TECHNICAL REPORT
R-57
June, 2020

Designing Optimal Dynamic Treatment Regimes: A Causal Reinforcement Learning Approach

Junzhe Zhang[1] **Elias Bareinboim**[1]

Abstract

A dynamic treatment regime (DTR) consists of a sequence of decision rules, one per stage of intervention, that dictates how to determine the treatment assignment to patients based on evolving treatments and covariates' history. These regimes are particularly effective for managing chronic disorders and is arguably one of the critical ingredients underlying more personalized decision-making systems. All reinforcement learning algorithms for finding the optimal DTR in online settings will suffer $\Omega(\sqrt{|\mathcal{D}_{\boldsymbol{X}\cup\boldsymbol{S}}|T})$ regret on some environments, where T is the number of experiments and $\mathcal{D}_{\boldsymbol{X}\cup\boldsymbol{S}}$ is the domains of the treatments $\boldsymbol{X}$ and covariates $\boldsymbol{S}$. This implies that $T = \Omega(|\mathcal{D}_{\boldsymbol{X}\cup\boldsymbol{S}}|)$ trials will be required to generate an optimal DTR. In many applications, the domains of $\boldsymbol{X}$ and $\boldsymbol{S}$ could be enormous, which means that the time required to ensure appropriate learning may be unattainable. We show that, if the *causal diagram* of the underlying environment is provided, one could achieve regret that is exponentially smaller than $\mathcal{D}_{\boldsymbol{X}\cup\boldsymbol{S}}$. In particular, we develop two online algorithms that satisfy such regret bounds by exploiting the causal structure underlying the DTR; one is the based on the principle of optimism in the face of uncertainty (`OFU-DTR`), and the other uses the posterior sampling learning (`PS-DTR`). Finally, we introduce efficient methods to accelerate these online learning procedures by leveraging the abundant, yet biased observational (non-experimental) data.

1. Introduction

In medical practice, a patient typically has to be treated at multiple stages; a physician sequentially assigns each treatment, repeatedly tailored to the patient's time-varying, dynamic state (e.g., infection's level, different diagnostic tests). Dynamic treatment regimes (DTRs, Murphy 2003) provide an attractive framework of personalized treatments in longitudinal settings. Operationally, a DTR consists of decision rules that dictate what treatment to provide at each stage, given the patient's evolving conditions and treatments' history. These decision rules are alternatively known as adaptive treatment strategies (Lavori & Dawson, 2000; 2008; Murphy, 2005a; Thall et al., 2000; 2002) or treatment policies (Lunceford et al., 2002; Wahed & Tsiatis, 2004; 2006).

Learning the optimal dynamic treatment regime concerns with finding a sequence of decision rules $\sigma_{\boldsymbol{X}}$ over a *finite* set of treatments $\boldsymbol{X}$ that maximizes a *primary outcome* Y. The main challenge is that since the underlying system dynamics are often unknown, it's not immediate how to infer the consequences of executing the policy $do(\sigma_{\boldsymbol{X}})$, i.e., the causal effect $E_{\sigma_{\boldsymbol{X}}}[Y]$. Most of the current work in the causal inference literature focus on the off-policy (offline) learning setting, where one tries to identify the causal effect from the combination of static data and qualitative assumptions about the data-generating mechanisms. Several criteria and algorithms have been developed (Pearl, 2000; Spirtes et al., 2001; Bareinboim & Pearl, 2016). For instance, a criterion called the *sequential backdoor* (Pearl & Robins, 1995) allows one to determine whether causal effects can be obtained by adjustment. This condition is also referred to as *sequential ignorability* (Rubin, 1978; Murphy, 2003). To ensure it, one could randomly assign values of treatments at each stage of the intervention and observe the subsequent outcomes; a popular strategy of this kind is known as the *sequential multiple assignment randomized trail* (SMART, Murphy 2005a). Whenever the backdoor condition can be ascertained, a number of efficient off-policy estimation procedures exist, including popular methods based on the propensity score (Rosenbaum & Rubin, 1983), inverse probability of treatment weighting (Murphy et al., 2001; Robins et al., 2008), and Q-learning (Murphy, 2005b).

More recently, (Zhang & Bareinboim, 2019) introduced

[1] Department of Computer Science, Columbia University, New York, USA. . Correspondence to: Junzhe Zhang <junzhez@cs.columbia.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

the first online reinforcement learning (RL, Sutton & Barto 1998) algorithm for finding the optimal DTR. Compared with the off-policy learning, an online learning algorithm learns through sequential, adaptive experimentation. It repeatedly adjusts the current decision rules based on the past outcomes; the updated decision rules are deployed to generate new observations. The goal is to identify the optimal treatment regime with low regret, i.e., the least amount of experimentation. Settings that allow some amount of online experimentation are increasingly popular, including, for instance, mobile and internet applications where continuous monitoring and just-in-time intervention are largely available (Chakraborty & Moodie, 2013)). For DTRs with treatments $\boldsymbol{X}$ and covariates' history $\boldsymbol{S}$, the strongest results of this kind establish $\tilde{\mathcal{O}}(\sqrt{|\mathcal{D}_{\boldsymbol{X} \cup \boldsymbol{S}}| T})$[1] for a particular algorithm introduced in (Zhang & Bareinboim, 2019), which is close to the lower bound $\Omega(\sqrt{|\mathcal{D}_{\boldsymbol{X} \cup \boldsymbol{S}}| T})$. However, when the cardinality of $\mathcal{D}_{\boldsymbol{X} \cup \boldsymbol{S}}$ is huge, even this level of regret (to guarantee appropriate learning) is somewhat unattainable in some critical settings, which suggests the need for investigating alternative and reasonable assumptions.

In many applications, one often has access to some causal knowledge about the underlying environment, represented in the form of *directed acyclic causal diagrams* (Pearl, 2000). When the causal diagram is sparse, e.g., some variables in $\boldsymbol{S}$ are affected by a small subset of treatments $\boldsymbol{X}$, the dimensionality of the learning problem could be reduced exponentially. There are RL algorithms exploiting the structural information in Markov decision processes (MDPs), where a finite state is statistically sufficient to summarize the treatments and covariates' history (Kearns & Koller, 1999; Osband & Van Roy, 2014). Unfortunately, the underlying environment of DTRs is often non-Markovian, and involves non-trivial causal relationships. For instance, in a treatment regime where patients receive multiple courses of chemotherapy, the initial treatment could affect the final remission via some unknown mechanisms, which are not summarizable by a prespecified state (Wang et al., 2012).

In this paper, we study the online learning of optimal dynamic treatment regimes provided with the causal diagram of the underlying, unknown environment. More specifically, our contributions are as follows. (1) We propose an efficient procedure (Alg. 1) reducing the dimensionality of candidate policy space by exploiting the functional and independence restrictions encoded in the causal diagram. (2) We develope two novel online reinforcement learning algorithms (Algs. 2 and 3) for identifying the optimal DTR, leveraging the causal diagram, and that consistently dominate the state-of-art methods in terms of the performance. (3) We introduce systematic methods to accelerate the proposed algorithms by extrapolating knowledge from the abundant, yet biased observational (non-experimental) data (Thms. 6 and 7). Our results are validated on multi-stage treatments regimes for lung cancer and dyspnoea. Given the space constraints, all proofs are provided in (Zhang & Bareinboim, 2020, Appendices A-C).

