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 [12–15] 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].

With A standard contextual bandit (CB) optimizes a policy \( \pi_{CB} \) of the four MPSes \( \pi_{CB} \) (Fig. 2), where Roughly speaking (to be formalized later on), equivalence class among scopes with respect to optimal rewards. every and can be classified based on two desiderata: different modes of interaction can be represented as induced graphs can become a context, there are 15 mixed policy scopes. These different expected reward in this environment is 1.0. this setting, the best policy is intervening only on \( X \) and \( Y \) and can be suboptimal, i.e., \( \mu_{\pi}^{*} = \mu_{\pi_{CB}} \) can be suboptimal, i.e., \( \mu_{\pi}^{*} = \mu_{\pi_{CB}} < \mu^{*} \) where \( \mu^{*} \) is the optimal expected reward. To ground what this means, let every variable be binary and \( U_{1} \) and \( U_{2} \), the UCs adjacent to \( X_{1} \) and \( X_{2} \), be fair coins and \( \epsilon \) be a noise over \( X_{1} \) following \( P(\epsilon = 1) = 0.2 \). Also, let the unobserved causal mechanisms be specified as \( X_{1} \leftarrow U_{1} \oplus \epsilon, C \leftarrow U_{1}, X_{2} \leftarrow U_{2} \oplus X_{1} \oplus C, \) and \( Y \leftarrow (1 - (X_{2} \oplus U_{2})) \vee C, \) where \( \oplus \) is the exclusive-or operator. Since the policy determines \( X_{2} \) irrelevant to \( U_{2} \) and the context \( C \) is also independent to \( U_{2}, \) we can elicit that \( \mu_{\pi_{CB}}^{*} = 0.75. \) 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: non-redundancy and optimality. We explain these desiderata through an illustration (Fig. 2) of the four MPSes \( S_{a} = \{\}, S_{b} = \{\{X_{2}, \{C, X_{1}\}\}\}, \) and \( S_{c} = \{\{X_{1}, \{C\}\}\}. \) We annotate their relationships with a superset symbol \( \supseteq, \) whether one scope has more actions or contexts than the other, and with a comparison symbol \( \supseteq_{\mu} \) (or \( =_{\mu} \)), whether one’s optimal reward is at least as good as (or equal to) the other’s in every world compatible with a causal graph. The equality forms an equivalence class among scopes with respect to optimal rewards. Roughly speaking (to be formalized later on), non-redundancy refers to the condition of a scope
such that removing any of its actions or contexts can negatively affect its maximum performance. In other words, given two scopes $S$ and $S'$, if $S \supsetneq S'$ and $S =_\mu S'$, then $S$ is said to be redundant. For instance, since $S_c \supset S_e$ while $S_c =_\mu S_e$, the CB policy (Fig. 13) is redundant and the CB agent wastes its resources not only for intervening on $X_1$ (a redundant action) but also for taking $X_1$ into account for $X_2$ (a redundant context). Furthermore, optimality of a scope $S$ represents that there exists no other scope $S'$ (not in the equivalence class of $S$) such that $S' \supsetneq \mu S$. 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 $S_n$ nor $S_d$. 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 [28].

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 $\hat{U}$, instead of $U$, to emphasize the union of two disjoint sets. We use structural causal models (SCMs) [11, 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 $V_i \leftarrow f_i(p_a_i, u_i)$ where $P_{A_i} \subseteq V \setminus \{V_i\}$ and $U_i \subseteq U$. The observational distribution $P(v) = \text{def} \sum_{u} \prod_{v \in V} P(v_i | p_a_i, u_i) P(u)$ 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 $M_{X^c}$, which is $M$ with $f_X$ replaced to $x$ for $X \in X$. Then, an interventional distribution $P_x(v \mid X)$ (or also $P(v \mid X, do(x))$) follows from $M_{X^c}$, and is such that $P_x(v \mid X) = \sum_u \prod_{v \in V \setminus X} P(v_i | p_a_i, u_i) P(u)$.

