{c})\pi'(x_1|c_1)\pi'(x_2|\mathbf{c})
\end{aligned}$$

A variant of the setting used in Figs. 7a to 7c where non-contextualizable variable W affects both X_1 and C_1 with the bidirected edge in X_1 and C_1 removed from the setting. In such case, $W \perp\!\!\!\perp \mathbf{U}_\mathbf{X}$ can be exploited to show that X_2 does not need to observe C_1. The implication of this example is that when we have to deal with a more setting where contextualizable variables are defined on a per-action basis.

$$\mu_{\mathcal{S}''} = \sum_{y,\mathbf{c},\mathbf{x},w,\mathbf{u}_\mathbf{x}} yP_\mathbf{x}(y|\mathbf{c})Q(c_1)Q(w|c_1)Q(\mathbf{u}_\mathbf{x}|c_1,w)Q(x_1|c_1,\mathbf{u}_\mathbf{x},w)Q(c_2|x_1,c_1,\mathbf{u}_\mathbf{x},w)Q(x_2|x_1,\mathbf{c},\mathbf{u}_\mathbf{x},w)$$



(a) $\mathcal{G}_{\mathcal{S}}$ (b) $\mathcal{G}_{\mathcal{S}'}$ (c) $\mathcal{G}_{\mathcal{S}''}$

Figure 17: A variant of the setting used in Figs. 7a to 7c where non-contextualizable variable W affects both X_1 and C_1 with the bidirected edge in X_1 and C_1 removed from the setting.

$$
\begin{aligned}
&= \sum_{y,\mathbf{c},\mathbf{x},w,\mathbf{u}_\mathbf{x}} y P_\mathbf{x}(y|\mathbf{c}) Q(c_1) Q(w|c_1) Q(\mathbf{u}_\mathbf{x}) Q(x_1|c_1,\mathbf{u}_{x_1},w) Q(c_2|x_1,c_1) Q(x_2|c_2,\mathbf{u}_{x_2}) \\
&= \sum_{y,\mathbf{c},\mathbf{x},w,\mathbf{u}_\mathbf{x}} y P_\mathbf{x}(y,\mathbf{c}) Q(w|c_1) Q(\mathbf{u}_\mathbf{x}) Q(x_1|c_1,\mathbf{u}_{x_1},w) Q(x_2|c_2,\mathbf{u}_{x_2}) \\
&\leq \sum_{y,\mathbf{c},\mathbf{x},\mathbf{u}_\mathbf{x}} y P_\mathbf{x}(y,\mathbf{c}) Q(\mathbf{u}_\mathbf{x}) Q(x_1|c_1,\mathbf{u}_{x_1},w^*(c_1)) Q(x_2|c_2,\mathbf{u}_{x_2}) \\
&\leq \sum_{y,\mathbf{c},\mathbf{x}} y P_\mathbf{x}(y,\mathbf{c}) Q(x_1|c_1,\mathbf{u}^*_{x_1},w^*(c_1)) Q(x_2|c_2,\mathbf{u}^*_{x_2}) \\
&\leq \sum_{y,\mathbf{c},\mathbf{x}} y P_\mathbf{x}(y,\mathbf{c}) \pi'(x_1|c_1) \pi'(x_2|c_2)
\end{aligned}
$$

Proposition 5. *Given $\langle \mathcal{G}, Y, \mathbf{X}^\star, \mathbf{C}^\star \rangle$, let $\mathbf{X}' \doteq \{X \in \mathbf{X}^\star \mid de(X)_\mathcal{G} \cap (\mathbf{C}^\star \cup \{Y\}) \neq \emptyset\}$, $\mathbf{X}'' \doteq de(Y)_{\mathcal{G}_{\underline{\mathbf{X}'}}} \cap \mathbf{X}'$, and $\mathbf{Z} \doteq de(Y)_{\mathcal{G}_{\overline{\mathbf{X}''}}}$. The POMPSes for $\langle \mathcal{G}, Y, \mathbf{X}^\star, \mathbf{C}^\star \rangle$ are the same as those for $\langle \mathcal{G} \backslash \mathbf{Z}, Y, \mathbf{X}', \mathbf{C}^\star \backslash \mathbf{Z} \backslash \mathbf{X}'' \rangle$.*

Proof. Consider a POMPS $\mathcal{S}$ for $\langle \mathcal{G}, Y, \mathbf{X}^\star, \mathbf{C}^\star \rangle$. Let $\mathcal{H} = \mathcal{G}_\mathcal{S}$. By definition, $\mathbf{X}(\mathcal{S}) \subseteq an(Y)_\mathcal{H}$ and, hence, $\mathbf{C}_X \subset an(Y)_\mathcal{H}$ for every $X \in \mathbf{X}(\mathcal{S})$.

