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].
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, \textit{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, \textit{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 \textit{causal graph} $G$ (Fig. 1a), where $C$, $X = \{X_1, X_2\}$, $Y$ 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 $\mu_\pi = E_\pi[Y]$ under a mixed policy (or simply, policy) $\pi \in \Pi$, where $\Pi$ is a mixed policy space. A mixed policy is associated with its \textit{scope}, called \textit{mixed policy scope} (MPS), which specifies the variables the policy is intervening, and the variables taken into account for each intervened variables.

A standard contextual bandit (CB) optimizes a policy $\pi_{CB}$ (Fig. 1b), a (stochastic) mapping from contexts to actions, which can be equally represented as a pair of decision rules $\pi_{CB} = \{\pi(x_1|c), \pi(x_2|x_1,c)\}$ (Fig. 1c). Traditionally, the policy is optimized within a restricted space $\Pi_{CB}$, characterized by policies following a scope $S_{CB} = \{\langle X_1, \{C\} \rangle, \langle X_2, \{X_1, C\} \rangle\}$ that $X_1$ is determined by $C$ and $X_2$ is decided based on $C$ and $X_1$. Unfortunately, the optimal policy $\pi_{CB}^* = \arg\max_{\pi \in \Pi_{CB}} \mu_\pi$ can be suboptimal, i.e., $\mu_{S_{CB}}^* \ngeq \mu^*$ where $\mu^*$ is the optimal expected reward. To ground 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 $\epsilon$ be a noise over $X_1$ following $P(\epsilon = 1) = 0.2$. Also, let the unobserved causal mechanisms be specified as $X_1 \leftarrow U_1 \circ \epsilon, C \leftarrow U_1, X_2 \leftarrow U_2 \circ X_1 \circ C$, and $Y \leftarrow (1 - (X_2 \circ U_2)) \circ C$, where $\circ$ 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 $\mu_{S_{CB}}^* = 0.75$. 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 $\epsilon$ 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: \textit{non-redundancy} and \textit{optimality}. We explain these desiderata through an illustration (Fig. 2) of the four MPSes $S_a = \{\}$, $S_{cb}$, $S_d = \{\langle X_1, \{C\} \rangle \}$, and $S_e = \{\langle X_2, \{C\} \rangle \}$. 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 $\geq$ whether one’s optimal reward is at least as good as (or equal to) the other’s in \textit{every} world compatible with a causal graph. The equality forms an equivalence class among scopes with respect to optimal rewards. Roughly speaking (to be formalized later on), \textit{non-redundancy} refers to the condition of a scope

---

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). $\pi$ nodes are intervention indicators, which will be left implicit throughout the paper.

Figure 2: Relationships between the MPSes
such that removing any of its actions or contexts can negatively affect its maximum performance. In other words, given two scopes \( S \) and \( S' \), if \( S \supseteq S' \) and \( S = \ast S' \), then \( S \) is said to be redundant. For instance, since \( S_c \supseteq S_e \) while \( S_c = \ast S_e \), 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 \( X_2 \) into account for \( X_2 \) (a redundant context). Furthermore, optimality of a scope \( S \) represents that there exists no other scope \( S' \) (not in the equivalence class of \( S \)) such that \( S' \supseteq S \). For example, \( S_d \), when optimized, is at least as good as \( S_a \) (i.e., \( \mu_{S_d} \geq \mu_{S_a} \)) in every environment, and can outperform it in some environments (i.e., \( \mu_{S_d} > \mu_{S_a} \)), which demonstrates that \( S_a \) does not meet the optimality criterion. Not every pair of scopes can be comparable: \( S_c \) is not comparable to \( S_n \) or \( S_i \). After a careful examination, we can indeed show that MPSes \( S_c, S_d, S_e \) meet the optimality criterion. Both non-redundancy and optimality are satisfied only by \( S_d \) and \( S_e \) 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 \( X \) with its value \( x \). We follow notational conventions from literature on measure theory, algebra of sets, and causal inference. We may use \( \cup \), instead of \( \bigcup \), 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 \( M \) is a quadruple \( \langle U, V, P(U), F \rangle \), where \( U \) is a set of exogenous variables determined by factors outside the model following a joint distribution \( P(U) \), and \( V \) is a set of endogenous variables whose values are determined following a collection of functions \( F = \{ f_i \}_{i \in V} \) such that \( V_i \leftarrow f_i(Pa_i, U_i) \) where \( Pa_i \subseteq V \setminus \{ V_i \} \) and \( U_i \subseteq U \). The observational distribution \( P(v) \) is defined as \( \sum_u \prod_{v_i \in V} P(v_i | pa_i, u_i) P(u) \). Further, \( do(X = x) \) represents the operation of fixing a set \( X \) to a constant \( x \) regardless of their original mechanisms.

Such intervention induces a submodel \( M_X \), which is \( M \) with \( f_X \) replaced to \( x \) for \( X \in X \). Then, an interventional distribution \( P_X(v|x) \) (or also \( P(v|x|do(x)) \)) follows from \( M_X \), and is such that \( P_X(v|x) = \sum_u \prod_{v_i \in V \setminus X} P(v_i | pa_i, u_i) P(u) \).

Graphically, each SCM (model, for short) is associated with a causal diagram \( G = \langle V, E \rangle \), 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 \( W \subseteq V \setminus \{ X \} \) such that \( U_W \perp \! \! \! \perp U_X \) and \( U_Y \not\perp \! \! \! \perp U_W \); From the agent’s perspective, only the causal graph \( G \) of the environment \( M \) is available, while its reward is validated through \( M \). We operate in the non-parametric setting, where no assumption about the form or shape of the pair \( \langle P(U), F \rangle \) is made, but for the structural knowledge encoded in \( G \). Whenever not even \( G \) is known, the agent can perform active interventions to learn it; for example, see [28, 29]. We denote by \( \hat{G}_X \) an edge subgraph of \( G \) which removes edges incoming to \( X \) and outgoing from \( Z \). A submodel \( M_X \) can be presented as \( \hat{G}_X \) with \( X \) fixed to \( x \). Hence, causal relationships among other variables are captured in \( G \setminus X \), which is the subgraph of \( G \) over \( V \setminus X \). We denote by \( G(V') \) the latent projection of \( G \) onto \( V' \), the causal graph retaining causal relationships among \( V' \) [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 A). Also, the omitted proofs and derivations are provided in Appendix.
2 Mixed Policies: Fundamentals & Basic Results

As discussed in the previous section, a causally 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 $G$ be a causal graph, $Y$ be a specific reward variable, $X^* \subseteq V \setminus \{Y\}$ a set of intervenable variables, and $C^* \subseteq V \setminus \{Y\}$ a set of contextualizable variables. A mixed policy scope $S$ is defined as a collection of pairs $(X, C_X)$ such that (i) $X \in X^*$, $C_X \subseteq C^* \setminus \{X\}$, and (ii) $G_S$ is acyclic, where $G_S$ is defined as $G$ with edges onto $X$ removed and directed edges from $C_X$ to $X$ added for every $(X, C_X) \in S$.

For concreteness, given a causal graph $G$ (Fig. 1a), the observational case is an MPS $\{\}$. An MPS $S_{cb} = \{(X_1, \{C\}), (X_2, \{X_1, C\})\}$ induces a graph (Fig. 1b) while $\{(X_2, \{C\})\}$ induces a graph in Fig. 1c. An MPS represents a class of mixed policies that share the same graphical characteristics manifested by $G_S$, an induced graph for $M_\pi$.

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

If we consider the MPS $S_{cb}$ discussed above, its mixed policy $\pi$ 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 $M_\pi$, a mixed policy $\pi$ induces a variant of SCM $M_\pi$ where the function for $X \in X(\pi)$ is replaced by the corresponding $\pi_{X \setminus C_X}$ (see [34] for a detailed account). We denote by $P_\pi$ the joint distribution over the variables from the system under the policy $\pi$. Throughout the paper, $G$, $Y$, $X^*$, and $C^*$ are oftentimes implicit including an underlying SCM $M \sim G$ and, thus, $\Pi$, as well.

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

$$\mu_\pi = \sum_{y, x, c} y P_x(y, c^-) \prod_{X \in X(\pi)} \pi(x | c_x).$$

(1)

The expression separates the atomic interventional probability (first factor), which is inherent to the underlying world and not affected by the policy $\pi$, from the likelihood of a specific intervention given contexts (second factor), which is optimizable and defined by $\pi$. The expected reward can also be written focusing only on a subset of intervened variables. Given $X' \subseteq X(\pi)$, let $C' = \bigcup_{X \in X'} C_X \setminus X'$, and $Q' = P_{\pi \setminus X'}$, where $\pi \setminus X'$ represents $\pi$ with decision rules over $X'$ removed. Then, $\mu_\pi = \sum_{y, x', c'} y Q'_x(y, c') \prod_{X \in X'} \pi(x | c_x)$. 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 $\pi$ is said to be optimal in the given environment if and only if $\mu_\pi \equiv \max_{\pi' \in \Pi} \mu_{\pi'}$. Restricting our attention to $\Pi_S \equiv \{\pi \in \Pi | \pi \sim S\}$, we define $\mu_S^* \equiv \max_{\pi' \in \Pi_S} \mu_{\pi'}$, an optimal policy $\pi$ with respect to $S$. We call a mixed policy deterministic if, for every $\pi_{X \setminus C_X} \in \pi$, $X$ is determined by a function of $C_X$.