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) $X \leftarrow Y$ if for a maximal subset $W \subseteq V \setminus \{X\}$ such that $U_W \perp U_X$ and $U_Y \not\perp U_W$. 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 [29, 50]. We denote by $G_{X^c}$ an edge subgraph of $G$ which removes edges incoming to $X$ and outgoing from $Z$. A submodel $M_{X^c}$ can be presented as $G_{X^c}$ with $X$ fixed to $x$. Hence, causal relationships among other variables are captured in $G_{X^c}$, 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 $V'$. [31]. We adopt familial notation, $ch, pa, an, dc$ for children, parents, ancestors, and descendants, respectively, with $Ch, Pa, An, De$ including arguments. Our work utilizes $d$-separation [32, 33] and do-calculus [34]. classic graphical rules to ascertain equalities between distributions. The omitted proofs and derivations are provided in [29].
2 Mixed Policies: Fundamentals & Basic Results

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

**Definition 1** (Mixed Policy Scope (MPS)). Let $G$ 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 $\{\}$, and $G_{\{\}}$, an induced graph for $M_{\{\}}$. Given an MPS $S_{\text{causal}} = \{(X_1, \{C\}), (X_2, \{X_1, C\})\}$, we induce 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 $M_S$.

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

If we consider the MPS $S_{\text{causal}}$ discussed above, its mixed policy $\pi = \{\pi_{X_1 | \{C\}}, \pi_{X_2 | \{C, X_1\}}\}$, 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 $M$, a mixed policy $\pi$ induces a variant of SCM $M_{\pi}$ where the function for $X \in X(\pi)$ is replaced by the corresponding $\pi_{X | C_X}$ (see [35] 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 $X^*$ 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 $X(S) = \{X \mid (X, C_X) \in S\}$ and active contexts $C(S) = \bigcup_{X \in X(S)} C_X$. Similarly, given $\pi \sim S$ (a mixed policy following the MPS), $X(\pi) = X(S)$ and $C(\pi) = C(S)$. 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 $Q' = P_{\pi | X'}$, where $\pi | X'$ 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 $\mu_{\pi} = \mu^* = \max_{\pi' \in \Pi} \mu_{\pi'}$. Restricting our attention to $\Pi_S = \{\pi \in \Pi \mid \pi \sim S\}$, we define $\mu_S^* = \max_{\pi' \in \Pi_S} \mu_{\pi'}$, an optimal policy $\pi$ with respect to $S$. We call a mixed policy deterministic if, for every $\pi_{X | C_X} \in \pi$, $X$ 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 $\mu_S^*$. Still, this result has a particular importance to the treatment provided here due to its implications to the d-separation criterion [49], which will be instrumental and discussed in depth in Sec. 4.1. 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 $\mu_{\pi} = \mu_S^*$. 


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. 5a. A mixed policy (Fig. 5b) includes X₂ listening to X₁, which enables systematic coordination between X₁ and X₂. The proposition implies that X₂ can rather listen to C (which is the context of X₁) directly (Fig. 5c). Further, in Fig. 5d X₂ utilizes both X₁ and C. However, it is sufficient to make use of only C. By noting that the policy relative to Fig. 5c can achieve optimality, while relying on lesser information than the one relative to Fig. 5d, 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 $C'_X \subseteq C_X$, for every $(X, C'_X) \in S'$. Further, we denote by $\pi' \subseteq \pi$, where $\pi' \sim S'$ and $\pi \sim S$ if $\pi'(x|c'_X) = \sum c_x \pi(x|c_X)p_X(c'_X|c_X)$, for every $X \in X(S')$ where $C'_X = C_X \setminus C'_X$.