1.1. Preliminaries

In this section, we introduce the basic notations and definitions used throughout the paper. We use capital letters to denote variables (X) and small letters for their values (x). Let $\mathcal{D}_X$ represent the domain of X and $|\mathcal{D}_X|$ its dimension. We consistently use the abbreviation $P(x)$ to represent the probabilities $P(X = x)$. $\boldsymbol{X}^{(i)}$ stands for a sequence $\{X_1, \dots, X_i\}$ ($\emptyset$ if $i < 1$). Finally, $I_{\{\boldsymbol{Z}=\boldsymbol{z}\}}$ is an indicator function that returns 1 if $\boldsymbol{Z} = \boldsymbol{z}$ holds true; otherwise 0.

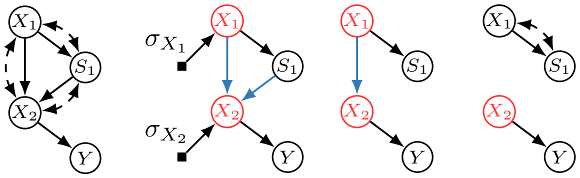
The basic semantical framework of our analysis rest on *structural causal models* (SCMs) (Pearl, 2000, Ch. 7). A SCM M is a tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, P(\boldsymbol{u}) \rangle$ where $\boldsymbol{V}$ is a set of endogenous (often observed) variables and $\boldsymbol{U}$ is a set of exogenous (unobserved) variables. $\mathcal{F}$ is a set of structural functions where $f_V \in \mathcal{F}$ decides values of an endogenous variable $V \in \boldsymbol{V}$ taking as argument a combination of other variables. That is, $V \leftarrow f_V(Pa_V, U_V), Pa_V \subseteq \boldsymbol{V}, U_V \subseteq \boldsymbol{U}$. Values of $\boldsymbol{U}$ are drawn from a distribution $P(\boldsymbol{u})$, which induces an observational distribution $P(\boldsymbol{v})$ over $\boldsymbol{V}$. An intervention on a subset $\boldsymbol{X} \subseteq \boldsymbol{V}$, denoted by $do(\boldsymbol{x})$, is an operation where values of $\boldsymbol{X}$ are set to constants $\boldsymbol{x}$, regardless of how they were ordinarily determined through the functions $\{f_X : \forall X \in \boldsymbol{X}\}$. For a SCM M, let $M_{\boldsymbol{x}}$ be a submodel of M induced by $do(\boldsymbol{x})$. The interventional distribution $P_{\boldsymbol{x}}(\boldsymbol{s})$ is the distribution over $\boldsymbol{S} \subseteq \boldsymbol{V}$ in submodel $M_{\boldsymbol{x}}$.

Each SCM M is associated with a directed acyclic graph (DAG) $\mathcal{G}$ (e.g., see Fig. 1a), called the causal diagram, where nodes correspond to endogenous variables $\boldsymbol{V}$, solid arrows represent arguments of each function f_V. A bi-directed arrow between nodes V_i and V_j indicates an unobserved confounder (UC) affecting both V_i and V_j, i.e., $U_{V_i} \cap U_{V_j} \neq \emptyset$. We will use the graph-theoretic family abbreviations, e.g. $An(\boldsymbol{X})_{\mathcal{G}}, De(\boldsymbol{X})_{\mathcal{G}}, Pa(\boldsymbol{X})_{\mathcal{G}}$ stand for the set of ancestors, descendants and parents of $\boldsymbol{X}$ in $\mathcal{G}$ (including $\boldsymbol{X}$). We omit the subscript $\mathcal{G}$ when it is obvious. A path from a node X to a node Y in $\mathcal{G}$ is a sequence of edges which does not include a particular node more than once. Two sets of nodes $\boldsymbol{X}, \boldsymbol{Y}$ are said to be d-separated by a third set $\boldsymbol{Z}$ in a DAG $\mathcal{G}$, denoted by $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z})_{\mathcal{G}}$, if every edge path from nodes in one set to nodes in another are "blocked". The criterion of blockage follows (Pearl, 2000, Def. 1.2.3).

In a causal diagram $\mathcal{G}$, variables $\boldsymbol{V}$ could be partitioned into disjoint groups, called *confounded components* (c-component), by assigning two variables to the same group if and only if they are connected by a path composed solely of bi-directed arrows (Tian & Pearl, 2002). The latent pro-

[1] $f = \tilde{\mathcal{O}}(g)$ if and only if $\exists k$ such that $f = \mathcal{O}(g \log^k(g))$.

jection $\texttt{Proj}(\mathcal{G}, \boldsymbol{S})$ is an algorithm that induces a causal diagram from $\mathcal{G}$ over a subset $\boldsymbol{S} \subseteq \boldsymbol{V}$ while preserving topological relationships among $\boldsymbol{S}$ (Tian, 2002, Def. 5). For example, in Fig. 1a, $\texttt{Proj}(\mathcal{G}, \{X_2, Y\})$ returns a subgraph $X_2 \to Y$; X_1, S_1, X_2 belong to the same c-component due to the bi-directed path $X_1 \leftrightarrow S_1 \leftrightarrow X_2$.



(a) $\mathcal{G}$ (b) $\mathcal{G}_{\sigma_{X_1,X_2}}$ (c) $\mathcal{G}_{\tilde{\sigma}_{X_1,X_2}}$ (d) $\mathcal{G}_{\sigma_{X_2}}$

Figure 1: (a) A causal diagram $\mathcal{G}$; (b) a manipulated diagram $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ with a policy space $\Pi = \{\mathfrak{D}_\emptyset \mapsto \mathfrak{D}_{X_1}, \mathfrak{D}_{\{S_1,X_1\}} \mapsto \mathfrak{D}_{X_2}\}$; (c) a diagram $\mathcal{G}_{\tilde{\sigma}_{X_1,X_2}}$ with a reduction $\tilde{\Pi} = \{\mathfrak{D}_\emptyset \mapsto \mathfrak{D}_{X_1}, \mathfrak{D}_{X_1} \mapsto \mathfrak{D}_{X_2}\}$; (c) a manipulated diagram $\mathcal{G}_{\sigma_{X_2}}$ with the minimal reduction $\Pi_{\text{MIN}} = \{\mathfrak{D}_\emptyset \mapsto \mathfrak{D}_{X_2}\}$.

2. Optimal Dynamic Treatment Regimes

We start the section by formalizing DTRs in the semantics of SCMs. We consider the sequential decision-making problem in a SCM $M^* = \langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, P(\boldsymbol{u}) \rangle$, where an agent (e.g., a physician) determines the values of a set of treatments $\boldsymbol{X} \subseteq \boldsymbol{V}$ with the goal of maximizing a primary outcome $Y \in \boldsymbol{V}$. Domains of $\boldsymbol{V}$ are discrete and finite.