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 $S$, there always exists a deterministic mixed policy $\pi \in \Pi$ such that $X(\pi)$ and $C(\pi)$ are disjoint and $\mu_\pi = \mu_S^*$. 

4
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. 3e $X_2$ utilizes both $X_1$ and $C$. However, it is sufficient to make use of only $C$. By noting that the policy relative to Fig. 3f can achieve optimality, while relying on lesser information than the one relative to Fig. 3e, we investigate how to capture non-redundancy within MPSes.

### 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 $S$ subsumes $S'$, denoted by $S' \subseteq S$, if $X(S') \subseteq X(S)$ and $C_{S'} \subseteq C_S$, for every $(X, C_X) \in S'$. Further, we denote by $\pi' \subseteq \pi$, where $\pi' \sim S'$ and $\pi \sim S$ if $\pi'(x|e_x) = \sum_{e'_x} \pi(x|e_x)e_xP_s(x|e'_x)$, for every $X \in X(S')$ where $C'_{X} = C_X \setminus C'_{S'}$.

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

The constraint on $\pi'$ 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—$\pi'(x|e'_x) = Q(x|e'_x)$ if $C_X \neq C_{S'}$ and $Q(x|e_x) = \pi(x|e_x)$, otherwise. Hence, the constraint provides a basis to characterize non-redundancy of MPS utilizing well-established graphical criteria.

**Theorem 1.** Let $S = \{X, C_X\}_{X \in X}$ be an MPS and let $H = G_S$. $S$ is non-redundant if and only if (i) $X \subseteq an(Y)_H$ and (ii) $(C \notin Y | C_X \setminus \{C\})$ in $H \setminus \{X\}$, for every $X \in X$ and $C \in C_X$.

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 $Y$. The condition (ii) coincides with rule 2 of do-calculus $Q(y|x, e_y, \{c\}) = Q_x(y|x, e_x, \{c\})$, where $Q = P_x$. In words, the path from $C$ to $Y$ can be concatenated with $X \rightarrow C$ to form a back-door path from $X$ to $Y$. 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.

### 3.1 Non-Redundancy under Optimality

Non-redundancy of MPS (Def. 3) based on a stringent constraint imposed on $\pi'$ is insufficient to understand, e.g., whether a context of an action would be still relevant even when $\pi \sim S$ 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. 3e where $X_2$ listens to $X_1$ as context. We showed that the dependence is vanished under the optimality (Fig. 3f). That is, the agent would better avoid learning $\pi(x_2|e_x, x_1)$ at the beginning, but optimize $\pi(x_2|e_x)$ instead.

**Definition 4 (Non-Redundancy under Optimality (NRO)).** Given $\langle G, Y, X^*, C^* \rangle$, an MPS $S$ is said to be non-redundant under optimality if there exists an SCM $M$ compatible with $G$ such that $\mu_S < \mu_{S'}$ for every strictly subsumed MPS $S' \subseteq S$, i.e., $\exists M \sim G S' \subseteq S (\mu_S < \mu_{S'})$.

---

1 This condition was leveraged in the atomic interventions case to establish minimalism [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., $|X(S)| = 1$), see [41, 42].
We present a general criterion to test such redundancies by first proposing a lemma to obtain an intermediate expression. Let \( \mathbf{X} \) denote a set of actions, \( \mathbf{S} \) a set of contexts, and \( \mathbf{C} \) the causal graph. The following proposition characterizes the optimal reward under determinism.

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

\[
\mu^\ast_{\mathbf{S}} = \sum_{x, c'} \mathbf{Q}(x|\mathbf{x}', c') \prod_{X \in \mathbf{X}'} \pi(x|\mathbf{x}' \cup c') \cap c_x,
\]

then, \( \mathbf{C}_{\mathbf{X}} \setminus (\mathbf{C} \cup \mathbf{X}') \) are jointly redundant to \( \mathbf{X}' \) under optimality, and \( \mathbf{S}' \subseteq (\mathbf{S} \setminus \mathbf{X}') \cup \{ (\mathbf{X}', \mathbf{C} \cap \mathbf{C}_{\mathbf{X}}) \} \) satisfies \( \mu^\ast_{\mathbf{S}} = \mu^\ast_{\mathbf{S}'} \).

**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 characterize 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 conditional variables, in which we explicitly represent with an operation \([\cdot]\) for clarity.

Given conditionals \( \mathbf{Z} \), the implied variables with respect to \( \mathbf{Z} \) is computed as follows. Initially setting \( [\mathbf{Z}] \leftarrow \mathbf{Z} \), we update \( [\mathbf{Z}] \leftarrow [\mathbf{Z}] \cup \{ X \in \mathbf{S} | \mathbf{C}_{\mathbf{X}} \subseteq [\mathbf{Z}] \} \) until it is converged. Then, given \( \pi \sim S \), an optimal policy with respect to \( \mathbf{S} \), a conditional independence statement \( \mathbf{W} \perp \perp \mathbf{T} | [\mathbf{Z}] \) for \( \mathbf{P}_{\pi} \) becomes \( \mathbf{W} \perp \perp \mathbf{T} | \mathbf{Z} \) in \( \mathcal{G}_{\pi} \). Consider \( C \subseteq \mathbf{C}_{\mathbf{X}} \) for some \( X \in \mathbf{X}(\pi) \). The redundancy of a single context can now be expressed as \( \mathbf{C} \perp \perp Y | [\mathbf{C}_{\mathbf{X}} \setminus \{C\}] \) for \( X \in \mathbf{X}(\pi) \). In Fig. 5a, \( C_2 \) as a context of \( X_2 \) is independent of \( Y \) given \( C_1 \) in a graph with \( X_2 \) removed since \( \{C_1\} = \{C_1, X_1\} \) and \( C_2 \leftarrow X_1 \rightarrow Y \) is not a valid trail anymore. Hence, \( C_2 \) is removable from \( \mathbf{C}_{\mathbf{X}_2} \).

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 \( \mu_\pi = \mathbb{E}_{c_3 \in \mathbf{X}_3} \mathbb{E}_\pi[y|c_3] \), we can derive that \( \mu_\pi \leq \mathbb{E}_\pi[y|c_3] \). Given that \( c_3 \) is merely a constant, new decision rules \( \pi'(x_i|c_1) = \mathbb{Q}_i(x_i|c_1, c_3) = \pi(x_i|c_1, c_3) \) for \( i = 1, 2 \) 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

\[
\mu_\pi = \sum_{c_1, c_2} \mathbb{Q}(c_2|c_1) \left( \sum_{y, x} x \mathbb{P}_x(y, c_1) \pi(x_1|c_1, c_2) \pi(x_2|c_2, c_2) \right) = \sum_{c_1, c_2} \mathbb{Q}(c_2|c_1) \mu_\pi(c_1, c_2).
\]

Let \( c_2 \) be a function taking \( c_1 \) such that \( c_2(c_1) = \arg \max_{c_2} \mu_\pi(c_1, c_2) \) for \( c_1 \in \mathbf{X}_1 \). Then,

\[
\leq \sum_{c_1} \mu_\pi(c_1, c_2(c_1)) = \sum_{y, x, c} y \mathbb{P}_x(y, c_1) \pi(x_1|c_1, c_2(c_1)) \pi(x_2|c_2, c_2(c_1)).
\]

By incorporating \( c_2(c_1) \) into \( \pi \), we can introduce \( \pi' \) 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., \( \pi(x_1|c_1, c_2) = \pi'(x_1|c_2) \).

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 \( V_{<V} \) denote a subset of \( V \) 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 $X' \subseteq X(S)$, actions of interest, $C' \subseteq C_{X' \setminus X'}$, non-action contexts of interest. If there exists a subset of exogenous variables $U'$ in $G_S$, a subset of endogenous variables $Z$ in $G_S$ that disjoint with $C' \cup X'$ and subsumes $C_{X' \setminus (C' \cup X')}$, and an order $\prec$ over $V' \cup C' \cup X' \cup Z$, then, the expected reward for $\pi$, a deterministic policy optimal with respect to $S$, can be written as

$$
\mu_\pi = \sum_{y, y', y' \prec} \sum_{x, z} Q(y', x, y) \prod_{z} Q(x) \prod_{x \in X'} \pi(x|e_x). \tag{3}
$$

Lemma 1 offers a sufficient condition for obtaining the intermediate expression (Eq. 3) for us to rewrite $\mu_\pi$ 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 $Q(y', x, y)$ from the rest. The third one is to obtain $\pi(x|e_x)$ from $Q(x|y' \prec z, u')$. We revisit Fig. 4 where we will ultimately show that, indeed $C_2$ and $C_3$ are redundant contexts under optimality. Given $C' = \{c_1\}$ and $X' = \{x_1, x_2\}$, consider $Z = \{c_2, c_3\}$, $U' = \emptyset$, and order $\prec = (c_3, c_1, x_2, c_2, x_1)$. We can derive the following expression for the expected reward (with subscripts concatenated),

$$
\mu_S = \sum_{y, y', y' \prec} \sum_{x, z} Q(y', x, y) \prod_{z} Q(x) \prod_{x \in X'} \pi(x|e_x). \tag{4}
$$

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 $U'$, $Z$, and $\prec$ satisfy Lemma 1. For $Z \in \mathfrak{z}$, let $V_Z$ be a minimal subset of $V' \prec Z \cup U'$ such that $Q(Z|V_Z) = Q(Z|V' \prec Z, U')$. We define $\hat{X}_{\mathfrak{z}}$ with respect to $(Z, V_Z)$, which is, with $\hat{X} = \hat{X}_{\mathfrak{z}} \cup \{Z \in \mathfrak{z} : V_Z \setminus U' \subseteq \hat{X}\}$, fixed($\hat{X}$) is $\mathfrak{z}$ if $\mathfrak{z}$ is $\mathfrak{z}$ and fixed($\hat{X}$), otherwise. If fixed($X_{\mathfrak{z}}|Z) \supseteq \mathfrak{z}$ for $X \in X_{\mathfrak{z}}$, then $\mathfrak{z} = (\mathfrak{s} \setminus \mathfrak{x}) \cup \{X_{\mathfrak{z}}, Z\} \mathfrak{x}$ satisfies $\mu_S = \mu_\mathfrak{s}$.