**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 $\pi \sim (S, M)$ 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—$\pi'(x|c'_X) = Q(x|c'_X)$ if $C'_X \neq C_X$ and $Q(x|c_x) = \pi'(x|c'_X) = \pi(x|c_x)$, 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) $(C \not\perp \! \! \! \perp Y | C_X \setminus \{C\})$ 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 $Q(y|x, c_X \setminus \{c\}) = Q_x(y|c_X \setminus \{c\})$, 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. 5) of a given MPS in Appendix E [28].

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. 5c where $X_2$ listens to $X_1$ as context. We showed that the dependence is vanished under the optimality (Fig. 5d). That is, the agent would better avoid learning $\pi(x_2|c, x_1)$ at the beginning, but optimize $\pi(x_2|c)$ instead.

**Definition 4 (Non-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., $\exists_{M \sim G} S' \subseteq S (\mu_{S'} > \mu_{S''})$.

---

1This condition was leveraged in the atomic interventions case to establish minimality [16] [17]; see also [20].
2The relevance of contextual information has been discussed in the influence diagrams literature [41] [20]. More recently, this condition was used in the case of singleton decisions (i.e., $|X(S)| = 1$), see [42] [43].
We will investigate a criterion more general than Thm. 1—weither, for a set of actions $X' \subseteq X^*$, a set of contexts $C' \subseteq C_{X^*} \backslash X'$ are relevant while taking account of deterministic relationships (Prop. 1). One approach is to characterize an opposite condition, i.e., $\mu_S^* = \mu_S^c$, for $S' \subseteq S$, as follows.

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

$$\mu_{S}^c = \sum_{y,c'} Q'_{X'}(y,c') \prod_{X \in X'} \pi'(x'(\cup c') \cap c_z),$$

then, $C_{X^*} \backslash (C' \cup X')$ are jointly redundant to $X'$ under optimality, and $S' \equiv (S \backslash X') \cup \{(X, C' \cap C_X)\}_{X \in X'}$ satisfies $\mu^*_{S} = \mu^c_{S}$.

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

To closely investigate a sufficient condition for Prop. 3, we start by discussing the implication of deterministic relationships, which characterizes an optimal policy, on the d-separation criterion. The graphical criterion handles deterministic mechanisms (i.e., *conditional intervention*) by excluding them appearing as common causes, e.g., $\leftarrow X \rightarrow$, in a trail [39]. This corresponds to adding those implied variables to the conditionals, in which we explicitly represent with an operation $\cdot$ for clarity.

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

Next, we illustrate contexts that unnecessarily induce correlations among actions without any implications on $Y$ (see Appendix E.1 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 $\mu_{\pi} = \mathbb{E}_{c_3} \mathbb{E}_{c_2} [y|c_3]$. Since there exists $c_3' = \arg \max_{c_3} \mathbb{E}_{c_3} \mathbb{E}_{c_2} [y|c_3]$, we can derive that $\mu_{\pi} \leq \mathbb{E}_{c_3} [y|c_3']$. Given that $c_3' \equiv c_1$ is merely a constant, new decision rules $\pi'(x_i|c_1) = Q(x_i|c_1, c_1') = \pi(x_i|c_1, c_1') \text{ for } i \in \{1, 2\}$ yield the same optimal reward. A more sophisticated example is shown in Fig. 5c where a redundant context can be fixed conditioned on the remaining contexts. The expected reward is expressed as

$$\mu_{\pi} = \sum_{c_1, c_2} Q(c_2|c_1) \left( \sum_{y,x} y P_x(y, c_1) \pi(x_1|c_1, c_2) \pi(x_2|c_2) \right) = \sum_{c_1, c_2} Q(c_2|c_1) \mu_{\pi}(c_1, c_2).$$

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 $c_1 \in X_{C_1}$. Then,

$$\leq \sum_{c_1} \mu_{\pi}(c_1, c_2'(c_1)) = \max_{c_1} \sum_{y,x} y P_x(y, c_1) \pi(x_1|c_1, c_2'(c_1)) \pi(x_2|c_2'),$$

By incorporating $c_2'$ into $\pi$, we can introduce $\pi'$ such that $\pi(x_1|c_1, c_2'(c_1))\pi(x_2|c_2') = \pi'(x_1|c_1)\pi'(x_2|c_1)$, satisfying Prop. 1. 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., $\pi(x_1|c_1, c_2') = \pi'(x_1|c_1)$.