A *dynamic treatment regime* (hereafter, policy) $\sigma_{\boldsymbol{X}}$ is a sequence of decision rules $\{\sigma_X : \forall X \in \boldsymbol{X}\}$. Each σ_X is a mapping from the values of the treatments and covariates' history $H_X \subseteq \boldsymbol{V}$ to the domain of probability distributions over X, denoted by $\sigma_X(x|h_X)$; we write $H_{X^+} = H_X \cup X$. An intervention $do(\sigma_{\boldsymbol{X}})$ following a policy $\sigma_{\boldsymbol{X}}$ is an operation that determines values of each $X \in \boldsymbol{X}$ following the decision rule σ_X, regardless of its original function f_X. Let $M^*_{\sigma_{\boldsymbol{X}}}$ be the manipulated SCM of M^* induced by $do(\sigma_{\boldsymbol{X}})$. We define the interventional distribution $P_{\sigma_{\boldsymbol{X}}}(\boldsymbol{v})$ as the distribution over $\boldsymbol{V}$ in the manipulated model $M^*_{\sigma_{\boldsymbol{X}}}$,

$$P_{\sigma_{\boldsymbol{X}}}(\boldsymbol{v}) = \sum_{\boldsymbol{u}} P(\boldsymbol{u}) \prod_{V \notin \boldsymbol{X}} P(v|pa_V, \boldsymbol{u}_V) \prod_{X \in \boldsymbol{X}} \sigma_X(x|h_X).$$

The collection of all possible $\sigma_{\boldsymbol{X}}$ defines a *policy space* Π, which we denote by $\{\mathfrak{D}_{H_X} \mapsto \mathfrak{D}_X : \forall X \in \boldsymbol{X}\}$. We are in search of an optimal policy $\sigma^*_{\boldsymbol{X}}$ maximizing the expected outcome $E_{\sigma_{\boldsymbol{X}}}[Y]$, i.e., $\sigma^*_{\boldsymbol{X}} = \arg\max_{\sigma_{\boldsymbol{X}} \in \Pi} E_{\sigma_{\boldsymbol{X}}}[Y]$.

Let $\mathcal{G}$ denote the causal diagram associated with M^* and let $\mathcal{G}_{\overline{\boldsymbol{X}}}$ be a subgraph of $\mathcal{G}$ by removing incoming arrows to $\boldsymbol{X}$. We denote by $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ a manipulated diagram obtained from $\mathcal{G}$ and Π by adding arrows from nodes in H_X to X in the subgraph $\mathcal{G}_{\overline{\boldsymbol{X}}}$. For example, Fig. 1b shows a manipulated graph $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ where treatments are highlighted in red and input arrows in blue. We assume that $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ does not include cycles. A DTR agent decides treatments following a topological ordering $\prec$ in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$. It does not forget previous treatments or information it once had, i.e., for any $X_i \prec X_j$, $H_{X_i^+} \subseteq H_{X_j}$. Such a property, called *perfect recall* (Koller & Friedman, 2009, Def. 23.5), ensures the following independence relationships among decision rules.

Definition 1 (Solubility)**.** A policy space Π is *soluble* w.r.t. $\mathcal{G}$ and Y if there exists a topological ordering $\prec$ on $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ (called the soluble ordering) such that whenever $X_i \prec X_j$, $(Y \cap De(X_j) \perp\!\!\!\perp \sigma_{X_i} | H_{X_j^+})_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, where σ_{X_i} is a new parent node added to X_i.

For instance, the policy space Π described in Fig. 1b is soluble relative to $X_1 \prec S_2 \prec X_2 \prec Y$ since $(Y \perp\!\!\!\perp \sigma_{X_1} | \{X_1, S_2, X_2\})_{\mathcal{G}_{\sigma_{X_1,X_2}}}$. When Π is soluble and M^* is known, there exist efficient dynamic programming planners (Lauritzen & Nilsson, 2001) that solve for the optimal policy $\sigma^*_{\boldsymbol{X}}$. Throughout this paper, we assume the parameters of M^* are unknown. Only the causal diagram $\mathcal{G}$, the policy space Π, and the primary outcome Y are provided to the learner, which we summarize as a signature $[\![\mathcal{G}, \Pi, Y]\!]$.

2.1. Reducing the Policy Space

In this section, we simplify the complexity of the learning problem by determining and exploiting irrelevant treatments and information for the candidate policies. We begin by defining the equivalence relationships among policy spaces.

Definition 2. Given $[\![\mathcal{G}, \Pi, Y]\!]$, a policy space $\tilde{\Pi}$ is *equivalent* to Π, if for any SCM M conforming to $\mathcal{G}$, $\max_{\tilde{\sigma}_{\boldsymbol{X}} \in \tilde{\Pi}} E_{M_{\tilde{\sigma}_{\boldsymbol{X}}}}[Y] = \max_{\sigma_{\boldsymbol{X}} \in \Pi} E_{M_{\sigma_{\boldsymbol{X}}}}[Y]$.

In words, two policy spaces are equivalent if they induce the same optimal performance. It is thus sufficient to optimize over a policy space that is in the same equivalence class of Π. We will introduce graphical conditions that identify such an equivalence class. Among equivalent policy spaces, we consistently prefer ones with smaller cardinality $|\Pi|$.

Definition 3. Given $[\![\mathcal{G}, \Pi, Y]\!]$, treatments $\tilde{\boldsymbol{X}} \subseteq \boldsymbol{X}$ are *irrelevant* if $\tilde{\boldsymbol{X}} = \boldsymbol{X} \setminus (\boldsymbol{X} \cap An(Y))_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$.

Intuitively, treatments $\tilde{\boldsymbol{X}}$ are irrelevant if they has no causal (functional) effect on the primary outcome Y. Therefore, the agent could choose not to intervene on $\tilde{\boldsymbol{X}}$ without compromising its optimal performance. Let $\Pi \setminus \tilde{\boldsymbol{X}}$ denote a partial policy space obtained from Π by removing treatments $\tilde{\boldsymbol{X}}$, i.e., $\{\mathfrak{D}_{H_X} \mapsto \mathfrak{D}_X : \forall X \notin \boldsymbol{X}\}$. The following proposition confirms the intuition of irrelevant treatments.

Lemma 1. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, $\Pi \setminus \tilde{\boldsymbol{X}}$ is equivalent to Π if treatments $\tilde{\boldsymbol{X}}$ are irrelevant.*

We will also utilize the notion of irrelevant evidences introduced in (Lauritzen & Nilsson, 2001, Def. 8).