Thm. 2 provides a condition where Eq. (3) can be transformed to $\mu_S$. To do so, it examines whether terms $Q(x|y' \prec z, u')$ can be removed by fixing $Z$ to $z^*$ conditional on $v _ z$ in connection with the context to be removed. That is,

$$
\mu_S = \sum_{y, y', y' \prec} \sum_{x, z} Q(y', x, y) \prod_{z} Q(x) \prod_{x \in X'} \pi(x|e_x). \tag{7}
$$

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

$$
\leq \sum_{y, y', y' \prec} \sum_{x, z} Q(y', x, y) \prod_{z} Q(x) \prod_{x \in X'} \pi(x|e_x). \tag{8}
$$

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

$$
\leq \sum_{y, y', y' \prec} \sum_{x, z} Q(y', x, y) \prod_{z} Q(x) \prod_{x \in X'} \pi(x|e_x). \tag{9}
$$

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

$$
\leq \sum_{y, y', y' \prec} \sum_{x, z} Q(y', x, y) \prod_{z} Q(x) \prod_{x \in X'} \pi(x|e_x). \tag{10}
$$

These steps correspond to checking fixed($C_{X_1}|Z) = fixed(\{C_1\}) = \{C_1, C_2, X_2, C_2\} \supseteq \{C_2, C_3\}$ for $X_1$. Since $\mu_{S'} \leq \mu_\mathfrak{s}$, by the existence of $\pi \in S$ that can emulate $\pi' \in S'$, and $\mu_{S'} \geq \mu_\mathfrak{s}$ by the derivation (Eq. 12), we can conclude that $\mu_{S'} = \mu_\mathfrak{s}$. As a consequence, MPS $\mathcal{S}$ is not non-redundant under optimality due to the ineffective contexts $\{C_2, C_3\}$ with respect to $\{X_1, X_2\}$.
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 $\mu \leq \mu_{\mathcal{G}_1}$ since there exists an optimal $X_1$ value, $x_1^*$. Next, $\mu_{\mathcal{G}_2} \leq \mu_{\mathcal{G}_1}$, there exists an optimal $X_2$ value, and can be determined without conditional on $X_1$, which is implied. Finally, $\mu_{\mathcal{G}_3} \leq \mu_{\mathcal{G}_1}$ since $X_1$ can better behave by taking an effective context $C$ into account. Therefore, the agent can only optimize parameters involving $\mathcal{G}_3$ (Fig. [6a]) to obtain 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 $(\mathcal{G}, X^*, C^*, Y)$, let $\mathbb{S}$ be a set of NRO MPSes. An MPS $S \in \mathbb{S}$ is said to be possibly-optimal if there exists $M \sim \mathcal{G}$ such that $\mu_S > \max_{S' \in \mathbb{S} \setminus \{S\}} \mu_{S'}$.

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 $S$ and $X \in X(S)$, adding $C \in C^* \setminus \{X\}$ as a context of $X$, resulting $S' = (S \setminus \{X\}) \cup \{(X, C_X \cup \{C\})\}$ improves $S$ if $C \notin \text{de}(X)_{\mathcal{G}_S}$ and $C \perp \perp Y \mid X_{\mathcal{H}_S}$ in $H \setminus \{X\}$.

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 $X \in X^*$ with an artificial one $\pi(x|z)$. 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 $Z$ carry information tantamount to the original one. If every parent of $X \in X^*$ is contextualizable (e.g., no UC), the problem becomes trivial (e.g., Markovian). Otherwise, we examine the existence of a back-door path. Let $P = P_x$ and $H = H_\pi$ for some $S \sim X$. Given $X \in X^* \setminus X(\pi)$ and $Z \subseteq C^* \setminus \{X\}$, if (i) $Y \perp \perp X \mid \{Z\}$ in $H_\pi$ and (ii) $X \notin \text{an}(Z)_{H_\pi}$, then

$$\mu_{\pi} = \sum_{y} yQ(y|x,z)Q(x|z)\pi(z) = \sum_{y,z} yQ'_y(y|z)Q(x|z)\pi(z) \leq \sum_{y,z} yQ'_y(y|z)Q(x|z)\pi'(z) \leq \mu_{\pi'}$$

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

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

8
Given an MPS $S$, let $S' \neq S$ be an MPS with $X(S) \subseteq X(S')$ such that $\mathcal{H}'$ the union of induced graphs $G_S \cup G_{S'}$ is acyclic. Let $X'$ be actions that the MPSes disagree on, i.e., $(X(S') \setminus X(S)) \cup \{X \in X(S) | C'_X \neq C_X \}$, and (invalid) MPS $S'' = \{ \langle X, pu(X)_{\mathcal{H}'} \cup U_X \rangle \}_{X \in X'}$. $\mu_{S''} = \mu_{S'}$ can be elicited by Thm. 2, with $U' = \emptyset$ and $\prec = \langle C_1, U_{X_1}, U_{X_2}, X_1, C_2, X_2 \rangle$ to demonstrate $\mu_{S''} = \mu_{S'}$. Since $\mu_{S'} \leq \mu_{S''}$, we can elicit $\mu_S \leq \mu_{S'}$, confirming that $S$ is not a POMPS. By allowing $X'$ to intersect with $X(S)$, 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. 7c) without inducing a cycle. This implies that contexts in the descendants of the reward becomes usable if interventions can break the ancestral relationships. Second, $X \in X^*$ that cannot affect $C^*$ or $Y$ is not intervenable — if $de(X) \cap (C^* \cup \{Y\}) = \emptyset$, there exists no MPS that makes $X \in an(C^* \cup \{Y\})$, and, thus, $X$ can be excluded from $X^*$.

**Proposition 5.** Given $(G, Y, X^*, C^*)$, let $X' = \{X \in X^* | de(X) \subseteq (C^* \cup \{Y\}) \neq \emptyset \}$, $X'' = de(Y) \cup X'$, and $Z = de(Y) \cup X''$. The POMPSes for $(G, Y, X^*, C^*)$ are the same as those for $(G, Z, X', C^* \setminus Z \setminus X'')$.

**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


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 \(X^* = \{X_1, X_2\}\) and \(\mathcal{C}^* = \{C, X_1\}\), 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. 2. 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:

(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 \(\mathcal{S}_\alpha \sim \mathcal{S}_\beta\) represents the equivalence in their maximum achievable expected rewards, (i.e., \(\mu^*_{\mathcal{S}_\alpha} = \mu^*_{\mathcal{S}_\beta}\)); 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 \(\mu\), the expected reward for the observational policy, and \(\mu^*_{\mathcal{S}(X_1,\{C\})}\), the maximum expected reward with the policy intervening on \(X_1\) given \(C\). We will show below \(\mu \leq \mu^*_{\mathcal{S}(X_1|c)}\) by using do-calculus,

\[
\mu = \sum_{y} y P(y) \quad \text{by definition (13)}
\]

\[
= \sum_{y,x_1} y P(y, x_1, c) \quad \text{marginal probability (14)}
\]

\[
= \sum_{y,x_1} y P(y|x_1, c) P(x_1|c) P(c) \quad \text{chain rule (15)}
\]

\[
= \sum_{y,x_1} y P_{x_1}(y|c) P(x_1|c) P_{x_1}(c) \quad \text{Rule 2 of do-calculus (16)}
\]

\[
= \sum_{y,x_1} y P_{x_1}(y, c) P(x_1|c) \quad \text{chain rule (17)}
\]

\[
= \sum_{y,x_1} y P_{x_1}(y, c) \pi(x_1|c) \quad \text{by definition (18)}
\]

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 \(\mathcal{S} = \{(X_1, \{C\}), (X_2, \{X_1\})\}\) and \(\mathcal{S}' = \{(X_2, \{C\})\}\) through the following derivation,

\[
\mu^*_{\mathcal{S}} = \sum_{y,x,c} y P_{x}(y, c) \pi(x_1|c) \pi(x_2|x_1) \quad \text{by Eq. (1)} (19)
\]

\[
= \sum_{y,x,c} y P_{x_2}(y, c) \pi(x_1|c) \pi(x_2|x_1) \quad \text{Rule 3 of do-calculus (20)}
\]

\[
= \sum_{y,x_2,c} y P_{x_2}(y, c) \sum_{x_1} \pi(x_1|c) \pi(x_2|x_1) \quad \text{algebra (21)}
\]

