Online Reinforcement Learning for Mixed Policy Scopes

Junzhe Zhang  
CausalAI Lab  
Columbia University  
junzhez@cs.columbia.edu

Elias Bareinboim  
CausalAI Lab  
Columbia University  
eb@cs.columbia.edu

Abstract

Combination therapy refers to the use of multiple treatments – such as surgery, medication, and behavioral therapy - to cure a single disease, and has become a cornerstone for treating various conditions including cancer, HIV, and depression. All possible combinations of treatments lead to a collection of treatment regimens (i.e., policies) with mixed scopes, or what physicians could observe and which actions they should take depending on the context. In this paper, we investigate the online reinforcement learning setting for optimizing the policy space with mixed scopes. In particular, we develop novel online algorithms that achieve sublinear regret compared to an optimal agent deployed in the environment. The regret bound has a dependency on the maximal cardinality of the induced state-action space associated with mixed scopes. We further introduce a canonical representation for an arbitrary subset of interventional distributions given a causal diagram, which leads to a non-trivial, minimal representation of the model parameters.

1 Introduction

The problem of policy learning is concerned with choosing actions based on the state of the environment with the goal of optimizing a certain measure of performance. Existing policy learning methods could be generally categorized into online learning and offline learning. Online reinforcement learning (RL) learns optimal policies by conducting sequential experimentation, while repeatedly adjusting the policy that is currently deployed based on the observed outcomes up to a certain point in time. Effective online algorithms have been developed for various canonical environments, including multi-armed bandits [26, 13, 14], Markov decision processes (MDPs) [30, 11, 34], and partially observable MDP [9, 3]. On the other hand, off-policy learning focuses on the offline setting with the goal of evaluating the effectiveness of candidate policies from history data collected following a different behavior policy [24, 32, 33]. More generally, the discipline of causal inference (CI) offers focuses a compelling set of tools and a formal language for offline learning. It allows the agent to draw conclusions about new policies from a combination of observations and knowledge about the data-generating mechanisms. Several methods and graphical criteria have been proposed [23, 29, 4].

By and large, almost all methods described above concerns optimizing over a parametric space of policies with a fixed state-action space, which are called the policy scope. That is, the set of actions and observed states are pre-specified a priori. However, in many practical applications, this is possibly somewhat stringent, and the agent has to optimize over candidate policies with varying state-action spaces; these are called mixed scopes. For concreteness, consider the causal model in Fig. 1a that describes possible treatments for alcohol use disorder [19]. Based on the condition of alcohol dependant patients $Z$, the physician may prescribe a medication $X_1$ to maximize the total days of abstinence $Y$. An alternative treatment is to prescribe a behavioral therapy $X_2$, which alters the social environment $W$ of the patient, and, in turn, changes the pattern of alcohol use. The goal
We will investigate the online learning of optimal policies with mixed scopes, provided with a causal

diagram associated with the underlying structural causal model (SCM) [23, Ch. 7]. In particular, our contributions are summarized as follows. (1) We develop a novel online learning algorithm that identifies an optimal policy with mixed scopes in an unknown SCM, and show that it achieves a sublinear cumulative regret. (2) We introduce a novel parametrization for SCMs with finite latent variables, which can represent an arbitrary subset of interventional distributions through a minimal collection of c-components. (3) Leveraging these results, we develop an alternative online algorithm that is more computationally efficient, while achieving the same asymptotic bound over the cumulative regret.

1.1 Preliminaries

In this section, we introduce the basic notations and definitions used throughout the paper. We use capital letters to denote random variables (X), small letters for their values (x) and |X| for the domain of X. For an arbitrary set X, let |X| be its cardinality. We denote by P(X) a probability distribution over variables X. Similarly, P(Y | X) represents a set of conditional distributions P(Y | X = x) for all realizations x. We will consistently use P(x) as abbreviations for probabilities P(X = x); so does P(Y = y | X = x) = P(y | x). Finally, 1{Z = z} is an indicator function that returns 1 if event Z = z holds true; otherwise, it returns 0.
The basic semantical framework of our analysis rests on structural causal models (SCMs) [23]. An SCM \( M \) is a tuple \( \{ V, U, F, P(U) \} \), where \( V \) is a set of endogenous variables and \( U \) is a set of exogenous variables. \( F \) is a set of functions s.t. each \( f_v \in F \) decides values of an endogenous variable \( V \in V \) taking as argument a combination of other variables in the system. That is, \( V \leftarrow f_v(PA_V, U_V) \), \( PA_v \subseteq V \), \( U_V \subseteq U \). Exogenous variables \( U \in U \) are mutually independent, values of which are drawn from the exogenous distribution \( P(U) \). Naturally, \( M \) induces a joint distribution \( P(V) \) over endogenous variables \( V \), called the observational distribution.