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 $V_{\prec V}$ denote a subset of $V$ preceding $V$ in $G$ given an order $\prec$ over $V$. Figure 5: Causal graphs exemplifying redundancies of (a) $C_2 \rightarrow X_2$ by deterministic relationships, edges to $X_1$ and $X_2$ from (b) $C_3$, (c) $C_2$, (d) $X_3$ due to marginally or conditionally fixable contexts; (e) represents a maximal, non-redundant MPS under an optimal condition for Fig. 4.
Lemma 1. Given an MPS $S$, which satisfies non-redundancy (Thm. 4), 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 $V' = C' \cup X' \cup Z$ such that

1. $(Y \perp \pi_X \mid [X' \cup C'])_{G_S},$
2. $(C \perp \pi_{X \prec C'}, Z \prec C', U' \mid [(X' \cup C')_{G_S}])_{G_S}$ for every $C \in C'$, and
3. $V_{X \prec C'}$ is disjoint with $\text{det}(X)_{G_S}$ and subsumes $\text{pa}(X)_{G_S}$ for every $X \in X'$,

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

$$\mu_\pi = \sum_{y, c', x'} yQ'_{x'}(y, c') \sum_{u' \in X'} \prod_{Z \in Z} Q(z | v'_{< Z}, u') \prod_{X \in X'} \pi(x | c_x).$$

(3)

Lemma 1 offers a sufficient condition for obtaining the intermediate expression (Eq. 3) for us to rewrite $\mu_\pi$ as proposed in Prop. 6. The order $\prec$ dictates how the chain rule is applied in deriving the expression and what variables will appear as condition for the probability terms. The first two conditions are relevant to separate $Q'_{x'}(y, c')$ from the rest. The third one is to obtain $\pi(x | c_x)$ from $Q(x | v'_{< Z}, u')$. We revisit Fig. 8 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 $\leq = (C_3, C_1, X_2, C_2, X_1)$. We can derive the following expression for the expected reward (with subscripts concatenated),

$$\mu^*_S = \sum_{y, c, x' \in X} yQ'_{x'}(y, c) \sum_{c_{123}} Q(c_{123}, x)$$

$$= \sum_{y, c, x' \in X} yQ'_{x'}(y, c) \sum_{c_{123}} Q(c_3) Q(c_{1} \mid c_3) Q(x_{2} \mid c_{13}) Q(x_{1} \mid c_{123}, x_2)$$

$$= \sum_{c_3} Q(c_3) \sum_{y, c, x' \in X} yQ'_{x'}(y, c) \sum_{c_{12}} Q(c_{12} \mid x_1, c_2) \pi(x_2 | c_2) \pi(x_1 | c_1, c_2).$$

(6)

We now provide a sufficient condition that further polishes the intermediate expression from Lemma 1 so as to represent it as the expected reward for a smaller MPS than the original one, fulfilling the condition presented in Prop. 8.

Theorem 2. Let $U'$, $Z$, and $\prec$ satisfy Lemma 1. For $Z \in Z$, let $V_Z$ be a minimal subset of $V_{X \prec Z} \cup U'$ such that $Q(Z \mid V_Z) = Q(Z \mid V'_{< Z}, U')$. We define fix$(T)$ with respect to $\{(Z, V_Z)\}_{Z \in Z}$, that is, with $T = [T] \cup \{Z \in Z \mid V_Z \setminus U' \subseteq [T]\}$. fixed$(T)$ is $T$ if $T = T$ and fixed$(T)$, otherwise. If fixed$(C_X \setminus Z) \subseteq C_X$ for $X \in X'$, then, $S' = (S \setminus X') \cup \{(X, C_X \setminus Z) | X \in X', S' \subseteq S \}$ satisfies $\mu^*_{S'} = \mu^*_S$.