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

\[
= \sum_{y,x_2,c} y P_{x_2}(y, c) \sum_{x_1} \pi(x_1|c) \pi'(x_2|x_1, c) \quad \text{by construction (22)}
\]
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,

$$
\leq \sum_{y,x_2,c} y P_{x_2}(y,c) \pi'(x_2|x_1, c)
$$

by definition

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

$$
= \sum_{y,x_2,c} y P_{x_2}(y,c) \pi(x_2|c)
$$

by construction

$$
\leq \mu_{S_\delta}
$$

Therefore, $\mu_{\{X_1, \{C\}\}, \{X_2, \{C, X_1\}\}} \leq \mu_{\{X_2, \{C\}\}}$. 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, $\{\{X_1, \{C\}\}\}$ (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.,

$$
\{\{X_2, \{C\}\}, \{X_1, \emptyset, X_2, \{C\}\}, \{X_1, \{C\}, X_2, \{C\}\}, \{X_2, \{C, X_1\}\}, \{X_1, \emptyset, X_2, \{C, X_1\}\}, \{X_1, \{C\}, X_2, \{C, X_1\}\}\}
$$

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), $\{\{X_2, \{C\}\}\}$ is subsumed by other
We empirically validate that the use of refined policies leads to a better performance measured. In particular, we demonstrate that the CB approach cannot achieve the optimal reward in a certain environment, 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., \( \{X_2, \{C\}\} \) and \( \{\emptyset, \{C\}\} \). Putting this information together, we can see in Fig. 10 that only \( \{X_1, \{C\}\} \) and \( \{X_2, \{C\}\} \) 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 \( \{X_1, \{C\}\} \) and \( \{X_2, \{X_1, C\}\} \). 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., \( \{X_1, \{C\}\} \) and \( \{X_2, \{C\}\} \). 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 \( X_1 \).
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 \( X, Y \) are said to be d-separated by a another set \( Z \) in a directed acyclic graph \( G \), denoted by \( (X \perp \perp Y \mid Z)_G \), if every path \( P \) from vertices in \( X \) to vertices in \( Y \) are blocked where blockage occurs when one of the following holds:
1. \( P \) contains at least one arrow-emitting node that is in \( Z \), or
2. \( P \) contains at least one collider that is outside \( Z \) and has no descendant in \( Z \).

**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):**
  \[
P_x(y|z, w) = P_x(y|w) \quad \text{if} \quad Z \perp Y \mid X, W \text{ in } \mathcal{G}_X
  \]

- **R2 (Action/observation exchange):**
  \[
P_x(y|z, w) = P_{x,x}(y|w) \quad \text{if} \quad Z \perp Y \mid X, W \text{ in } \mathcal{G}_{XZ}
  \]

- **R3 (Insertion/deletion of actions):**
  \[
P_x(y|w) = P_{x,x}(y|w) \quad \text{if} \quad Z \perp Y \mid X, W \text{ in } \mathcal{G}_{XZ(W)}
  \]

where \( Z(W) \) is a subset of \( Z \) that is not ancestor of \( W \) in \( \mathcal{G}_X \). For convenience of some of the proofs, let \( \mathcal{H} = \mathcal{G} \setminus X \) and \( Q = P_x \), and consider the following (rewritten) what rules:

- **R1 (Insertion/deletion of observations):**
  \[
  Q(y|z, w) = Q(y|w) \quad \text{if} \quad Z \perp Y \mid W \text{ in } \mathcal{H}
  \]

- **R2 (Action/observation exchange):**
  \[
  Q(y|z, w) = Q_{x,x}(y|w) \quad \text{if} \quad Z \perp Y \mid W \text{ in } \mathcal{H}_{XZ}
  \]

- **R3 (Insertion/deletion of actions):**
  \[
  Q(y|w) = Q_{x,x}(y|w) \quad \text{if} \quad Z \perp Y \mid W \text{ in } \mathcal{H}_{XZ(W)}
  \]

This representation 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 \( \mathcal{G}_x \), and let \( X \prec C \) be \( \{ X \in X \mid X \prec C \} \). We may write capital subscripts of a value as lowercase, e.g., \( c \) instead of \( C \) for a value for \( C_X \).

With \( X = X(\pi), C = C(\pi), \) and \( C^- = C \setminus X \), we start by writing the expected reward of \( \pi \),

\[
\mu_\pi \triangleq \mathbb{E}_\pi[Y] \quad \text{def} \quad (27)
\]

\[
= \sum_{y,x,c^-} y P_\pi(y, x, c^-) \quad \text{MP} \quad (28)
\]

\[
= \sum_{y,x,c^-} y P_\pi(y|x, c^-) P_\pi(x, c^-) \quad \text{CR} \quad (29)
\]

\[
= \sum_{y,x,c^-} y P(y|x, c^-) P_\pi(x, c^-) \quad \text{R1} \quad (30)
\]

\[
= \sum_{y,x,c^-} y P_x(y|c^-) P_\pi(x, c^-) \quad \text{R2} \quad (31)
\]

\[
= \sum_{y,x,c^-} y P_x(y|c^-) \prod_{x \in X} P_\pi(x|x_{\prec X}, c_{\prec x}) \prod_{c \in C^-} P_\pi(c|x_{\prec C}, c_{\prec c}) \quad \text{CR} \quad (32)
\]

\[
= \sum_{y,x,c^-} y P_x(y|c^-) \prod_{x \in X} P_\pi(x|x_c, c_{\prec c}) \prod_{c \in C^-} P_\pi(c|x_{\prec C}, c_{\prec c}) \quad \text{R1} \quad (33)
\]

\[
= \sum_{y,x,c^-} y P_x(y|c^-) \prod_{x \in X} \pi(x|x_c) \prod_{c \in C^-} P_\pi(c|x_{\prec C}, c_{\prec c}) \quad \text{def} \quad (34)
\]

\[
= \sum_{y,x,c^-} y P_x(y|c^-) \prod_{x \in X} \pi(x|x_c) \prod_{c \in C^-} P_\pi(c|x_{\prec C}, c_{\prec c}) \quad \text{R1} \quad (35)
\]

\[
= \sum_{y,x,c^-} y P_x(y|c^-) \prod_{x \in X} \pi(x|x_c) \prod_{c \in C^-} P_{x_{\prec C}}(c|x_{\prec c}) \quad \text{R2} \quad (36)
\]

\[
= \sum_{y,x,c^-} y P_x(y|c^-) \prod_{x \in X} \pi(x|x_c) \prod_{c \in C^-} P_x(c|x_{\prec c}) \quad \text{R3} \quad (37)
\]
where one attempts to optimize will allow for different sensing capabilities for agents. The results presented in this paper can be applied to the regime nodes (Eq. (30)) is identical to Rule 2 applied to X in Eq. (31). The derivation for a subset of 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 C. 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 $\pi \sim S$ given an MPS $S$. Let $X = X(\pi)$ and $C^- = C(\pi) \setminus X$. Given a topological order among $X$ defined over $G_S$ such that $X_i \prec X_j$ if $i < j$, let $Q' = P_{\pi \setminus X}$ where $P_{\pi \setminus X}$ denotes a policy $\pi$ with actions $X' \subseteq X$ removed. Then,

$$
\mu_{\pi} = \sum_{y, x_1, c_1} yQ'_{x_1}(y, c_1)\pi(x_1|c_1) = \sum_{x_1, c_1} \pi(x_1|c_1) \sum_y yQ'_{x_1}(y, c_1).
$$

If $\pi_{X_1|C_1}$ is not deterministic with respect to $c_1$ where $P_{\pi}(c_1) = Q'(c_1) > 0$, there must be at least two values of $x_1'$ and $x_1''$ such that

$$
\pi(x_1'|c_1) \sum_y yQ'_{x_1'}(y, c_1) = \pi(x_1''|c_1) \sum_y yQ'_{x_1''}(y, c_1).
$$

Otherwise, if one value is larger than the other, this contradicts the optimality since $\pi_{X_1|C_1}$ can select the value that yields a larger value than the other. In case of $Q'(c_1) = 0$, 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 $S$, there always exists a deterministic mixed policy $\pi \in \Pi$ such that $X(\pi)$ and $C(\pi)$ are disjoint and $\mu_{\pi} = \mu_S^*$. 