A policy \( \pi \) over a subset \( X \subseteq V \) is a sequence of decision rules \( \{ \pi(X \mid S_X) \}_{X \in X} \), where every \( \pi(X \mid S_X) \) is a probability distribution mapping from domains of a set of covariates \( S_X \subseteq V \) to the domain of an action \( X \). An intervention following a policy \( \pi \) over variables \( X \), denoted by \( do(\pi) \), is an operation which sets values of every \( X \in X \) to be decided by policy \( X \sim \pi(X \mid S_X) \) \[27\], replacing the functions \( f_{X} = \{ f_{X} : \forall X \in X \} \) that would normally determine their values. For an SCM \( M \), let \( M_{\pi} \) be a submodel of \( M \) induced by intervention \( do(\pi) \). For a set \( Y \subseteq V \), the interventional distribution \( P(Y \mid do(\pi)) \) is defined as the distribution over \( Y \) in the submodel \( M_{\pi} \), i.e., \( P_{M_{\pi}}(Y \mid do(\pi)) = P_{M_{\pi}}(Y) \). Subscript \( M \) is left implicit when it is obvious from the context.

Each SCM \( M \) is also associated with a causal diagram \( G \) (e.g., Fig. 1a), which is a directed acyclic graph (DAG) where solid nodes represent endogenous variables \( V \), empty nodes represent exogenous variables \( U \), and arrows represent the arguments \( PA_V, U_V \) of each structural function \( f_v \). We will use standard graph-theoretic family abbreviations to represent graphical relationships such as parents, children, descendants, and ancestors. For example, the set of parent nodes of \( X \) in \( G \) is denoted by \( pa(X) = \cup_{S \in X} pa(X) \); \( ch, de \) and \( an \) are similarly defined. Capitalized versions \( P_{a}, Ch, De, An \) include the argument as well, e.g. \( Pa(X)_{\hat{G}} = pa(X) \cup X \). For a subset \( X \subseteq V \), \( \hat{G}[X] \) is a vertex-induced subgraph from \( G \), which contains nodes \( X \) and edges among them.

For convenience, we define a bidirected arrow \( V_i \leftrightarrow V_j \) between endogenous nodes \( V_i, V_j \in V \) as a sequence of arrows \( V_i \leftarrow U_k \rightarrow V_k \) where \( U_k \in U \) is an exogenous parent shared by \( V_i, V_j \). A bi-directed path is a consecutive sequence of bi-directed arrows. We will utilize a special type of clustering of nodes in the diagram \( G \), called the c-component \[38\]. Formally, a subset \( C \subseteq V \) is a c-component in a causal diagram \( G \) if any pair \( V_i, V_j \in C \) is connected by a bi-directed path in \( G \). For instance, there exist bidirected paths \( X_1 \leftrightarrow X_2 \leftrightarrow Y \) and \( Z \leftrightarrow W \) in the causal diagram \( G \) in Fig. 1a, \( G \) thus contains c-components \( \{ X_1, X_2, Y \}, \{ Z, W \} \). On the other hand, the diagram \( G_S \) in Fig. 1b contains c-components \( \{ X_1 \}, \{ X_2, Y \}, \{ Z, W \} \), since bi-directed arrows \( X_1 \leftrightarrow X_2 \) and \( Y \leftrightarrow X_1 \) are removed. For a detailed survey on SCMs, we refer readers to [23, Ch. 7].

## 2 Optimizing Mixed Policy Spaces

We are concerned with the decision-making setting where an agent interacts with an SCM to optimize a reward \( Y \). The agent could intervene on an arbitrary subset of actions \( X \), called intervenable variables. A policy scope \( S \) is a collections of pairs \( \{ X, S_X \} \) where \( X \in X \) and covariates \( S_X \subseteq V \subseteq \{ Y \} \). Let \( X(S) = \{ X \}_{(X, S_X) \in S} \) denote the subset of actions in scope \( S \); similarly, \( S(S) = \bigcup_{(X, S_X) \in S} S_X \). For instance, for a scope \( S_1 = \{ (X_1, \{ Z \}) \} \) in Fig. 1b, \( X(S_1) = \{ X_1 \} \) and \( S(S_1) = \{ Z \} \). Formally, a policy \( \pi \) associated with scope \( S \) is a sequence of decision rules \( \{ \pi(X \mid S_X) \}_{(X, S_X) \in S} \). The collection of such policies \( \pi \) define a policy space \( \Pi_S \). All policies in \( \Pi_S \) share the same scope \( S \); they intervene on all actions in \( X(S) \) based on observed values of covariates \( S(S) \). Policy space \( \Pi_S \) models a general class of decision rules and treatment regimens in many classic sequential decision making settings, including policies in Markov decision processes [23], strategy profiles in influence diagrams [12], and dynamic treatment regimes [20].

This paper investigates a more general space of policies that are not restricted to a single scope \( S \). A mixed policy scope \( \Pi \) is a combinatorial set of policy scopes over intervenable variables \( X \). Formally,

**Definition 1.** A mixed policy scope \( \Pi \) is a collection of policy scopes \( S \) such that \( X(S) \subseteq X \).