Definition 4. Given $[\![\mathcal{G}, \Pi, Y]\!]$, evidences $\tilde{\boldsymbol{S}} \subseteq H_X$ for $X \in \boldsymbol{X}$, denoted by $\tilde{\boldsymbol{S}} \mapsto X$, are *irrelevant* if $(Y \cap De(X) \perp\!\!\!\perp \tilde{\boldsymbol{S}} | H_{X^+} \setminus \tilde{\boldsymbol{S}})_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$.

Def. 4 states that evidences $\tilde{\boldsymbol{S}} \mapsto X$ have no value of information on the outcome Y if the remaining evidences are known. Let $\Pi \setminus \{\tilde{\boldsymbol{S}} \mapsto X\}$ denote a policy space obtained from Π by removing $\tilde{\boldsymbol{S}}$ from input space of σ_X, i.e, $\{\mathfrak{D}_{H_X \setminus \tilde{\boldsymbol{S}}} \mapsto \mathfrak{D}_X\} \cup (\Pi \setminus \{X\})$. Our next result corroborates the definition of irrelevant evidence.

Lemma 2. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, $\Pi \setminus \{\tilde{\boldsymbol{S}} \mapsto X\}$ is equivalent to Π if evidences $\tilde{\boldsymbol{S}} \mapsto X$ are irrelevant.*

Lems. 1 and 2 allow us to search through the equivalence class of Π with reduced cardinality.

Definition 5. Given $[\![\mathcal{G}, \Pi, Y]\!]$, a policy space $\tilde{\Pi}$ is a reduction of Π if it is obtainable from Π by successively removing irrelevant evidences or treatments.

Lemma 3. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, a reduction $\tilde{\Pi}$ of the policy space Π is soluble if Π is soluble.*

Lem. 3 shows that $\tilde{\Pi}$ satisfies some basic causal constraints of Π, i.e., the solubility is preserved under reduction. In general, computational and sample complexities of the learning problem depend on cardinalities of candidate policies. Naturally, we want to solve for the optimal policy in a function space that is reduced as much as possible.

Definition 6. Given $[\![\mathcal{G}, \Pi, Y]\!]$, a reduction Π_{MIN} of Π is *minimal* if it has no irrelevant evidence and treatment.

One simple algorithm for obtaining a minimal reduction Π_{MIN} is to remove irrelevant treatments and evidences iteratively from Π until no more reduction could be found. An obvious question is whether the ordering of removal affects the final output, i.e., there exist multiple minimal reductions. Fortunately, the following theorem implies the opposite.

Theorem 1. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, there exists a unique minimal reduction Π_{MIN} of the policy space Π.*

We describe in Alg. 1 the `Reduce` algorithm that efficiently finds the minimal reduction. More specifically, let $\prec$ be a soluble ordering in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$. `Reduce` examines the treatments in $\boldsymbol{X}$ following a reverse ordering regarding $\prec$. For each treatment X_i, it iteratively reduce the policy space by removing irrelevant evidences. Finally, it obtains the minimal reduction by removing all irrelevant treatments.

Theorem 2. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, `Reduce` returns the minimal reduction Π_{MIN} of a soluble policy space Π.*

As an example, we apply `Reduce` on the policy space Π described in Fig. 1b. Since $(Y \perp\!\!\!\perp S_1 | X_1, X_2)_{\mathcal{G}_{\sigma_{X_1,X_2}}}$, evidence $S_1 \mapsto X_2$ is irrelevant. Removing S_1 leads to a reduction $\tilde{\Pi} = \Pi \setminus \{S_1 \mapsto X_2\}$ described in Fig. 1c. Similarly, we could remove $X_1 \mapsto X_2$ since $(Y \perp\!\!\!\perp X_1 | X_2)_{G_{\tilde{\sigma}_{X_1,X_2}}}$. Treatment X_1 is now irrelevant since there exists no path from X_1 to Y. Removing X_1 gives the minimal reduction Π_{MIN} described in Fig. 1d. Suppose policies in Π are deterministic. The cardinality of Π is $|\mathfrak{D}_{X_1}||\mathfrak{D}_{\{X_1,X_2,S_2\}}|$; while $|\Pi_{\text{MIN}}|$ could be much smaller, equating to $|\mathfrak{D}_{X_2}|$.

Algorithm 1 `Reduce`

```
Input: Signature $[\![\mathcal{G}, \Pi, Y]\!]$.
Let $\prec$ be a soluble ordering in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ and let treatments in $\boldsymbol{X}$ be ordered by $X_1 \prec \cdots \prec X_n$.
for all $i = n, \dots, 1$ do
    for all irrelevant evidence $S \mapsto X_i$ in $\Pi$ do
        Let $\Pi = \Pi \setminus \{S \mapsto X_i\}$.
    end for
end for
Return $\Pi = \Pi \setminus \tilde{\boldsymbol{X}}$ where $\tilde{\boldsymbol{X}}$ are irrelevant treatments.
```

3. Online Learning Algorithms

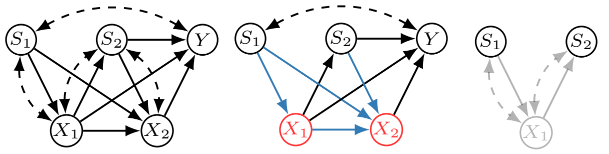
The goal of this section is to design online RL algorithms that find the optimal DTR $\sigma^*_{\boldsymbol{X}}$ in an unknown SCM M^* based solely on the information summarized in $[\![\mathcal{G}, \Pi, Y]\!]$.

An online learning algorithm learns the underlying system dynamics of M^* through repeated episodes of interactions $t = 1, \dots, T$. At each episode t, the agent picks a policy $\sigma^t_{\boldsymbol{X}}$, assigns treatments $do(\boldsymbol{X}^t)$ following $\sigma^t_{\boldsymbol{X}}$, and receives subsequent outcome Y^t. The cumulative regret up to episode T is defined as $R(T, M^*) = \sum_{t=1}^{T} (E_{\sigma^*_{\boldsymbol{X}}}[Y] - Y^t)$, i.e, the loss due to the fact that the algorithm does not always follow the optimal policy $\sigma^*_{\boldsymbol{X}}$. A desirable asymptotic property is to have $\lim_{T \to \infty} R(T, M^*)/T = 0$, meaning that the agent eventually converges and finds the optimal policy $\sigma^*_{\boldsymbol{X}}$. We also consider the Bayesian settings where the actual SCM M^* is sampled from a distribution ϕ^* over a set of candidate SCMs in $\mathfrak{M}$. The Bayesian regret up to episode T is defined as $R(T, \phi^*) = E[R(T, M^*) | M^* \sim \phi^*]$. We will assess and compare the performance of online algorithms in terms of the cumulative and Bayesian regret.