(1. a step to $\langle \mathcal{G}, Y, \mathbf{X}', \mathbf{C}^\star \rangle$): First, we justify the reduction of $\mathbf{X}^\star$. Consider $X \in \mathbf{X}^\star$ having no $\mathbf{C}^\star$ or Y as descendants (exclusive) in $\mathcal{G}$. It is impossible for X to become an ancestor of Y since none of its descendant can become a context for other $X' \in an(Y)_\mathcal{H}$. Further, due to the action-context separation (Prop. 2) X as an action cannot be simultaneously a context for other X'. Thus, we can exclude actionable variables that is not an ancestor of other contextualizable variables or Y. Hence, $\mathbf{X}' = \{X \in \mathbf{X}^\star \mid de(X)_\mathcal{G} \cap (\mathbf{C}^\star \cup \{Y\}) \neq \emptyset\}$ is the subset of $\mathbf{X}^\star$ that can become an action in POMPS.

(2. a step to $\langle \mathcal{G}, Y, \mathbf{X}', \mathbf{C}^\star \backslash \mathbf{Z} \rangle$): Next, we explain the reduction of $\mathbf{C}^\star$ focusing on removing $\mathbf{Z}$. Any $C \in \mathbf{C}^\star$ that is a descendant of Y without any $\mathbf{X}'$ present in the directed path from Y to C cannot become $an(Y)_\mathcal{H}$ due to an induced cycle caused by $C \rightarrow X$, $X \in an(Y)_\mathcal{H}$, and $Y \in an(C)_\mathcal{H}$. Hence, contextualizable variables that cannot change their descendant relationships with Y, i.e., $de(Y)_{\mathcal{G}_{\overline{\mathbf{X}'}}} \cap \mathbf{C}^\star$ are not contextualizable. With $\mathbf{X}'' = de(Y)_{\mathcal{G}_{\underline{\mathbf{X}'}}} \cap \mathbf{X}'$ (action variables that are descendants of Y with no other action variables in between), we can elicit $de(Y)_{\mathcal{G}_{\overline{\mathbf{X}'}}} \cap \mathbf{C}^\star = de(Y)_{\mathcal{G}_{\overline{\mathbf{X}''}}} \cap \mathbf{C}^\star$. Hence, $\mathbf{C}^\star$ can be reduced to $\mathbf{C}^\star \backslash \mathbf{Z}$ where $\mathbf{Z} = de(Y)_{\mathcal{G}_{\overline{\mathbf{X}''}}}$.

(3. a step to $\langle \mathcal{G} \backslash \mathbf{Z}, Y, \mathbf{X}', \mathbf{C}^\star \backslash \mathbf{X}'' \backslash \mathbf{Z} \rangle$): Finally, we examine eliminating $\mathbf{Z}$ from the graph. Since $\mathbf{Z}$ can only be descendant of Y, the expected reward of any valid MPS in the setting is free from the influence of the mechanism changes for $\mathbf{Z}$. Hence, the removal of $\mathbf{Z}$ does not affect the reward landscape among the MPSes. However, once if $\mathbf{Z}$ is removed from the setting, then $\mathbf{X}''$ is no longer descendant of Y in the resulting setting, and the constraint that $de(\mathbf{X}'')_\mathcal{G} \backslash \mathbf{X}''$ cannot become a context without excluding them from descendants of Y will be lifted. That is, the new setting induces more MPSes which is not valid in the original setting. We claim that removing $\mathbf{X}''$ from the contextualizable variables is sufficient. Any MPSes valid in the new setting but invalid in the original setting are those without a proper intervention on a subset of $\mathbf{X}'$ that would otherwise properly break the cyclicity (the reason for invalidity). By the way, without intervening them, they cannot become POMPSes in the new setting since, simply, intervening on $\mathbf{X}''$ leads to a better policy scope, thus, they are indeed non-POMPSes. As a consequence, any POMPS in the new setting satisfies the constraint and is a valid POMPS in the original setting (vice versa). $\square$