A policy space \( \Pi_S \) with regard to a mixed policy scope \( S \) is the set of all policies compatible with every scope in \( S \), i.e., \( \Pi_S = \bigcup_{S \in \Pi} \Pi_S \). Henceforth, we will consistently refer to \( \Pi_S \) as a mixed policy space [18]. Our goal is to learn an optimal policy in \( \Pi_S \) that maximize reward \( Y \) in an SCM \( M \), i.e.,

\[
\pi^* = \arg \max_{\pi \in \Pi_S} E_M[Y \mid do(\pi)]
\]  

(1)
The agent does not have access to the parametrization of the underlying SCM $M$. Instead, the agent can observe endogenous variables $V$, the mixed policy space $\Pi_S$, and the causal diagram $G$ representing qualitative knowledge encoded about $M$. Throughout this paper, we assume that endogenous variables $V$ are discrete and finite, while exogenous variables $U$ could take any (continuous) value. Distributions $P(V)$ and $P(V \mid \text{do}(\pi))$ are thus categorical probability measures.

2.1 A Causal Upper Confidence Bound Algorithm

A typical online algorithm evaluates the effectiveness of candidate policies $\pi \in \Pi_S$ by directly deploying policies in the actual environment (SCM $M$) through repeated rounds of interventions $t = 1, 2, \ldots, T$. For each round $t$, the algorithm selects a policy $\pi_t \in \Pi_S$, performs an intervention $\text{do}(\pi_t)$, and receives a subsequent reward $Y_t$. The cumulative regret for an online algorithm in an SCM $M$ after $T$ rounds of interventions is defined as $R(T, M) = \sum_{t=1}^{T} (\mathbb{E}_M[Y \mid \text{do}(\pi^*)] - Y_t)$. In words, $R(T, M)$ measures the cumulative loss incurred since the online algorithm does not always select the optimal policy $\pi^*$. A reasonable, or desirable property for an online algorithm is to have a sublinear regret for any SCM $M$, i.e., $\lim_{T \to \infty} R(T, M)/T = 0$. If the regret is sublinear, the agent is known to be selecting the optimal policy almost all of the time as $T$ goes to infinity.

We will introduce a novel online algorithm that optimizes the mixed policy space while achieving a sublinear regret. It follows the well-celebrated principle of optimism in the face of uncertainty. That is, for each episode, the agent picks a policy based on upper confidence bounds (UCB) $[1, 2, 39, 40, 41]$, which evaluate the effectiveness of candidate policies in the most optimistic models, compatible with past observations, that induces the maximal expected reward. One innovation of our approach is to leverage the invariances embedded in the causal structure $G$, which allows us to evaluate a candidate policy using samples induced by other policies with different scopes.

Formally, let $C(G)$ denote a c-collection containing all (maximal) c-components in a causal diagram $G$. Consequently, $C(G)$ forms a partition $\{C_1, \ldots, C_n\}$ over nodes in $G$ where there is not bi-directed arrow between every pair $C_i$ and $C_j$ whenever $i \neq j$. Consider again the causal diagram $G$ in Fig. 14, c-collection $C(G)$ contains c-components $C_1 = \{X_1, X_2, Y\}$, $C_2 = \{Z, W\}$, which forms a partition over nodes in $G$. For an arbitrary subset $C \subseteq V$, the c-factor $Q[C]$ is a function defined as $Q[C](v) = P(e \mid \text{do}(v \cup c))$ [33], where $\text{do}(v \cup c)$ is an atomic intervention setting values of $V \setminus C$ to a constant $v \setminus c$ [23]. For convenience, we often omit input $v$ and write $Q[C]$. Fix a policy scope $S \in S$. For any policy $\pi \in \Pi_S$, let set $Z = A\pi(Y|\hat{G}_S \setminus X(S))$. The interventional distribution $P(Y \mid \text{do}(\pi))$ could be decomposed over c-components in subgraph $G[Z]$ as follows:

$$P(y \mid \text{do}(\pi)) = \sum_z \prod_{x \in X(S)} \pi(x \mid s_x) \prod_{C \in C_S} Q[C], \tag{2}$$

where the c-collection $C_S = C(G[Z])$ contains all c-components in $G[Z]$. Let a c-collection $C = \cup_{S \in S} C_S$. Among quantities in the above equation, decision rules $\pi(x \mid s_x)$ are fixed. To evaluate effects $P(Y \mid \text{do}(\pi))$ of candidate policies in $\Pi_S$, it is thus sufficient to learn parameters of $Q[C]$ for components $C \in C$. Consider again the causal diagram $G$ in Fig. 14, the mixed policy scope $S$ contains elements $S_1 = \{(X_1, \{Z\})\}, S_2 = \{(X_2, \{Z\})\}$ and $S_{1,2} = \{(X_1, \{Z\}), (X_2, \{X_1, Z\})\}$. For a policy $\pi_1(x_1|z)$ characterized with scope $S_1 = \{(X_1, \{Z\})\}$, $P(Y \mid \text{do}(\pi_1))$ decomposes as:

$$P(y \mid \text{do}(\pi_1)) = \sum_{x_1, x_2, z, w} \pi(x_1 \mid z)Q[X_2, Y|Z, W]. \tag{3}$$

The c-collection $C_S = \{(X_2, Y), (Z, W)\}$. Similarly, enumerating c-components associated with every scope $S \in S$ gives a c-collection $C = \{(Z, W), (X_1, Y), (X_2, Y), \{Y\}\}$. For any $C \in C$, let $S(C)$ be a subset of scopes $S \in S$ such that $C \subseteq C_S$. Our online algorithm evaluates $Q[C]$ by directly performing intervention $\text{do}(\pi)$ in the actual SCM $M$.