---

**Algorithm 1 Separation of actions and contexts of an MPS**

1. function sep-mps($S, G$)
   2. input: a mixed policy scope $S$, a causal graph $G$
   3. output: an updated, action-context separated mixed policy scope $	ilde{S}$
   4. for $X \in$ topological-order($X(S); G_S$) do
      5. Replace $C_X$ in $S$ by $(\bigcup_{X' \in C_X \cap X} C_{X'}) \cup (C_X \setminus X)$.  
   6. return $S$

$$
= \sum_{y, x, c} yP_x(y|c^-) \prod_{x \in X} \pi(x|c_x) P_x(c^-) \quad \text{CR (38)}
$$

$$
= \sum_{y, x, c} yP_x(y, c^-) \prod_{x \in X} \pi(x|c_x), \quad \text{CR (39)}
$$
Proof. Let a mixed policy $\rho \sim \mathcal{S}$ be optimal with respect to $\mathcal{S}$. First, there exists an optimal deterministic mixed-policy $\rho'$ equivalent to $\rho$ with respect to the expected reward (Prop.[1]). Since the graph $\mathcal{G}_\rho$ is acyclic, there exists a topological order among $X$. Consider $X \in X$ such that $C_{X'} \cap X = \emptyset$ for every $X' \in X \cap X$. We can create a new function $\pi_X$ based on $\rho_X$ and $\rho'_{X'}$:

$$\pi_X((c_X \{x'\}) \cup c_X) = \rho_X' (c_X \{x'\} = c_x \{x'\}, X' = \rho'_{X'}(c_X \{x'\})).$$

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

**E Non-Redundant Mixed Policy**

**Theorem 1.** Let $S = \{(X, C_X)\}_{X \in X}$ be an MPS and let $\mathcal{H} = \mathcal{G}_S$. $S$ is non-redundant if and only if (i) $X \subseteq an(Y)_H$ and (ii) $(C \perp Y \mid C_X \{C\})$ in $\mathcal{H}\{X\}$, for every $X \in X$ and $C \in C_X$.

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

$$\mu_{\pi} = \sum_{y,x,c_x} yQ_x'(y,c_x)\pi(x|c_x) \quad \text{def}$$

$$= \sum_{y,x,c_x} yQ_x'(y|c_x)Q_x'(c_x)\pi(x|c_x) \quad \text{CR}$$

$$= \sum_{y,x,c_x} yQ_x'(y|c_x)Q'(c_x)\pi(x|c_x) \quad \text{violation of (i)}$$

$$= \sum_{y,x,c_x} yQ'(y, c_x)\pi(x|c_x) \quad \text{CR}$$

$$= \sum_{y,c_x} yQ'(y, c_x) \sum x \pi(x|c_x) \quad \text{algebra}$$

$$= \sum_{y} yQ'(y) \quad \text{MP}$$

$$= \mu_{\pi \{X\}} \quad \text{def}.$$ 

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

$$\mu_{\pi} = \sum_{y,x,c_x} yQ_x(y|c_x)Q_x'(c_x)\pi(x|c_x) \quad \text{def, CR}$$

$$= \sum_{y,x,c_x} yQ_x(y|c_x)Q_x'(c_x)\pi(x|c_x) \quad \text{violation of (ii)}$$

$$= \sum_{y,x,c_x} yQ_x(y|c_x)Q_x'(c_x)Q_x(c_x)\pi(x|c_x) \quad \text{CR}$$

$$= \sum_{y,x,c_x} yQ_x(y|c_x) \sum_{c} Q_x'(c|c_x)\pi(x|c_x) \quad \text{CR, algebra}$$

$$= \sum_{y,x,c_x} yQ_x(y|c_x) \sum_{c} Q_x(c|c_x)\pi(x|c_x) \quad \text{R1}$$

$$= \sum_{y,x,c_x} yQ_x(y|c_x) \sum_{c} P_{\pi}(c|c_x)\pi(x|c_x) \quad \text{R3}.$$
We show, for an arbitrary MPS \((I)\) (e.g., variables pass bits to the downstream. This principle is also applied to different groups since the 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.

\[
\sum_{y, x, c \in \mathbb{Z}} yQ'_f(y, c^-)\pi''(x | c^-) = \mu_{\pi},
\]

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

**Proof.** We show, for an arbitrary MPS \(S' \subseteq S\), that we can construct an SCM \(M \sim G\) such that a mixed policy \(\pi \sim S\) satisfies \(\mu_\pi = \mu_{\pi'}\) for \(\pi' \subseteq \pi\). Let \(S'\) and \(S\) differ on \(\langle X, C_X \rangle\) (either only a subset of \(C_X\) or \(X\) itself). We consider a minimal edge subgraph \(H\) of \(G\) such that the above conditions are satisfied for \(X\) and \(C_X\). The graph is characterized as groups of context variables \(\{C_X\}\); where, for each group \(C_X\), there exists a subset of context variables \(C_X'\) connected via bidirected paths and each of the rest \(C_X \setminus C_X'\) has a directed path towards the subset \(C_X'\). More precisely speaking, there exists no \(C_X\) appearing in the (bi)directed paths as non-ends, while the paths in general can intersect. If the group is a singleton \(C_X = \{C\}\), either a bidirected or directed path towards \(Y\), not passing \(X\), exists. Otherwise, a bidirected path exists between \(C \in C_X'\) and another (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. 12(d) above mentioned paths can intersect with each other. We consider each bidirected and directed path maintains its own ‘channel’ where the path in \(H\) 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 \(C_X \setminus C_X'\)) behaves as a fair coin or a vector of independent fair coins if it involves in multiple paths (e.g., \(Z\) in Fig. 12(d) has 6 bits for every pair among \(\{C_1, C_2, C_3, C_4\}\)). We design the function for every \(C \in C_X\) 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 \(\pi\) becomes 1.0 as every bit-parity is counted twice and complemented at \(Y\). But any \(\pi'\), whose \(X\) does not listen to \(C_X\) as a whole, makes its expected reward 0.5.

**Proposition 6.** Given an MPS \(S\), every intermediate MPS \(nr\text{-mps}\) is valid MPS.

**Proof.** Let \(S\) be an intermediate MPS at a point of time in the execution of \(nr\text{-mps}\). Since the algorithm removes a subset of \(X(S)\) and that of contexts \(C_X\) for some \(X \in X(S)\), 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

![Diagram](image-url)
Algorithm 2 Non-Redundant Mixed Policy Scope $\mathcal{S}$

1: function NR-MPS($\mathcal{G}, Y, S$)
    input: a mixed policy scope $S$, a causal graph $\mathcal{G}$
    output: an updated, non-redundant mixed policy scope $\hat{S}$
2: $S \leftarrow S \setminus (X \setminus \text{an}(Y))_{|\mathcal{G}_{\mathcal{S}}}$.
3: for $X \in \text{reverse-order}(X(S); \mathcal{G}_{\mathcal{S}})$ do
4:     if $X \not\in \text{an}(Y)_{|\mathcal{G}_{\mathcal{S}}}$ then:
5:         $S \leftarrow S \setminus \{X\}$ and continue.
6:     for $C \in \text{pa}_X$ do
7:         if $(C \perp Y | C_X \setminus \{C\})$ in $\mathcal{G}_{\mathcal{S}} \setminus \{X\}$ then
8:             $S \leftarrow (S \setminus \{X\}) \cup \{(X, C_X \setminus \{C\})\}$.
9:     return $S$.

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 $S$ results in an induced graph $\mathcal{G}_{S'}$ which is an edge subgraph of $\mathcal{G}_S$ where $S'$ is an MPS after Line 8. Hence, Line 8 does not create a cycle from acyclic $\mathcal{G}_{S'}$. Through Lines 2 and 8, a subset of or an element of $X(S)$ is removed. Not only are its induced edges (i.e., $C \rightarrow X$ for $C \in \text{pa}_X$) 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($X$) to $X$ with respect to the acyclicity of the updated induced graph after Line 2 and 8.

(Line 2) Let $S$ and $S'$ be MPSes before and after removing $X' \subseteq \text{an}(Y)_{|\mathcal{G}_{\mathcal{S}}}$ at Line 2. If $X' = \emptyset$, then done. Otherwise, let $\mathcal{H}$ be $\mathcal{G}_{S}[\text{an}(Y)_{|\mathcal{G}_{\mathcal{S}}}]$ the induced graph restricted to the ancestors of $Y$, which is the same after the removal, $\mathcal{G}_{S'}[\text{an}(Y)_{|\mathcal{G}_{\mathcal{S}'}}]$. Since $\mathcal{G}_{S}$ is acyclic, its subgraph $\mathcal{H}$ is acyclic, as well. Let $\mathcal{H}'$ be $\mathcal{G}_{S'} \setminus \mathcal{V}(\mathcal{H})$, the vertex-induced subgraph of $\mathcal{G}_{S'}$ by excluding the ancestors of $Y$. By definition of $S'$, there is no $X \in \mathcal{X}(S')$ 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 $S'$, 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 $\mathcal{G}_{S'}$.

(Line 8) Let $X$ be the action to be removed. Let $\mathcal{H}' = \mathcal{G}_{S'}[\text{de}(X)_{|\mathcal{G}_{\mathcal{S}'}}]$ and $\mathcal{H} = \mathcal{G}_{S} \setminus \mathcal{V}(\mathcal{H}')$. $\mathcal{H}$ is acyclic since it is the same as $\mathcal{G}_{S} \setminus \mathcal{V}(\mathcal{H})$, the subgraph of acyclic $\mathcal{G}_{S'}$. By construction, there exists no $X' \in \mathcal{X}(S')$ 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 $\mathcal{V}(\mathcal{H}')$ to $\mathcal{V}(\mathcal{H})$ in $\mathcal{G}_{S'}$. Consequently, $\mathcal{G}_{S'}$ 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 $S$, nr-mps returns a unique, maximal non-redundant MPS of $S$.