With a slight abuse of notation, we denote by $\Pi_{\text{MIN}} = \{\mathfrak{D}_{H_X} \mapsto \mathfrak{D}_X : \forall X \in \boldsymbol{X}\}$, the minimal reduction obtained from `Reduce`$(\mathcal{G}, \Pi, Y)$. Let $\boldsymbol{S} = (\cup_{X \in \boldsymbol{X}} H_X) \setminus \boldsymbol{X}$. For any policy $\sigma_{\boldsymbol{X}} \in \Pi_{\text{MIN}}$, $E_{\sigma_{\boldsymbol{X}}}[Y]$ could be written as

$$E_{\sigma_{\boldsymbol{X}}}[Y] = \sum_{\boldsymbol{s}, \boldsymbol{x}} E_{\boldsymbol{x}}[Y | \boldsymbol{s}] P_{\boldsymbol{x}}(\boldsymbol{s}) \prod_{X \in \boldsymbol{X}} \pi_X(x | h_X). \tag{1}$$

Among quantities in the above equation, only transitional probabilities $P_{\boldsymbol{x}}(\boldsymbol{s})$ and immediate outcome $E_{\boldsymbol{x}}[Y | \boldsymbol{s}]$ are unknown. It thus suffices to learn $P_{\boldsymbol{x}}(\boldsymbol{s})$ and $E_{\boldsymbol{x}}[Y | \boldsymbol{s}]$ to identify the optimal policy. In the remainder of this paper, we will focus on the projection $\mathcal{G}_{\text{MIN}}$ from $\mathcal{G}$ over variables



(a) $\mathcal{G}$ (b) $\mathcal{G}_{\sigma_{X_1,X_2}}$ (c) $\mathcal{G}_{[S_1,S_2]}$

Figure 2: (a) A causal diagram $\mathcal{G}$; (b) the manipulated diagram $\mathcal{G}_{\sigma_{X_1,X_2}}$ with $\Pi = \{\mathcal{D}_{S_1} \mapsto \mathcal{D}_{X_1}, \mathcal{D}_{\{S_1,X_1,S_2\}} \mapsto \mathcal{D}_{X_2}\}$; (c) the subgraph $\mathcal{G}_{[S_1,S_2]}$.

$\{\boldsymbol{S},\boldsymbol{X},Y\}$, i.e., $\mathcal{G}_{\text{MIN}} = \texttt{Proj}(\mathcal{G},\{\boldsymbol{S},\boldsymbol{X},Y\})$. We will consistently use Π and $\mathcal{G}$, respectively, to represent the minimal reduction Π_{MIN} and the projection $\mathcal{G}_{\text{MIN}}$. For convenience of analysis, we will assume that outcome $E_{\boldsymbol{x}}[Y|\boldsymbol{s}]$ are provided. However, our methods extend trivially to settings where $E_{\boldsymbol{x}}[Y|\boldsymbol{s}]$ are unknown.

3.1. Optimism in the Face of Uncertainty

We now introduce a new online algorithms, `OFU-DTR`, for learning the optimal dynamic treatment regime in an unknown SCM. `OFU-DTR` follows the celebrated principle of *optimism in the face of uncertainty* (OFU). Like many other OFU algorithms (Auer et al., 2002; Jaksch et al., 2010; Osband & Van Roy, 2014), `OFU-DTR` works in phases comprised of optimistic planning, policy execution and model updating. One innovation in our work is to leverage the causal relationships in the underlying environment that enables us to obtain tighter regret bounds.

The details of the `OFU-DTR` algorithm are described in Alg. 2. During initialization, it simplifies the policy space Π and causal diagram $\mathcal{G}$ using `Reduce` and `Proj`. `OFU-DTR` interacts with the environment through policies in Π in repeated episodes of $t = 1,\dots,T$. At each episode t, it maintains a confidence set $\mathcal{P}_t$ over possible parameters of $P_{\boldsymbol{x}}(\boldsymbol{s})$ from samples collected prior to episode t. We will discuss the confidence set construction later in this section. Given a confidence set $\mathcal{P}_t$, `OFU-DTR` computes a policy $\sigma^t_{\boldsymbol{X}}$ by performing optimistic planning. More specifically, let $V_{\sigma_{\boldsymbol{X}}}(P_{\boldsymbol{x}}(\boldsymbol{s}))$ denote the function for $E_{\sigma_{\boldsymbol{X}}}[Y]$ given by Eq. (1). `OFU-DTR` finds the optimal policy $\sigma^t_{\boldsymbol{X}}$ for the most optimistic instance $P^t_{\boldsymbol{x}}(\boldsymbol{s})$ from $\mathcal{P}_t$ that induces the maximal outcome $V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s}))$. Since Π is soluble, one could solve for $\sigma^t_{\boldsymbol{X}}$ by extending the standard single policy update planner (Lauritzen & Nilsson, 2001), which we describe in (Zhang & Bareinboim, 2020, Appendix D). Finally, `OFU-DTR` executes $\sigma^t_{\boldsymbol{X}}$ throughout episode t and new samples $\boldsymbol{X}^t, \boldsymbol{S}^t$ are collected.

```
Algorithm 2 OFU-DTR
1: Input: Signature $[\![\mathcal{G},\Pi,Y]\!]$, $\delta \in (0,1)$.
2: Initialization: Let $\Pi = \texttt{Reduce}(\mathcal{G},\Pi,Y)$ and let
   $\mathcal{G} = \texttt{Proj}(\mathcal{G},\{\boldsymbol{S},\boldsymbol{X},Y\})$.
3: for all episodes $t = 1,2,\dots$ do
4:     Define counts $n^t(\boldsymbol{z})$ for any event $\boldsymbol{Z} = \boldsymbol{z}$ prior to
       episode $t$ as $n^t(\boldsymbol{z}) = \sum_{i=1}^{t-1} I_{\{\boldsymbol{Z}^i=\boldsymbol{z}\}}$.
5:     For any $S_k \in \boldsymbol{S}$, compute estimates
           $$\hat{P}^t_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) = \frac{n^t(\bar{\boldsymbol{x}}_k,\bar{\boldsymbol{s}}_k)}{\max\{n^t(\bar{\boldsymbol{x}}_k,\bar{\boldsymbol{s}}_k \setminus \{s_k\}),1\}}.$$
6:     Let $\mathcal{P}_t$ denote a set of distributions $P_{\boldsymbol{x}}(\boldsymbol{s})$ such that
       its factor $P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$ in Eq. (2) satisfies
           $$\left\|P_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k\setminus\{s_k\}) - \hat{P}^t_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k\setminus\{s_k\})\right\|_1 \leq f_{S_k}(t,\delta),$$
       where $f_{S_k}(t,\delta)$ is a function defined as
           $$f_{S_k}(t,\delta) = \sqrt{\frac{6|\mathcal{D}_{S_k}|\log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|t/\delta)}{\max\{n^t(\bar{\boldsymbol{x}}_k,\bar{\boldsymbol{s}}_k \setminus \{s_k\}),1\}}}.$$
7:     Find the optimistic policy $\sigma^t_{\boldsymbol{X}}$ such that
           $$\sigma^t_{\boldsymbol{X}} = \arg\max_{\sigma_{\boldsymbol{X}}\in\Pi} \max_{P^t_{\boldsymbol{x}}(\boldsymbol{s})\in\mathcal{P}_t} V_{\sigma_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) \qquad (3)$$
8:     Perform $do(\sigma^t_{\boldsymbol{X}})$ and observe $\boldsymbol{X}^t, \boldsymbol{S}^t$.
9: end for
```

Confidence Set Consider a soluble ordering $\prec$ on $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$. Let $\boldsymbol{S}$ be ordered by $S_1 \prec \cdots \prec S_m$. For any $\boldsymbol{S}^{(k)}$, let $\mathcal{G}_{[\boldsymbol{S}^{(k)}]}$ be a subgraph of $\mathcal{G}$ which includes $\boldsymbol{S}^{(k)}$ and edges among its elements. It follows from (Tian, 2002, Lem. 11) that $P_{\boldsymbol{x}}(\boldsymbol{s})$ factorize over c-components in $\mathcal{G}$.