**Lemma 1.** For a causal diagram $G$, let $C \subseteq C$ be a c-component in $G$. Then, c-factor $Q[C]$ factorizes over a topological ordering $\prec$ in $G$ as follows:

$$Q[C] = \prod_{V \in C} q(v \mid PA^+_V) \tag{6}$$

where extended parents $PA^+_V = Pa(C_V) \setminus \{V\}; C_V$ is the c-component containing $V$ in $G[V \setminus \{V' \prec V\}]$. Moreover, $q(V \mid PA^+_V) = \Pr(V \mid PA^+_V, \text{do}(\pi))$ for any policy $\pi \in \Pi_S(C)$. 

\footnote{Generally, a c-collection $C$ is a collection of c-components $C$ contain in a causal diagram $G$.}
As an example, consider again the causal diagram $\mathcal{G}$ and policy scopes $S_1, S_2, S_{1,2}$ described in Fig. 1. For a c-component $C_3 = \{X_2, Y\} \in \mathbb{S}_{S_1}$, this implies the policy scope $S_1 \in \mathbb{S}(C_3)$. Observe that $C_V = \{X_2, Y\}, C_{X_2} = \{X_2\}$ and extended parents $PA_C = \{X_2, Z, W\}, PA_{X_2} = \{Z\}$. Lem. 1 implies $Q[C_3]$ is computable from the sampling process following any policy $\pi_1 \in \Pi_{S_1}$, i.e.,

$$Q[X_2, Y] = P(y \mid x_2, z, w, do(\pi_1))P(x_2 \mid z, do(\pi_1)).$$

(7)

Details of our proposed algorithm, CAUSAL-UCB*, are provided in Alg. 1. It interacts with the underlying SCM through policies in $\Pi_{S_1}$ in repeated episodes $t = 1, \ldots, T$. At episode $t$, it computes an empirical estimate of c-factor $Q[C]$ for each c-component $C \in \mathcal{C}$. More specifically, the empirical estimate of c-factor $\hat{Q}_t[C]$ prior to episode $t$ is given by:

$$\hat{Q}_t[C](v) = \prod_{V \in C} \hat{q}_t(v \mid pa^+_V), \quad \text{where } \hat{q}_t(v \mid pa^+_V) = \frac{n_t(v, pa^+_V)}{\max \{n_t(pa^+_V), 1\}}. \quad (8)$$

Among quantities in the above equation, $n_t(v, pa^+_V)$ is the event count for observing $V = v, PA^+_C = pa^+_V$ after deploying policy $\pi \in \Pi_{S(C)}$ prior to episode $t$. That is, $n_t(v, pa^+_V) = \sum_{t'=1}^{t-1} \mathbb{1}_V = v, (PA^+_C), \pi \in \Pi_{S(C)}$ and $n_t(pa^+_V) = \sum_v n_t(v, pa^+_V)$. At Step 3, CAUSAL-UCB* maintains a confidence set $\mathcal{M}$ of all possible SCM compatible with $\mathcal{G}$ using convex intervals centered around each estimate $\hat{Q}_t[C]$. It then finds the optimal policy $\pi_t$ for the most optimistic instance $M_t$ from $\mathcal{M}$ that induces the maximal expected reward $\mathbb{E}[Y \mid do(\pi)]$ (Step 4).

Finally, Step 7 performs $do(\pi_t)$ throughout episode $t$ and collect new samples $V_t$. Let $V(C)$ be a union $\bigcup_{C \in \mathcal{C}} C$. It is possible to derive a regret bound of CAUSAL-UCB* after $T > 1$ episodes.

**Theorem 1.** For a causal diagram $\mathcal{G}$ and a mixed policy scope $\mathcal{S}$, fix a $\delta \in (0,1)$. With probability at least $1 - \delta$, it holds for any $T > 1$, the regret of CAUSAL-UCB* is bounded by

$$R(T, M) \leq 19\Delta(\mathcal{G}, \mathcal{S})\sqrt{|S|T\ln(\frac{1}{\delta})},$$

where function $\Delta(\mathcal{G}, \mathcal{S}) = \max_{S \in \mathcal{S}} \Delta(\mathcal{G}, S)$ and $\Delta(\mathcal{G}, S) = \sum_{V \in V(C_{S})} \sqrt{|\Omega_{V, U, PA_V}|}$.