**Proof.** The algorithm refines a given mixed policy scope (MPS) $S$ by iterating over $\mathcal{X}(S)$ and $\mathcal{C}(S)$ 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 $S^\perp$. Let $S'$ be an intermediate MPS when an arbitrary $X$ is under examination (Line 4). $X \in \text{an}(Y)$ in $\mathcal{G}_{S'}$ is preserved in $\mathcal{G}_{S^\perp}$ because the later
we would like to ensure that the edges in the path are intact throughout the changes made by the policy. We restrict our attention to a collider-minimal, shortest path. Every collider \( W \) in the path has a directed path towards \( Y \) through \( W \rightearrow C_i \rightearrow X \rightearrow Y \), where \( C_i \in C_X \) by the testing criteria and \( W \rightearrow C_i \) can be of zero length. Let \( \mathcal{G}_S \subseteq \mathcal{G}_{S'} \) be the path graph together with directed paths between colliders and conditionals. Let \( S'' \) be the MPS at the end of testing every \( C_i \in C_X \). We show that \( (C \not\perp\!
mid Y | C''_X) \) in \( \mathcal{G}_{S''} \{ X \} \) holds true (with \( C''_X \) as in \( S'' \)). Let \( C_i \) be the subset of \( C_X \) that associates with the colliders in the path. Since every \( C_i \in C_i \) will have a back-door path to \( Y \) (by concatenating a directed path between \( C_i \) and \( W \) and the subpath of \( \rho \) between \( W \) and \( Y \)) given \( C_i \} \subset C_i \). Hence, the result follows (for an illustrative example, please see Fig. [14]).

Now we investigate whether the path \( \rho \) between \( C \) and \( Y \) is still valid in \( \mathcal{G}_S \) given \( C_X \). Specifically, we would like to ensure that the edges in the path are intact throughout the changes made by the algorithm. The removal of \( X' \in X(\tilde{S}) \) may affect the parents of \( X' \) and the removal of \( C' \in C_X \) affects \( C' \rightarrow X' \). That is, for both cases, we investigate whether \( C' \rightarrow X' \) in \( \rho \) 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.** \( C' \not\perp\!
mid Y \) in \( \mathcal{G}_\rho \{ X' \} \) demonstrates the existence of a collider-free \( d \)-connection \( \phi \), which is disjoint with \( C_X \{ C' \} \).

The directed path between \( X' \) and \( Y \) will be valid in \( \mathcal{G}_{S''} \) if every \( C'' \rightarrow X'' \) appeared in the path is intact, which depletes its validity to the bottom-most \( X'' \in X(\tilde{S}) \). \( C' \in C_X \) will be dependent to \( Y \) given \( C_X \{ C' \} \) as no \( C_X \{ C' \} \) 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, \( \mathcal{G}_\rho \subseteq \mathcal{G}_{S''} \) and \( C' \in C_X \).

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 \( \rho \) can be one of the following two forms

\[
C \bullet \bullet \cdots \bullet \rightarrow X' \bullet \leftarrow Y \quad \text{or} \quad C' \bullet \bullet \cdots \bullet \leftarrow X' \bullet \rightarrow Y
\]

where \( \times \) represents that the path may have colliders. \( X' \) has a directed path to (i) \( C \), (ii) \( Y \), or (iii) some \( C_a \in C_X \} \) via a collider \( W \) (which can be \( X' \) itself) in the path. Then, a directed path can be of the form: (i) \( X' \leftarrow \rightarrow C \rightarrow X \leftarrow \rightarrow Y \), (ii) \( X' \rightarrow \rightarrow Y \), or (iii) \( X' \rightarrow \rightarrow \rightarrow W \rightarrow \rightarrow C_a \rightarrow X \leftarrow \rightarrow Y \).

**Claim 2.** \( C' \not\perp\!
mid Y \) in \( \mathcal{G}_\rho \{ X' \} \) demonstrates the existence of a collider-free \( d \)-connection \( \phi \), which is disjoint with \( C_X \{ C' \} \).

**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) \( C \bullet \bullet \cdots \bullet \rightarrow C' \rightarrow \rightarrow Y \) where both avoids passing through \( X' \).

For (a), if the subpath does not contain a collider, (a1) \( C' \bullet \bullet \cdots \bullet \rightarrow C \rightarrow X \rightarrow \rightarrow Y \) is a valid trail signaling \( C' \not\perp\!
mid Y \) in \( \mathcal{G}_\rho \{ X' \} \). Otherwise, there exists a collider \( W \) in the subpath, which uses a shortcut to \( Y \) through \( C_a \in C_X \{ X \} \): (a2) \( C' \bullet \bullet \cdots \bullet \rightarrow W \leftarrow \rightarrow C_a \rightarrow X \leftarrow \rightarrow Y \).

In case of (a1), \( C'' \in C_X \{ C' \} \) cannot appear in between \( C' \) and \( C \) since otherwise it violates the fact that \( \rho \) being the shortest—one can create \( \rho ' \) by replacing \( \cdots C'' \rightarrow C' \rightarrow X \rightarrow \rightarrow Y \) by \( \cdots C'' \rightarrow X \rightarrow \rightarrow Y \).

In case of (a2), similarly, \( C' \) cannot reside along \( C' \bullet \bullet \cdots \bullet \rightarrow W \). In addition, we prove that \( C'' \) does not exist in-between \( W \rightarrow \rightarrow C_a \) since otherwise there exists \( \rho ' \) which does not require \( W \) in the path \( \rho \) as a collider making use of \( C'' \rightarrow X' \), contradicting the collider-minimality of \( \rho \). In either cases (a1) and (a2), \( C' \rightarrow X \rightarrow Y \) or \( C_a \rightarrow X \leftarrow \rightarrow Y \) are \( C'' \)-free otherwise it contradicts the topological order between \( X' \) and \( X \).
In case of (b), either there exists (b1) a path towards $Y$, $C' \rightarrow Y$ without a collider, or (b2) through a collider as seen in the case of (a2), where the same proof is applicable. Given (b1), again, if $C''$ exists in the path, we can shorten $\rho$ by connecting $X' \leftarrow C''$ while replacing $X' \leftarrow C' \rightarrow C''$, which violates $\rho$ being the shortest.

Hence, the result follows.

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.

\[
\mu_\pi = \sum_{y,x,c} y Q(y|x,c) Q(x,c) \quad \text{MP,CR (40)}
\]

\[
= \sum_{y,x,c} y Q(y|x,c_1,c_2) Q(x,c) \quad \text{R1 (41)}
\]

\[
= \sum_{y,x,c} y Q'_x(y|c_1,c_2) Q(x,c) \quad \text{R1,R2 (42)}
\]

\[
= \sum_{y,x,c} y Q'_x(y|c_1,c_2) Q(x_1|c_3,c_1) Q(x_2|c_3,c_2) Q(c) \quad \text{CR,R1 (43)}
\]

\[
= \sum_{y,x,c_1,c_2} y Q'_x(y|c_1,c_2) Q(c_1,c_2) \sum_{c_3} Q(x_1|c_3,c_1) Q(x_2|c_3,c_2) Q(c_3) \quad \text{MP,CR (44)}
\]
\[ \sum_{c_3} Q(c_3) \sum_{y,x,c_1,c_2} y Q'_x(y|c_1,c_2) Q(c_1,c_2) Q(x_1|c_3,c_1) Q(x_2|c_3,c_2) \quad \text{algebra (45)} \]
\[ \leq \sum_{y,x,c_1,c_2} y Q'_x(y|c_1,c_2) Q(c_1,c_2) Q(x_1|x_3^*,c_1) Q(x_2|x_3^*,c_2) \quad \text{def (46)} \]
\[ = \sum_{y,x,c_1,c_2} y Q'_x(y|c_1,c_2) Q(c_1,c_2) \pi'(x_1|c_1) \pi'(x_2|c_2) \quad \text{def (47)} \]
\[ = \sum_{y,x,c_1,c_2} y Q'_x(y|c_1,c_2) Q_x^* (c_1,c_2) \pi'(x_1|c_1) \pi'(x_2|c_2) \quad \text{R1,R3 (48)} \]
\[ = \sum_{y,x,c_1,c_2} y Q'_x(y|c_1,c_2) \pi'(x_1|c_1) \pi'(x_2|c_2) \quad \text{CR (49)} \]

where \( c_3^* \in X_{c_3} \) is the value maximizing the inner sum.