Corollary 1. *Given $[\![\mathcal{G},\Pi,Y]\!]$, for any $S_k \in \boldsymbol{S}$, let $\bar{\boldsymbol{S}}_k$ denote a c-component in $\mathcal{G}_{[\boldsymbol{S}^{(k)}]}$ that contains S_k and let $\bar{\boldsymbol{X}}_k = Pa(\bar{\boldsymbol{S}}_k)_{\mathcal{G}} \setminus \bar{\boldsymbol{S}}_k$. $P_{\boldsymbol{x}}(\boldsymbol{s})$ could be written as:*

$$P_{\boldsymbol{x}}(\boldsymbol{s}) = \prod_{S_k\in\boldsymbol{S}} P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\}). \qquad (2)$$

Consider the causal diagram $\mathcal{G}$ of Fig. 2a as an example. By definition, the policy space Π described in Fig. 2b is minimal. Thus, $\boldsymbol{S} = \{S_1,S_2\}, \boldsymbol{X} = \{X_1,X_2\}$. We observes in Fig. 2c that $\{S_2\}$ is the c-component in subgraph $\mathcal{G}_{[S_1,S_2]}$ that contains S_2; c-component $\{S_1\}$ contains S_1 in $\mathcal{G}_{[\{S_1\}]}$. Corol. 1 implies $P_{x_1,x_2}(s_1,s_2) = P(s_1)P_{x_1}(s_2)$, which gives $P_{x_1,x_2}(s_2|s_1) = P_{x_1}(s_2)$ and $P_{x_1,x_2}(s_1) = P(s_1)$.

At each episode t, `OFU-DTR` computes the empirical estimator $\hat{P}^t_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k\setminus\{s_k\})$ for each factor in Eq. (2). Specifically, for samples $\mathcal{H}_t = \{\boldsymbol{X}^i,\boldsymbol{S}^i\}_{i=1}^{t-1}$ collected prior to episode t, $\hat{P}^t_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k\setminus\{s_k\})$ is the relative frequency of event $S^t_k = s_k$ at the state $\bar{\boldsymbol{S}}^t_k \setminus \{S^t_k\} = \bar{\boldsymbol{s}}_k \setminus \{s_k\}, \bar{\boldsymbol{X}}^t_k = \bar{\boldsymbol{x}}_k$. The confidence set $\mathcal{P}_t$ is defined as a series of convex intervals centered around estimates $\hat{P}^t_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$ (Step 6). The adaptive sampling process of OFU-DTR ensures the identifiability of interventional probabilities $P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$.

Lemma 4. *Given* $[\![\mathcal{G},\Pi,Y]\!]$*, for any* $S_k \in \boldsymbol{S}$ *and any* $\sigma_{\boldsymbol{X}} \in \Pi$, $P_{\sigma_{\boldsymbol{X}}}(s_k|\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\}) = P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$.

We are now ready to analyze asymptotic properties of OFU-DTR, which will lead to a better understanding of their theoretical guarantees.

Theorem 3. *Given* $[\![\mathcal{G},\Pi,Y]\!]$*, fix a* $\delta \in (0,1)$*. With probability (w.p.) at least* $1-\delta$*, it holds for any* $T > 1$*, the regret of* OFU-DTR *is bounded by*

$$R(T, M^*) \leq \Delta(T,\delta) + 2|\boldsymbol{S}|\sqrt{T\log(2|\boldsymbol{S}|T/\delta)}, \tag{4}$$

where $\Delta(T,\delta)$ *is a function defined as*

$$\Delta(T,\delta) = \sum_{S_k \in \boldsymbol{S}} 17\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}| T \log(|\boldsymbol{S}|T/\delta)}.$$

OFU-DTR improves over the state-of-art online algorithms for DTRs. Consider again the policy space Π in Fig. 2b. Oblivious of the causal diagram $\mathcal{G}$, the algorithm developed in (Zhang & Bareinboim, 2019) leads to a near-optimal regret $\tilde{\mathcal{O}}(\sqrt{|\mathcal{D}_{\{S_1,S_2,X_1\}}|T})$ [2] [3]. Thm. 3 implies that OFU-DTR achieves a regret bound $\tilde{\mathcal{O}}(\sqrt{|\mathcal{D}_{\{S_2,X_1\}}|T})$, removing the factor of $\sqrt{|\mathcal{D}_{\{S_1\}}|}$. In general, if $|\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}| < |\mathcal{D}_{\boldsymbol{S}\cup\boldsymbol{X}}|$ for some S_k, OFU-DTR outperforms state-of-art methods by exploiting the causal knowledge of $\mathcal{G}$.

3.2. Posterior Sampling

We now introduce an alternative algorithm, PS-DTR, based on the heuristics of posterior sampling (Thompson, 1933; Strens, 2000; Osband et al., 2013). We will focus on the Bayesian settings where the actual M^* is drawn from a set of candidate SCMs $\mathcal{M}$ following a distribution ϕ^*. The details of PS-DTR are described in Alg. 3. In addition to $[\![\mathcal{G},\Pi,Y]\!]$, PS-DTR assumes the access to a prior ϕ over the interventional probabilities $P_{\boldsymbol{x}}(\boldsymbol{s})$, i.e.,

$$\phi(\boldsymbol{\theta}) = \sum_{M\in\mathcal{M}} I_{\{P_{M_{\boldsymbol{x}}}(\boldsymbol{s})=\boldsymbol{\theta}\}}\phi^*(M). \tag{5}$$

In practice, for the discrete domains, ϕ could be the product of a series of uninformative Dirichlet priors. Similar to OFU-DTR, PS-DTR first simplifies the policy space Π and causal diagram $\mathcal{G}$ and proceeds in repeated episodes. At each episode t, PS-DTR updates the posterior $\phi(\cdot|\mathcal{H}_t)$ from collected samples $\mathcal{H}_t = \{\boldsymbol{X}^i, \boldsymbol{S}^i\}_{i=1}^{t-1}$. It then draws an sampled estimate of $P^t_{\boldsymbol{x}}(\boldsymbol{s})$ from the updated posteriors.