Thm. 1 implies that CAUSAL-UCB* is able to achieve a sublinear regret; therefore, policy $\pi_t$ eventually converges to an optimal policy as episode $t \to \infty$. For example, in the causal diagram $\mathcal{G}$ and the mixed policy scope $\mathcal{S} = \{S_1, S_2, S_{1,2}\}$ described in of Fig. 1, applying Thm. 1 gives a regret bound $O\left(\Delta(\mathcal{G}, \mathcal{S})\sqrt{T\ln T}\right)$. Evaluating $\Delta(\mathcal{G}, \mathcal{S})$ with the cardinality $|\Omega_{X_2}| > |\Omega_{X_1}|$ gives

$$\Delta(\mathcal{G}, \mathcal{S}) = \Delta(\mathcal{G}, S_1) = \sqrt{|\Omega_{Y,X_2,Z,W}|} + \sqrt{|\Omega_{X_2,Z}|} + \sqrt{|\Omega_{W,Z,X_1}|} + \sqrt{|\Omega_{Z}|}. \quad (10)$$

In comparison, applying standard UCB with deterministic polices in $\Pi_{S}$ as arms leads to a regret bound $O\left(\sqrt{|\Pi|T \ln T}\right)$ where $|\Pi_{S}| = |\Omega_{X_1}|^{|\Omega_S|} + |\Omega_{X_2}|^{|\Omega_S|} + |\Omega_{X_1}|^{|\Omega_S|} \times |\Omega_{X_2}|^{|\Omega_S|}$, which is dominated by that of CAUSAL-UCB* as the number of endogenous states increases.
Accelerate Learning Processes

In the regret bound of Thm. [1], the function $\Delta(G, S)$ can be seen as a measure evaluating the difficulty of the learning task. It has a dependency on the maximal cardinality of the state-action space associated with policy scopes $S$ in $\mathcal{S}$. There exist a general characterization of the mixed scope with respect to properties that allow the agent to detect redundant and suboptimal policy scopes [13]. Doing so, we obtain a refinement of the agent’s candidate policy scopes so that it converges to the optimal strategy faster and more robustly. For instance, consider the causal diagram $G$ described in Fig. [2]. We are interested in evaluating policies with mixed scope $S = \{S_1, S_2, S_{1,2}\}$ where $S_1 = \{(X_1, \emptyset)\}$, $S_2 = \{(X_2, \emptyset)\}$ and $S_{1,2} = \{(X_1, \emptyset), (X_2, \{X_1\})\}$. It is possible to show that $S_{1,2}$ is redundant since there always exists an optimal policy with scope $S_1$ or $S_2$. Removing $S_{1,2}$ leads to a more refined mixed scope $S^* = \{S_1, S_2\}$. Since $\Delta(G, S^*) = \max\{\Delta(G, S_1), \Delta(G, S_2)\} \leq \max\{\Delta(G, S_1), \Delta(G, S_1), \Delta(G, S_{1,2})\} = \Delta(G, S)$, applying CAUSAL-UCB* with the refined scope $S^*$ achieves a smaller regret than that with $S$, thus accelerating the online learning process.

3 Optimistic Planning over Mixed Policy Scopes

The optimization problem entailed by CAUSAL-UCB* in the algorithm discussed earlier (Eq. [5]) requires the learner to search over all possible SCMs compatible with the causal diagram and experimental data. In principle, this entails a challenging task since one does not have access to the parametric forms of the underlying structural functions $\mathcal{F}$ nor the exogenous distribution $P(U)$. Consequently, it is not clear how the existing optimization procedures can be used.

When domains of endogenous variables are discrete and finite, there exists a canonical family of SCMs that parameterizes all observational and interventional distributions in any causal diagram using a finite number of exogenous states [43, 22]. A c-component $C \in \mathcal{C}$ is said to cover an exogenous variable $U \in U$ if $U \in \cup_{V \in C} U_V$. Let $\mathcal{C}(U)$ denote the set of c-components in $\mathcal{C}$ covering $U$. Consider again the causal diagram in Fig. [1a] for the c-collection $\mathcal{C} = \{\{Z, W\}, \{X_1, Y\}, \{X_2, Y\}, \{Y\}\}$, exogenous node $U_2$ is covered by $\mathcal{C}(U_2) = \{\{X_1, Y\}, \{X_2, Y\}, \{Y\}\}$; similarly, $\mathcal{C}(U_1) = \{\{Z, W\}\}$. Next we present a novel finite-state decomposition representing c-factors in any c-collection $\mathcal{C}$.

Theorem 2. For any SCM $M = \langle \mathcal{V}, \mathcal{U}, \mathcal{F}, P(U) \rangle$, let $\mathcal{C}$ be an arbitrary c-collection. For any $\mathcal{C} \in \mathcal{C}$, c-factor $Q[C]$ decomposes as follows:

$$Q[C](v) = \sum_{U \in U} \sum_{u=1, \ldots, d_U} \prod_{V \in \mathcal{C}} \mathbb{I}\{f_V(p_{a_V}, u_V) = v\} \prod_{u \in U} P(u)$$

(11)

where for every exogenous $U \in U$, $P(U)$ is a discrete distribution over a finite domain $\{1, \ldots, d_U\}$ with cardinality $d_U = \sum_{C \in \mathcal{C}(U)} |\Omega_{Pa(C)}|$; $\mathcal{C}(U) \subseteq \mathcal{C}$ are c-components covering $U$.