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

\[ \mu_\pi = \sum_{y,x,c} y Q(y|x,c) Q(x,c) \quad \text{def (50)} \]
\[ = \sum_{y,x,c} y Q(y|x_1,c) Q(x,c) \quad \text{R1 (51)} \]
\[ = \sum_{y,x,c} y Q'_x(y|c_1) Q(x,c) \quad \text{R1,R2 (52)} \]
\[ = \sum_{y,x,c} y Q'_x(y|c_1) Q(c_2|c_1) Q(x_1|c) Q(x_2|c) \quad \text{CR (53)} \]
\[ = \sum_{y,x,c} y Q'_x(y|c_1) Q_x^* (c_1) Q(c_2|c_1) Q(x_1|c) Q(x_2|c) \quad \text{R1,R3 (54)} \]
\[ = \sum_{y,x,c} y Q'_x(y|c_1) Q(c_2|c_1) Q(x_1|c) Q(x_2|c) \quad \text{CR (55)} \]
\[ = \sum_{y,x,c} y Q'_x(y,c_1) \sum_{c_2} Q(c_2|c_1) Q(x_1|c) Q(x_2|c) \quad \text{algebra (56)} \]
\[ = \sum_{y,x} \sum_{c_1,c_2} Q(c_2|c_1) \sum_{y,x} y Q'_x(y,c_1) Q(x_1|c) Q(x_2|c) \quad \text{algebra (57)} \]
\[ \leq \sum_{c_1,c_2} y Q'_x(y,c_1) Q(x_1|c_1,c_2^*(c_1)) Q(x_2|c_1,c_2^*(c_1)) \quad \text{def (58)} \]
\[ = \sum_{y,x,c_1} y Q'_x(y,c_1) \pi'(x_1|c_1) \pi'(x_2|c_1) \quad \text{def (59)} \]

The derivation for Fig. 5d is as follows. Let \( Q' = P_{\pi'|\{x_1,X_2}\} \).

\[ \mu_\pi = \sum_{y,x_{12},c_2} y Q(y|x_{12},c_2) Q(x_{12},c_2) \quad \text{MP,CR (60)} \]
\[ = \sum_{y,x_{12},c_2} y Q'_{x_{12}}(y|c_2) Q(x_{12},c_2) \quad \text{R1,R2 (61)} \]
\[ = \sum_{y,x_{12},c_2} y Q'_x(y|c_2) \sum_{x_3} Q(x,c_2) \quad \text{MP (62)} \]
\[ = \sum_{y,x_{12},c_2} y Q'_{x_{12}}(y|c_2) \sum_{x_3} Q(c_2) Q(x|x_3|c_2) Q(x_1|x_3,c_2) Q(x_2|x_3,c_2,x_1) \quad \text{CR (63)} \]
\[ = \sum_{y,x_{12},c_2} y Q'_x(y|c_2) Q'_{x_{12}}(c_2) \sum_{x_3} Q(x|c_2) Q(x_1|x_3,c_2) Q(x_2|x_3,c_2) \quad \text{R2,R3 (64)} \]
\[ = \sum_{y,x_{12},c_2} y Q'_x(y|c_2) \sum_{x_3} Q(x|c_2) Q(x_1|x_3,c_2) Q(x_2|x_3,c_2) \quad \text{CR (65)} \]
The first condition can be viewed as both (i) rule 1 of do-calculus with respect to the regime nodes \(\pi_X\) and (ii) rule 1 for non-action contexts of interest. If there exists a subset of exogenous variables \(U'\) in \(G_S\), a subset of endogenous variables \(Z\) in \(G_S\) that disjoint with \(C' \cup X'\) and subsumes \(C_{X'} \cap (C' \cup X')\), and an order \(\prec\) over \(V' = C' \cup X' \cup Z\) such that

1. \((Y \perp \pi_X | [X' \cup C'])_{G_S}\),
2. \((C \perp \pi_{X_{<C}}, X_{<C}, U' | ([X' \cup C']_{<C}])_{G_S}\) for every \(C \in C'\), and
3. \(V'_{<X}\) is disjoint with \(de(X)_{G_S}\) and subsumes \(pa(X)_{G_S}\) for every \(X \in X'\),

then, the expected reward for \(\pi\), a deterministic policy optimal with respect to \(S\), can be written as

\[
\mu_\pi = \sum_{y,c',x'} yQ_{X'}(y,c') \sum_{u',z} Q(u') \prod_{Z \in Z} Q(z|V'_{<z}, u') \prod_{X' \in X'} \pi(x|c_x). \tag{3}
\]

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

\[
\mu_\pi = \sum_{y} yQ(y) \tag{69}
\]

\[
= \sum_{y,c',x'} yQ(y, c', x') \tag{70}
\]

\[
= \sum_{y,c',x'} yQ(y|c', x')Q(c', x') \tag{71}
\]

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

\[
= \sum_{y,c',x'} yQ_{X'}(y|c') Q(c', x') \tag{72}
\]

\[
= \sum_{y,c',x'} yQ_{X'}(y|c') \sum_{z,u'} Q(c', x', z, u') \tag{73}
\]

\[
= \sum_{y,c',x'} yQ_{X'}(y|c') \sum_{z,u'} \prod_{C' \in C'} Q(c|V'_{<c}, u') \prod_{V \in V'_{<V}} Q(v|V'_{<v}, u') \tag{74}
\]

The second condition corresponds to (i) rule 1 for \(Z_{<C}\) and \(U'\) and (ii) rule 2 for \(X'_{<C}\),

\[
= \sum_{y,c',x'} yQ_{X'}(y|c') \prod_{C' \in C'} Q'(c|X'_{<c'}, C'_{<C'}) \sum_{z,u'} \prod_{V \in V'_{<V}} Q(v|V'_{<v}, u') \tag{75}
\]

\[
= \sum_{y,c',x'} yQ_{X'}(y|c') \prod_{C' \in C'} Q'(c|X'_{<c'}, C'_{<C'}) \sum_{z,u'} \prod_{V \in V'_{<V}} Q(v|V'_{<v}, u') \tag{76}
\]

\[
= \sum_{y,c',x'} yQ_{X'}(y|c') \prod_{C' \in C'} Q'(c|X'_{<c'}, C'_{<C'}) \sum_{z,u'} \prod_{V \in V'_{<V}} Q(v|V'_{<v}, u') \tag{77}
\]

\[
= \sum_{y,c',x'} yQ_{X'}(y|c') Q'(c') \sum_{z,u'} \prod_{V \in V'_{<V}} Q(v|V'_{<v}, u') \tag{78}
\]

\[
= \sum_{y,c',x'} yQ_{X'}(y|c') \sum_{z,u'} \prod_{V \in V'_{<V}} Q(v|V'_{<v}, u') \tag{79}
\]
The third condition ensures that the conditionals for a term \(Q(x|\cdot)\) can be refined only to the parents of \(X\), which is \(C_X\),

\[
= \sum_{y,c',x'} yQ_{X'}^{c}(y,c') \sum_{z,u'} Q(u') \prod_{z \in Z} Q(z|v_{<z}, u') \prod_{x \in X'} Q(x|c_x) \qquad \text{C3.R1} \quad (80)
\]

\[\square\]

**Theorem 2.** Let \(U', Z, \) and \(<\) satisfy Lemma 1. For \(Z \in \mathbf{Z}\), let \(V_Z\) be a minimal subset of \(v_{<z} \cup U'\) such that \(Q(Z | V_Z) = Q(Z | V_{<z} \cup U')\). We define \(\text{fix}(T)\) with respect to \(\{\langle Z, V_Z \rangle \} \in \mathbf{Z},\) that is, with \(T = \{T\} \cup \{Z \in \mathbf{Z} | V_Z \cup U' \subseteq \{T\}\}, \) \(\text{fix}(T)\) is \(T\) if \(T = T\) and \(\text{fix}(T)\), otherwise. If \(\text{fix}(C_X \setminus Z) \supseteq C_X\) for \(X \in X'\), then, \(S' = (S \setminus X') \cup \{(X, C_X \setminus Z)\} \in X'\) satisfies \(\mu_{S'} = \mu_S\).

**Proof.** Let \(Q = P_\pi\) and \(Q' = P_{\pi'|X'}\).

\[
\mu_\pi = \sum_{y,c',x'} yQ_{X'}^{c}(y,c') \sum_{z,u'} Q(u') \prod_{z \in Z} Q(z|v_{<z}, u') \prod_{x \in X'} Q(x|c_x) \quad \text{Lemma 1}
\]

\[
= \sum_{y,c',x'} yQ_{X'}^{c}(y,c') \sum_{z,u'} Q(u') \prod_{z \in Z} Q(z|v_{<z}) \prod_{x \in X'} Q(x|c_x, z, c_x \cap z) \quad \text{def}
\]

\[
= \sum_{u'} Q(u') \left( \sum_{y,c',x'} yQ_{X'}^{c}(y,c') \sum_{z \in Z} Q(z|v_{<z}) \prod_{x \in X'} Q(x|c_x, z, c_x \cap z) \right) \quad \text{algebra}
\]

Choose \(u'^*\) maximizing the term in the parentheses (note that \(V_Z\) may intersect \(U'\)),

\[
\leq \sum_{y,c',x'} yQ_{X'}^{c}(y,c') \sum_{z \in Z} Q(z|v_{<z}) \prod_{x \in X'} Q(x|c_x, z, c_x \cap z) \quad \text{def}
\]

Take \(Z_1 \in \mathbf{Z}\), the first element among \(Z\) with respect to \(<\).

\[
= \sum_{y,c',x'} yQ_{X'}^{c}(y,c') \sum_{z_1} Q(z_1|v_{<z_1}) \prod_{x \in X'} Q(x|c_x, z_1, c_x \cap z_1) \cdots
\]

Choose \(z_1^*\) for each free variable \(v_{z_1} \setminus U'\),

\[
\leq \sum_{y,c',x'} yQ_{X'}^{c}(y,c') \sum_{z \setminus \{z_1\}} Q(z|v_{<z}) \prod_{z \in Z \setminus \{Z_1\}} Q(|v_{<z} \setminus U' \cup \{Z_1\} \cap v_{z_1} \setminus U'|, |v_{z_1} \setminus U'|, v_{z_1} \cap u'^*)_z \prod_{x \in X'} Q(x|c_x, z, c_x \cap z, z_1^*|v_{z_1} \setminus U'|, v_{z_1} \cap u'^*)_z
\]