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

$$\mu^*_{S'} = \sum_{u' \in X'} \sum_{y, c', x'} yQ'_{x'}(y, c') \prod_{Z \in Z} Q(z | v_{< Z}, u') \prod_{X \in X'} \pi(x | c_x)$$

(7)

defines dependency
given to infer
marginally fixable
irrelevant to $Z$

We explain the theorem by deriving further from Eq. 10. $C_3$ can be fixed to a constant $c_3^*$ so that

$$\leq \sum_{y, c, x \in X} yQ'_{x}(y, c) \sum_{c_{12}} Q(c_{12} | x_1, c_2) \pi(x_2 | c_2) \pi(x_1 | c_1, c_2).$$

(8)

There exists $x_3^* \in X_{X_2}$ where we can replace $\pi(x_2 | c_3^*)$ with $\pi'(x_2)$ such that $\pi'(x_2^*) = 1$.

$$\leq \sum_{y, c, x \in X} yQ'_{x}(y, c) \sum_{c_{12}} Q(c_{12} | x_1, c_2) \pi'(x_2) \pi(x_1 | c_1, c_2).$$

(9)

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

$$= \sum_{c_{12}} Q(c_{12} | x_1, c_2) \pi(x_2 | c_1, c_2)$$

(10)

$$\leq \sum_{y, c, x \in X} yQ'_{x}(y, c) \pi'(x_2) \pi(x_1 | c_1, c_2^*)$$

(11)

$$= \sum_{y, c, x \in X} yQ'_{x}(y, c) \pi'(x_2) \pi(x_1 | c_1, c_2^*) = \mu^*_{S'}.\pi'(x_2) \pi(x_1 | c_1, c_2^*) = \mu^*_{S'}.$$
4 A Partial Order over Mixed Policies and Possible-Optimality

Equipped with the notion of non-redundancy under optimality (NRO, Def. 3), an agent can more efficiently optimize its policy than relying on generic non-redundancy (Def. 1). 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. 3a to 3d which represent various NRO MPSes. However, even without interacting with an environment, we can claim \( \mu \leq \mu_{S'} \leq \mu_{S''} \leq \mu_{S'''} \), that is, the next MPS is better than or equal to (simply better or improved hereinafter) the one regarding their optimal expected rewards in any model: First, \( \mu \leq \mu_{S'} \) since there exists an optimal \( X_1 \) value, \( x^*_1 \); Next, \( \mu_{S'} \leq \mu_{S''} \), there exists an optimal \( X_2 \) value, and can be determined without conditional on \( X_1 \), which is implied; Finally, \( \mu_{S'} \leq \mu_{S''} \) since \( X_1 \) can better behave by taking an effective context \( C \) into account. Therefore, the agent can only optimize parameters involving \( S'' \) (Fig. 3d) 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 \textit{possible-optimality} of MPS.

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

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 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. These two operations offer sufficient conditions for identifying non-POMPSes.

**Proposition 4.** Given an MPS \( S \) and \( X \in \mathcal{X}(S) \), adding \( C \in \mathcal{C}^* \setminus \{X\} \) as a context of \( X \), resulting \( S' = (S \setminus \{X\}) \cup \{(X, C_X \cup \{C\})\} \) improves \( S \) if \( C \notin \text{de}(X)_S \) and \( C \perp Y \mid (\mathcal{C}_X \cup \{C\}) \) in \( \text{H} \setminus \{X\} \).