Henceforth, we will refer to the family of SCMs with finite exogenous domain defined in Thm. 2 as C-canonical SCMs. For concreteness, consider the causal diagram in Fig. [1a] with $X_1, X_2, Y, Z, W \in \{0, 1\}$. Recall that for $\mathcal{C} = \{\{Z, W\}, \{X_1, Y\}, \{X_2, Y\}, \{Y\}\}$, $\mathcal{C}(U_1) = \{\{Z, W\}\}$ and $\mathcal{C}(U_2) = \{\{X_1, Y\}, \{X_2, Y\}, \{Y\}\}$. The cardinality of $U_1$ in C-canonical SCMs is $d_1 = |\Omega_{Pa(Z, W)}| = |\Omega_{Z, X_1, W}| = 8$. Similarly, the cardinality of $U_2$ in C-canonical SCMs is given by:

$$d_2 = |\Omega_{Pa(X_1, Y)}| + |\Omega_{Pa(X_2, Y)}| + |\Omega_{Pa(Y)}| = 2^5 + 2^4 + 2^4 = 64.$$  

(12)

Thm. 2 implies that $Q[Z, W], Q[X_2, Y]$ in the diagram of Fig. [1a] could be written as:

$$Q[Z, W] = \sum_{u_1 = 1, \ldots, d_1} \mathbb{I}\{f_Z(u_1) = z\} \mathbb{I}\{f_W(z, x_1, u_1) = w\} P(u_1)$$

(13)

$$Q[X_2, Y] = \sum_{u_2 = 1, \ldots, d_2} \mathbb{I}\{f_{X_2}(z, u_2) = x_2\} \mathbb{I}\{f_Y(x_2, z, w, u_2) = y\} P(u_2)$$

(14)

where $P(U_1), P(U_2)$ are distributions over finite domains $\{1, \ldots, d_1\}, \{1, \ldots, d_2\}$, respectively. Following the decomposition in Eq. (3), for any policy $\pi_1 \in \Pi_{S_1}$, the interventional distribution $P(Y | do(\pi_1))$ is a function of c-factors $Q[X_2, Y], Q[Z, W]$ and is given by:

$$P(y | do(\pi_1)) = \sum_{x_1, x_2, z, w} \pi(x_1 | z) \sum_{u_1 = 1, \ldots, d_1} \mathbb{I}\{f_Z(u_1) = z\} \mathbb{I}\{f_W(z, x_1, u_1)\} P(u_1)$$

$$\sum_{u_2 = 1, \ldots, d_2} \mathbb{I}\{f_{X_2}(z, u_2) = x_2\} \mathbb{I}\{f_Y(x_2, z, w, u_2) = y\} P(u_2)$$

(15)
Observe in Thm. 2 cardinalities of exogenous domains rely on the total number of c-components in \( C \). Next we introduce an effective method to reduce the model complexity of canonical SCMs by exploring identifiable relationships among c-factors \( Q[C] \) contained in \( C \).

**Definition 2.** For a causal diagram \( G \) and a c-collection \( C \), c-factor \( Q[C] \), \( \forall C \in C \), is identifiable if \( Q[C] \) is uniquely computable from other c-factors \( Q[C'] \) in \( C \setminus \{C\} \). That is, \( Q_{M_1}[C] = Q_{M_2}[C] \) for every pair of SCMs \( M_1, M_2 \) with \( G_{M_1} = G_{M_2} = G \), and \( Q_{M_1}[C'] = Q_{M_2}[C'] \) for any \( C' \in C \setminus \{C\} \).

In words, \( Q[C] \) is identifiable w.r.t. \( \langle G, C \rangle \) if it could be written as a function of c-factors in the remainder of the collection \( C' = C \setminus \{C\} \). For instance, consider the causal diagram of Fig. 1a and the c-collection \( C = \{\{Z, W\}, \{X_1, Y\}, \{X_2, Y\}, \{Y\}\} \). \[ \text{[36, Lem. 10]} \] implies that c-factor \( Q[Y] \) is identifiable from \( Q[X_1, Y] \) and is given by \( Q[Y] = \sum_{i} Q[X_1, Y_i] \). For every pair of SCMs that generates the same set of c-factors in \( C' = \{\{Z, W\}, \{X_1, Y\}, \{X_2, Y\}\} \), they must also coincide in parameters of \( Q[Y] \). It is thus sufficient to focus on representing parameters of c-factors in the subset \( C' \), which we call a reduction of the original c-collection \( C \).

**Definition 3.** For a causal diagram \( G \) and a c-collection \( C \), a subset \( C' \subseteq C \) is said to be a reduction of \( C \) if it is obtained by successive removals of identifiable c-component \( C \).