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

\[
z_1^* = z_1^*(v_{z_1} \setminus U', v_{z_1} \cap u'^*) \quad \text{base case}
\]

\[
z_i^* = z_i^*(v_{z_i} \setminus (U' \cup Z_{<i}), v_{z_i} \cap (u'^* \cup \{z_j^*(\cdot)\}_{j=1}^{i-1})) \quad \text{for } i > 1
\]

Then,

\[
\mu_\pi \leq \sum_{y,c',x'} yQ_{X'}^{c}(y,c') \prod_{x \in X'} Q(x|c_x, \{z_i^*(\cdot)\}_{i=1}^m).
\]

We want to ensure that \(c_x \cap \{z_i^*(\cdot)\}_{i=1}^m\) is a function of \(c_x \setminus Z\). 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 \(X(S)\). If the values for \(C_X \setminus Z\) can be determined from the value of \(C_X \setminus Z\), then, for some \(\pi\),

\[
\mu_\pi \leq \sum_{y,c',x'} yQ_{X'}^{c}(y,c') \prod_{x \in X'} \pi(x|c_x, Z).
\]

This completes the proof. \[\square\]
Figure 15: A more involved example for redundancies in a mixed policy scope

Figure 16: (a) a dependency specified $Q(Z|\cdot)$ terms with additional ‘implicating’ relationships (directed dashed edges onto $X_1$ and $X_1$) and exogenous variables; (b) $X_1$ requires to fix $C_2$ given $C_1$ (blue for given) where an unknown $u_2^*$ can be marginally fixed and $c_2^2(c_1, u_2^*)$ 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 $c_2^2(c_1, u_2^*)$. 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 $\prec = (C_1, C_4, C_2, C_5, X_1, X_3, C_3, C_6, X_2, X_4)$.

\[
\begin{align*}
\mu_{\pi} &= \sum_{y, x, c_{14}} y Q(y | x, c_{14}) Q(x, c_{14}) \\
&= \sum_{y, x, c_{14}} y Q_s^y(y | c_{14}) Q(x, c_{14}) \\
&= \sum_{y, x, c_{14}} y Q_s^y(y | c_{14}) \sum_{c_{2356}, x, c_{14}, u^{'}} Q(c_{2356}, x, c_{14}, u^{'}) \\
&= \sum_{y, x, c_{14}} y Q_s^y(y | c_{14}) \sum_{c_{2356}, u^{'}} Q(u^{'}) Q(c_{14}) Q(c_2 | c_1, u_2) Q(c_5 | c_4, u_5) Q(x_1 | c_{12}) \\
&= \sum_{y, x, c_{14}} \prod_{c_{2356}, u^{'}} Q(x_3 | c_{45}) Q(c_3, c_2, u_4) Q(c_5, x_3, u_6) Q(x_2 | c_{123}) Q(x_4 | c_{456}) \quad \text{CR.R1} \\
&= \sum_{y, x, c_{14}} \prod_{c_{2356}, u^{'}} Q(x_3 | c_{45}) Q(c_3, c_2, u_4) Q(c_5, x_3, u_6) Q(x_2 | c_{123}) Q(x_4 | c_{456}) \quad \text{R3.CR} \\
&= \sum_{c_{2356}, u^{'}} Q(u^{'}) Q(c_3, c_2, u_4) Q(c_5, c_1, u_5) Q(x_1 | c_{12}) \quad \text{defines dependency} \\
&= \sum_{y, x, c_{14}} y Q_s^y(y, c_{14}) Q(x_1 | c_1, c_2) Q(x_2 | c_1, c_2) Q(x_3 | c_4, c_5) Q(x_4 | c_4, c_{56}) \quad \text{R1}
\end{align*}
\]

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

\[
\leq \sum_{y, x, c_{14}} y Q_s^y(y, c_{14}) \pi'(x_1 | c_1) \pi'(x_2 | c_1) \pi'(x_3 | c_4) \pi'(x_4 | c_4)
\]
A variant of the setting used in Figs. 7a to 7c where non-contextualizable variables are treated as if they are contextualizable (proof).

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 $X$ 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 $X$ 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. 

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.

A Partial Order over Mixed Policies and Possible-Optimality

Theorem 3. Given an MPS $S$, let $S' \neq S$ be an MPS with $X(S) \subseteq X(S')$ such that $H''$ the union of induced graphs $G_S \cup G_{S'}$ is acyclic. Let $X'$ be actions that the MPSes disagree on, i.e., $(X(S') \setminus X(S)) \cup \{X \in X(S) \mid C'_X \neq C_X\}$, and (invalid) MPS $S'' = \{\langle X, pa(X)_{H''} \cup U_X \rangle \}_{X \in x'}$. $\mu_{S''} = \mu_{S'}$ can be elicited by Thm. 2 then, $\mu_S \leq \mu_{S''}$. 

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 ($C^*$) 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 $S''$ induced graph. Since more or the same contexts are used for $S''$ than both $S$ and $S'$, $\mu_S \leq \mu_{S''}$ and $\mu_{S'} \leq \mu_{S''}$. If we elicit $\mu_{S''} \leq \mu_{S'}$, we can conclude that $\mu_S \leq \mu_{S''} \leq \mu_{S'}$. 

We present an example for $\mu_{S'} \leq \mu_{S''}$ where Fig. 7a and Fig. 7b present $G_S$ and $G_{S'}$, respectively.

\[
\mu_{S''} = \sum_{y, c, u_{x, z}} yP_x(y|c)Q(c_1)Q(u_{x_1}|c_1)Q(u_{x_2}|c_1, u_{x_1})Q(x_1|c_1, u_{x_2})Q(x_2|c_1, u_{x_2})Q(x_3|c_1, u_{x_2})Q(x_4|c_1, u_{x_2})
\]

A variant of the setting used in Figs. 1a to 1c 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, $W \indep X$ 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.

\[
\mu_{S''} = \sum_{y, c, u_{x, z}} yP_x(y|c)Q(c_1)Q(w|c_1)Q(u_{x}|c_1, w)Q(x_1|c_1, u_{x}, w)Q(x_2|c_1, x_1, u_{x}, w)Q(x_2|c_1, u_{x}, w)
\]
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.

\[
\begin{align*}
&= \sum_{y, c, x, w, u_x} y P_x(y|c)Q(x|c_1)Q(w|c_1)Q(u_x|x_1, u_{x_1}, w)Q(c_2|x_1, c_1)Q(x_2|c_2, u_{x_2}) \\
&= \sum_{y, c, x, w, u_x} y P_x(y, c)Q(x_1|c_1, u_{x_1}, w)Q(x_2|c_2, u_{x_2}) \\
&\leq \sum_{y, c, x} y P_x(y, c)Q(x_1|c_1, u_{x_1}, w^*(c_1))Q(x_2|c_2, u_{x_2}) \\
&\leq \sum_{y, x} y P_x(y, c)\pi' (x_1|c_1)\pi'(x_2|c_2)
\end{align*}
\]

**Proposition 5.** Given $\langle G, Y, X^*, C^* \rangle$, let $X' = \{ X \in X^* \mid de(X)_G \cap (C^* \cup \{ Y \}) \neq \emptyset \}$. $X'' = de(Y)_G \cap X'$, and $Z = de(Y)_{G_{\bar{X}}}$. The POMPSs for $\langle G, Y, X^*, C^* \rangle$ are the same as those for $\langle G, Z, Y', X', C^* \cap Z \setminus X'' \rangle$.

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

1. a step to $\langle G, Y, X', C^* \rangle$: First, we justify the reduction of $X^*$. Consider $X \in X^*$ having no $C^*$ or $Y$ as descendants (exclusive) in $G$. 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. 3) $X$ 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, $X' = \{ X \in X^* \mid de(X)_G \cap (C^* \cup \{ Y \}) \neq \emptyset \}$ is the subset of $X^*$ that can become an action in POMPS.

2. a step to $\langle G, Y, X', C^* \setminus Z \rangle$: Next, we explain the reduction of $C^*$ focusing on removing $Z$. Any $C \in C^*$ that is a descendant of $Y$ without any $X'$ 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., $de(Y)_{G_{\bar{X}}} \cap C^*$ are not contextualizable. With $X'' = de(Y)_{G_{\bar{X}}} \cap X'$ (action variables that are descendants of $Y$ with no other action variables in between), we can elicit $de(Y)_{G_{\bar{X}}} \cap C^* = de(Y)_{G_{\bar{X}}} \cap C^*$. Hence, $C^*$ can be reduced to $C^* \setminus Z$ where $Z = de(Y)_{G_{\bar{X}}}$.

3. a step to $\langle G \setminus Z, Y, X', C^* \setminus X'' \setminus Z \rangle$: Finally, we examine eliminating $Z$ from the graph. Since $Z$ 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 $Z$. Hence, the removal of $Z$ does not affect the reward landscape among the MPSes. However, once if $Z$ is removed from the setting, then $X''$ is no longer descendant of $Y$ in the resulting setting, and the constraint that $de(X'')_{G \setminus X''}$ 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 $X''$ 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 $X'$ 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 $X''$ leads to a better policy scope, thus, they are indeed non-POMPSes. As a consequence, any MPSes in the new setting satisfies the constraint and is a valid POMPS in the original setting (vice versa).