Algorithm 3 PS-DTR

```
1: Input: Signature $[\![\mathcal{G},\Pi,Y]\!]$, prior $\phi$.
2: Initialization: Let $\Pi$ = Reduce($\mathcal{G},\Pi,Y$) and let
   $\mathcal{G}$ = Proj($\mathcal{G},\{\boldsymbol{S},\boldsymbol{X},Y\}$).
3: for all episodes $t = 1,2,\dots$ do
4:     Sample $P^t_{\boldsymbol{x}}(\boldsymbol{s}) \sim \phi(\cdot|\mathcal{H}_t)$.
5:     Compute the optimal policy $\sigma^t_{\boldsymbol{X}}$ such that
           $\sigma^t_{\boldsymbol{X}} = \arg\max_{\sigma_{\boldsymbol{X}}\in\Pi} V_{\sigma_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})).$   (7)
6:     Perform $do(\sigma^t_{\boldsymbol{X}})$ and observe $\boldsymbol{X}^t, \boldsymbol{S}^t$.
7: end for
```

In Step 5, PS-DTR computes an optimal policy $\sigma^t_{\boldsymbol{X}}$ that maximizes the expected outcome $V_{\sigma_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s}))$ induced by the sampled $P^t_{\boldsymbol{x}}(\boldsymbol{s})$. Finally, $\sigma^t_{\boldsymbol{X}}$ is executed throughout episode t and new samples $\boldsymbol{X}^t, \boldsymbol{S}^t$ are collected.

Theorem 4. *Given* $[\![\mathcal{G},\Pi,Y]\!]$ *and a prior* ϕ*, if* ϕ *satisfies Eq. (5), it holds for any* $T > 1$*, the regret of* PS-DTR *is bounded by*

$$R(T,\phi^*) \leq \Delta(T, 1/T) + 1, \tag{6}$$

where function $\Delta(T,\delta)$ *follows the definition in Thm. 3.*

Compared with Thm. 3, the regret bound in Thm. 4 implies that PS-DTR achieves the similar asymptotic performance as OFU-DTR. In OFU-DTR, one has to find an optimal policy $\sigma^t_{\boldsymbol{X}}$ for the most optimistic instance in a family of SCMs, whose distribution $P_{\boldsymbol{x}}(\boldsymbol{s})$ are imprecise, bounded in a convex polytope $\mathcal{P}_t$ (Eq. (3)). On the other hand, the policy $\sigma^t_{\boldsymbol{X}}$ in PS-DTR is a solution for SCMs with fixed probabilities $P^t_{\boldsymbol{x}}(\boldsymbol{s})$. Since Π is soluble, such policy $\sigma^t_{\boldsymbol{X}}$ could be obtained using the standard dynamic program solvers (Nilsson & Lauritzen, 2000; Koller & Milch, 2003). Preliminary analysis reveals that solving for the optimal policy with with imprecise probabilities performs at least the double of the number of arithmetic operations required with fixed-point values (Cabañas et al., 2017). This suggests that PS-DTR is more computationally efficient compared to OFU-DTR.

4. Learning From Observational Data

Algorithms introduced so far learn the optimal policy through repeated experiments from scratch. In many applications, however, conducting experiments in the actual environment could be extremely costly and undesirable due to unintended consequences. A natural solution is to extrapolate knowledge from the observational data, so that the future online learning process could be accelerated.

Given the causal diagram $\mathcal{G}$, one could apply standard causal identification algorithms (Tian, 2002; Tian & Pearl, 2002; Shpitser & Pearl, 2006; Huang & Valtorta, 2006) to esti-

[2] $\mathcal{D}_{\{X_2\}}$ is omitted since we assume $E_{\boldsymbol{x}}[Y|\boldsymbol{s}]$ is provided.

[3] To the best of our knowledge, the family of algorithms proposed in (Zhang & Bareinboim, 2019) are the first adaptive strategies that work regardless of the causal graph, which extends results for bandits found in the literature (Zhang & Bareinboim, 2017).

mate the causal effect (e.g., $P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$) from the observational distribution $P(\boldsymbol{v})$. However, challenges of non-identifiability could arise and the target effects may be not uniquely computable from the data.

Inferring about treatment effects in non-identifiable settings has been a target of growing interest in the domains of causal inference (Balke & Pearl, 1995; Chickering & Pearl, 1996; Richardson et al., 2014; Zhang & Bareinboim, 2017; Kallus & Zhou, 2018; Kallus et al., 2018; Cinelli et al., 2019). To address this challenge, we consider a partial identification approach which reduces the parameter space of causal effects from the observational data, called the *causal bounds*. Following (Tian & Pearl, 2002), for any $\boldsymbol{S} \subseteq \boldsymbol{V}$, we define function $Q[\boldsymbol{S}](\boldsymbol{v}) = P_{\boldsymbol{v} \setminus \boldsymbol{s}}(\boldsymbol{s})$. Also, $Q[\boldsymbol{V}](\boldsymbol{v}) = P(\boldsymbol{v})$ and $Q[\emptyset](\boldsymbol{v}) = 1$. For convenience, we often omit input $\boldsymbol{v}$ and write $Q[\boldsymbol{S}]$. Our first result derives inequality relationships among Q functions.

Lemma 5. *For a SCM $\langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, P(\boldsymbol{u})\rangle$, let subsets $\boldsymbol{S} \subseteq \boldsymbol{C} \subseteq \boldsymbol{V}$. For a topological ordering $\prec$ in $\mathcal{G}$, let $\boldsymbol{S}$ be ordered by $S_1 \prec \cdots \prec S_k$. $Q[\boldsymbol{S}]$ is bounded from $Q[\boldsymbol{C}]$ as:*

$$Q[\boldsymbol{S}] \in \left[A(\boldsymbol{S}, Q[\boldsymbol{C}]), B(\boldsymbol{S}, Q[\boldsymbol{C}])\right],$$

where $A(\boldsymbol{S}, Q[\boldsymbol{C}]), B(\boldsymbol{S}, Q[\boldsymbol{C}])$ are functions defined as follows. Let $\boldsymbol{W} = An(\boldsymbol{S})_{\mathcal{G}_{[\boldsymbol{C}]}}$. If $\boldsymbol{W} = \boldsymbol{S}$,

$$A(\boldsymbol{S}, Q[\boldsymbol{C}]) = B(\boldsymbol{S}, Q[\boldsymbol{C}]) = Q[\boldsymbol{W}],$$

where $Q[\boldsymbol{W}] = \sum_{\boldsymbol{c} \setminus \boldsymbol{w}} Q[\boldsymbol{C}]$; otherwise,

$$\begin{aligned} A(\boldsymbol{S}, Q[\boldsymbol{C}]) &= \max_{\boldsymbol{z}} Q[\boldsymbol{W}], \\ B(\boldsymbol{S}, Q[\boldsymbol{C}]) &= \min_{\boldsymbol{z}} \left\{ Q[\boldsymbol{W}] - \sum_{s_k} Q[\boldsymbol{W}] \right\} \\ &\quad + B(\boldsymbol{S} \setminus \{S_k\}, Q[\boldsymbol{C}]), \end{aligned}$$

where $\boldsymbol{Z} = Pa(\boldsymbol{W})_{\mathcal{G}} \setminus Pa(\boldsymbol{S})_{\mathcal{G}}$.