A reduction \( C' \) is minimal if there exists no proper subset \( C'' \subset C' \) such that \( C'' \) is a reduction of \( C' \). To reduce the model complexity of candidate SCMs, it is always preferable to present c-components in a minimal reduction. Next we describe a systematic procedure to obtain a minimal reduction of c-collection \( C \) in a causal diagram \( G \). Details of our algorithm, \textsc{MINCOLLECT}, are described in Alg. 2. It repeatedly removes identifiable c-factors \( Q[C] \) from collection \( C \). The subroutine \textsc{Identify} \[ \text{[36]} \] is a complete algorithm in determining the identifiability of \( Q[C] \) from another c-factor \( Q[C'] \) such that \( C \subseteq C' \). In particular, \textsc{Identify} returns a formula estimating \( Q[C] \) from \( Q[C'] \) if the target query is identifiable; otherwise, it returns “Fail”. It is possible to show that the greedy procedure in Alg. 2 always returns a minimal reduction of \( C \).

**Proposition 1.** For a causal diagram \( G \) and a c-collection \( C \), \textsc{MINCOLLECT}(\( G, C \)) returns a minimal reduction \( C^* \) of \( C \).

A natural question arising at this point is whether the order of removing c-components \( C \) could affect the output of \textsc{MINCOLLECT}. Fortunately, the next proposition shows that this is not the case.

**Proposition 2.** For a causal diagram \( G \), any c-collection \( C \) has a unique minimal reduction.

Consider again the c-collection \( C = \{\{Z, W\}, \{X_1, Y\}, \{X_2, Y\}, \{Y\}\} \) in the causal diagram \( G \) of Fig. 1a. Applying procedure \textsc{MINCOLLECT}(\( G, C \)) gives a minimal reduction \( C^* = \{\{Z, W\}, \{X_1, Y\}, \{X_2, Y\}\} \). Therefore, it is sufficient to consider \( C^*-\)canonical SCMs compatible with \( G \) when computing the optimistic policy in Eq. (5). Compared with \( C \)-canonical SCMs, the cardinality of \( U_1 \) in a \( C^*-\)canonical SCM remains the same and equates to \( d_1 = |\Omega_{Pa(Z, W)}| = 8 \). On the other hand, \( C^*(U_2) = \{\{X_1, Y\}, \{X_2, Y\}\} \) and the cardinality of \( U_2 \) is given by

\[
d_2 = |\Omega_{Pa(X_1, Y)}| + |\Omega_{Pa(X_2, Y)}| = 8^2 + 8^2 = 48,
\]

which is smaller than the cardinality of \( U_2 \) in \( C \)-canonical SCMs given by Eq. (12). One could then obtain an optimistic policy in Eq. (5) by solving a series of equivalent polynomial programs. For a more detailed survey on canonical SCMs and related work, we refer readers to the complete technical report \[ \text{[42, Appendix C]} \].

### 3.1 Thompson Sampling

The canonical representation of c-factors allows the optimization problem in Eq. (5) to be reducible to a series of equivalent polynomial programs. Nevertheless, solving polynomial optimization is NP-hard in general \[ \text{[8]} \], which means that applying \textsc{CAUSAL-UCB*} is still computationally challenging. This section introduces an alternative online algorithm that is computationally feasible, while achieving a similar asymptotic bound on the cumulative regret.
Our algorithm is based on the heuristics of Thompson sampling (TS) [35,31,21]. It maintains a prior distribution \(\rho(M)\) over all possible SCMs \(M\) compatible with the causal diagram \(\mathcal{G}\). One may surmise that it is challenging to define such a prior \(\rho\) since the agent does not have access to the parametric forms of the underlying SCM. Fortunately, it follows from the decomposition in Thm. 2 that we could assume the exogenous domain to be discrete and finite without loss of generality. Particularly, we will consider a family of \(\mathcal{C}^*\)-canonical SCMs where \(\mathcal{C}^*\) is a minimal reduction of \(c\)-collection \(\mathcal{C}\) in \(\mathcal{G}\). The cardinality of domain of every \(U \in \mathcal{U}\) is given by \(d_U = \sum_{C \in \mathcal{C}^*(U)} |\Omega_{Pa(C)}|\). We assume that probabilities of \(P(U)\) are drawn from uninformative Dirichlet priors; and \(\mathcal{F}\) are drawn uniformly from the finite class of possible structural functions. That is, for every \(U \in \mathcal{U}\) and every \(V \in \mathcal{V}\),

\[
P(U) \sim \text{Dir}(\alpha_1, \ldots, \alpha_{d_U}), \quad f_V \sim \text{Unif}(\Omega_{PA_V} \times \Omega_{UV} \rightarrow \Omega_V),
\]

where \(\alpha_1 = \cdots = \alpha_{d_U} = 1\); \(\Omega_{PA_V} \times \Omega_{UV} \rightarrow \Omega_V\) is a class containing all functions mapping from finite domains of \(PA_V, U_V\) to \(V\), which must also contain only a finite number of elements.

Details of our algorithm, CAUSAL-TS*, are described in Alg. 3. At each episode \(t\), it updates the posterior distribution \(\rho(M|\mathcal{V}_t)\) from collected samples \(\mathcal{V}_t = \{\mathcal{V}_1, \ldots, \mathcal{V}_{t-1}\}\) prior to episode \(t\), and draws an estimate \(M_t\) from the updated posteriors. Similar to Bayesian causal inference approaches [7,43], we will obtain a posterior sample \(M \sim \rho(M | \mathcal{V}_t)\) using Gibbs sampling [10]. In Step 4, CAUSAL-TS* computes an optimistic policy \(\pi_t\) that maximizes the expected outcome \(\mathbb{E}[Y \mid \text{do}(\pi)]\) induced by the sampled SCM \(M_t\). Finally, the agent executes \(\pi_t\) throughout episode \(t\) and new samples \(\mathcal{V}_t\) are collected.