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. 3 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 \( X \in \mathcal{X}^* \) 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 \( X \in \mathcal{X}^* \) is contextualizable (e.g., no UC), the problem becomes trivial (e.g., Markovian). Otherwise, we examine the existence of a back-door path. Let \( Q = P_x \) and \( H = \mathcal{G}_\pi \) for some \( \mathcal{S} \sim \pi \). Given \( X \in \mathcal{X}^* \setminus \mathcal{X}(\pi) \) and \( Z \subseteq \mathcal{C}^* \setminus \{X\} \), if (i) \((Y \perp \perp X)_{H_X} \) and (ii) \( X \notin \text{an}(Z)_H \), then

\[
\mu_{\pi'} = \sum_{y,x,z} yQ(y|x,z)Q(x|z)Q(z) = \sum_{y,x,z} yQ'_y(y|x)Q(x|z)Q(z) \quad \text{with} \quad yQ'_y(y|x)Q(x|z)Q(z) \quad \text{and} \quad yQ'_y(y,z)Q'(z) \quad \text{for some} \ \pi'.
\]

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 \( X_2 \)): \( C_2 \) cannot be used as \( Z \) since \( X_1 \in \text{an}(C_2)_G \); \( X_2 \leftrightarrow C_2 \rightarrow Y \) is an open back-door path. We propose a solution for adding interventions simultaneously, powered by Thm. 2.

---

16112 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.
Theorem 3. 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., $(X(S') \setminus X(S)) \cup \{X \in X(S) \mid C'_X \neq C_X\}$, and (invalid) MPS $S'' = \{\langle X, pu(X)_{\mathcal{H}'}, U_X \rangle \}_{X \in X'}$. $\mu_{S''} = \mu_{S'}$. Given an MPS $S$, an intermediate MPS is constructed adding new contexts to a subset of $X^*$ while assuming that any non-contextualizable variables can be used as contexts. Consider comparing $S = \emptyset$ (Fig. 7a) and $S'$ (Fig. 7b) where we employ an intermediate representation $S''$ (Fig. 7c) to ultimately inspect $\mu_{S'} \leq \mu_{S''}$. Thm. 2 is applicable with $U' = \emptyset$ and $\leq \langle C_1, U_{X_1}, U_{X_2}, X_1, C_2, X_2 \rangle$ 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) to obtaining POMPSes (or a superset of) is that contexts are interleaved in both terms in Eq. (4) 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 $C_1$ take place after the reward is evaluated, remarkably demonstrates the opposite. With $X_1$ intervened on, $C_1$ 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 intervene-worthy — if $de(X)_{G} \cap \{C^* \cup \{Y\}\} = \emptyset$, there exists no MPS that makes $X \in an(C^* \cup \{Y\})_G$, and, thus, $X$ can be excluded from $X^*$.

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

5 Conclusions

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

Our work investigates the efficiency and effectiveness of AI agents to explore the environments and ultimately achieve optimality. Our results provide a tool for AI engineers and researchers to identify where the inefficiency of a policy may be coming from, including potentially unintended side effects. Further, the characterization provided in this work can suggest how systemic improvements are possible given non-parametric causal understanding of the underlying systems. The very topic of our paper about efficiency and effectiveness has been studied for several decades in diverse fields: bandits, reinforcement learning, design of experiments, etc. Hence, it is not difficult to imagine that our work will share the common problems with other automated decision making tools and methods such as (i) the system optimized based on an ill-defined reward may harm ‘unknown unknowns’ (e.g., increasing the revenue of alcoholic beverage companies based on targeted advertising over recovering alcoholics if their health is not properly modeled) or (ii) the optimization can be impossible due to the participants of adversarial players (e.g., rewarding the number of software bugs fixed makes software engineers to create more bugs to fix, see Goodhart’s law). Mitigating the first kind of risks will require deploying proper countermeasure through, e.g., regulations by governments. The second kind of risks (errors or failures) can be detected through examining possible changes of underlying mechanisms (i.e., anomaly detection). However, the current work does not consider multiple adversarial participants (e.g., game-theoretic settings), which is a subject of future research.

**Acknowledgments and Disclosure of Funding**

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

**References**