Following previous work [28,21,22], we will assess the performance of TS using the Bayesian cumulative regret up to episode \(T\), i.e., \(R(T, \rho) = \mathbb{E}[R(T, M) \mid M \sim \rho(M)]\), where the expectation is taken with respect to the prior distribution over \(M\). There exists a general relationship between TS and UCB algorithms in many model classes [28]. This allows one to convert regret bounds developed for CAUSAL-UCB* into Bayesian regret bounds for CAUSAL-TS*. Formally,

**Theorem 3.** Given a causal diagram \(\mathcal{G}\), a mixed policy scope \(\mathbb{S}\), and a prior distribution \(\rho\), it holds for any \(T > 1\), the regret of CAUSAL-TS* is bounded by

\[
R(T, \rho) \leq 26\Delta(\mathcal{G}, \mathbb{S})\sqrt{|\mathbb{S}| T \ln \left(\frac{\left|\mathbb{V}(\mathcal{C})\right|T}{\rho}\right)}.
\]

Compared with Thm. 1, the above regret bound implies that CAUSAL-TS* achieves a similar asymptotic performance as CAUSAL-UCB*. Particularly, Alg. 3 requires one only has to find an optimal policy \(\pi_t\) in a specific SCM \(M_t\), while in Alg. 1 parameters of \(M_t\) are imprecise, bounded in the hypothesis class \(\mathcal{M}_t\). There exist effective planning algorithms in finding optimal policies in a structured environment provided with detailed parameterization of the underlying SCM [15,12]. This implies that CAUSAL-TS* is more computationally feasible compared to CAUSAL-UCB*.

### 4 Simulations

In this section, we evaluate the performance of our algorithms on randomly generated SCMs in various types of causal diagrams. After all, our algorithms can consistently find the corresponding optimal policies with mixed scopes. Further leveraging causal relationships in the underlying environment accelerates the convergence rate of online learners. In all experiments, we evaluate the novel CAUSAL-TS*, with uninformative Dirichlet priors over exogenous probabilities and uniform priors over structural functions, which we label as c-ts*. As a baseline, we also include randomized trials (rct) allocating treatments in all possible scopes uniformly at random, standard Thompson sampling algorithm (ts) using all deterministic policies as arms, and Thompson sampling over a simplified mixed scope (ts*), which is obtained by applying graphical conditions in [18]. For all algorithms, we measure their cumulative regrets over \(T = 1.1 \times 10^3\) episodes. We refer readers to the technical report [42, Appendix B] for a more detailed discussion on the experimental set-up.
This paper investigated the online reinforcement learning for selecting an optimal treatment regimen from a policy space characterized with mixed state-action scopes. We first presented an online algorithm (Alg. 1) that achieves a sublinear regret but is computationally intractable for any moderate-size instance. We further introduced a novel type of parametrizations for general SCMs (Thm. 2), with finite observed and latent domains, that can represent an arbitrary subset of interventional distributions in a causal diagram using the minimal number of decomposing factors, called c-components. We then developed a more computationally efficient online algorithm (Alg. 3), based on the heuristics of Thompson sampling, that identifies an optimal policy with the same sample complexity.

The performance of these algorithms on random SCMs in Fig. 3 reveals that c-ts* consistently outperforms ts and ts* since it is able to reuse samples induced by policies with mixed scopes by exploiting underlying causal knowledge. The performance of ts and ts* coincides since none of scopes \( S_1, S_2, S_1,2 \) is consistently redundant or suboptimal. Unsurprisingly, rct performs the worst among all algorithms.

Simulation results, shown in Fig. 3c, reveal that the performance of ts and ts* coincide; the causal approach c-ts* consistently dominates ts and ts*; and finally, rct performs the worst among all strategies.

This paper investigated the online reinforcement learning for selecting an optimal treatment regimen from a policy space characterized with mixed state-action scopes. We first presented an online algorithm (Alg. 1) that achieves a sublinear regret but is computationally intractable for any moderate-size instance. We further introduced a novel type of parametrizations for general SCMs (Thm. 2), with finite observed and latent domains, that can represent an arbitrary subset of interventional distributions in a causal diagram using the minimal number of decomposing factors, called c-components. We then developed a more computationally efficient online algorithm (Alg. 3), based on the heuristics of Thompson sampling, that identifies an optimal policy with the same sample complexity.

In today’s healthcare, the growing use of combination therapies opens new opportunities in designing effective regimes by combining multiple treatments. The additional degrees of freedom present challenges in comparing different treatment regimes. We believe that our results constitute a significant step towards the development of a more principled science of personalized medicine.
Acknowledgments and Disclosure of Funding

This research was supported in part by the NSF, ONR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References