While this result may appear non-trivial, Lem. 5 generalizes the natural bounds in (Manski, 1990) to longitude settings. For instance, in Fig. 2a, $P_{x_1}(s_1, s_2)$ is not identifiable due to the presence of UCs (i.e., $X_1 \leftrightarrow S_1$). Let $\boldsymbol{S} = \{S_1, S_2\}$ and $\boldsymbol{C} = \{S_1, S_2, X_1\}$. Lem. 5 allows us to bound $P_{x_1}(s_1, s_2)$ from $P(s_1, s_2, x_1)$ as $P_{x_1}(s_1, s_2) \geq P(s_1, s_2, x_1)$ and $P_{x_1}(s_1, s_2) \leq P(s_1, s_2, x_1) - P(s_1, x_1) + P(s_1)$.

Theorem 5 (C-component Bounds)**.** *Given $[\![\mathcal{G}, \Pi, Y]\!]$, for any $S_k \in \boldsymbol{S}$, let $\boldsymbol{C}$ be a c-component in $\mathcal{G}$ that contains $\bar{\boldsymbol{S}}_k$. Let $\boldsymbol{C}_k = \boldsymbol{C} \cap \boldsymbol{S}^{(k)}$ and let $\boldsymbol{Z} = Pa(\boldsymbol{C}_k)_{\mathcal{G}} \setminus Pa(\bar{\boldsymbol{S}}_k)_{\mathcal{G}}$. $P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$ is bounded in $\left[a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}, b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}\right]$ where*

$$\begin{aligned} a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k} &= \max_{\boldsymbol{z}} \left\{ A(\boldsymbol{C}_k, Q[\boldsymbol{C}]) / B(\boldsymbol{C}_k \setminus \{S_k\}, Q[\boldsymbol{C}]) \right\}, \\ b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k} &= \min_{\boldsymbol{z}} \left\{ B(\boldsymbol{C}_k, Q[\boldsymbol{C}]) / B(\boldsymbol{C}_k \setminus \{S_k\}, Q[\boldsymbol{C}]) \right\}. \end{aligned}$$

Among quantities in the above equation, $Q[\boldsymbol{C}]$ is identifiable from the observational data $P(\boldsymbol{v})$ following (Tian, 2002, Lem. 7). Thm. 5 improves the DTR bounds in (Zhang & Bareinboim, 2019) by exploiting the independence relationships among variables $\boldsymbol{S}$. For example, in Fig. 2a, S_1 and S_2 are independent under $do(x_1)$. That is, $P_{x_1}(s_2) = P_{x_1}(s_2, s_1)/P(s_1)$ for any s_1. By Thm. 5, $\boldsymbol{C} = \{S_1, S_2, X_1\}$ and $\boldsymbol{C}_k = \{S_1, S_2\}$. Bounding $Q[\boldsymbol{C}_k]$ from $Q[\boldsymbol{C}]$ gives $P_{x_1}(s_2) \geq \max_{s_1} P(x_1, s_2|s_1)$ and $P_{x_1}(s_2) \leq \min_{s_1} P(x_1, s_2|s_1) - P(x_1|s_1) + 1$.

4.1. Online Learning with Causal Bounds

We next introduce efficient methods to incorporate the causal bounds into online learning algorithms. For any $S_k \in \boldsymbol{S}$, let $\mathfrak{C}_{S_k}$ denote a parameter family of $P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$ induced by causal bounds $\left[a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}, b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}\right]$. We denote by $\mathfrak{C}$ a sequence $\{\mathfrak{C}_{S_k} : \forall S_k \in \boldsymbol{S}\}$. Naturally, $\mathfrak{C}$ defines a family $\mathcal{P}_c$ of parameters for the interventional distribution $P_{\boldsymbol{x}}(\boldsymbol{s})$. To incorporate the causal bounds $\mathfrak{C}$, `OFU-DTR` finds the optimal policy $\sigma^t_{\boldsymbol{X}}$ of the most optimistic instance in the family of probabilities $\mathcal{P}_c \cap \mathcal{P}_t$. That is, we replace the optimization problem defined in Eq. (3) with the following:

$$\sigma^t_{\boldsymbol{X}} = \arg\max_{\sigma_{\boldsymbol{X}} \in \Pi} \max_{P^t_{\boldsymbol{x}}(\boldsymbol{s}) \in \mathcal{P}_c \cap \mathcal{P}_t} V_{\sigma_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) \tag{8}$$

Let $|\mathfrak{C}_{S_k}|$ denote the maximal L1 norm of any pair of probability distributions in $\mathfrak{C}_k$, i.e.,

$$|\mathfrak{C}_{S_k}| = \max_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\}} \sum_{s_k} \left| a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k} - b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k} \right|.$$

We are now ready to derive the regret bound of `OFU-DTR` that incorporate causal bounds $\mathfrak{C}$ through Eq. (8).

Theorem 6. *Given $[\![\mathcal{G}, \Pi, Y]\!]$ and causal bounds $\mathfrak{C}$, fix a $\delta \in (0, 1)$. W.p. at least $1 - \delta$, it holds for any $T > 1$, the regret of `OFU-DTR` is bounded by*

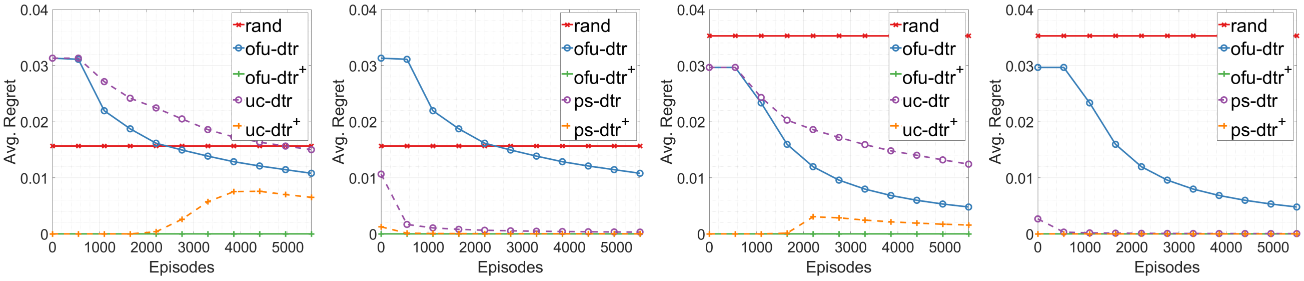
$$R(T, M^*) \leq \Delta(T, \mathfrak{C}, \delta) + 2|\boldsymbol{S}|\sqrt{T \log(2|\boldsymbol{S}|T/\delta)},$$

where function $\Delta(T, \mathfrak{C}, \delta)$ is defined as

$$\sum_{S_k \in \boldsymbol{S}} \min \left\{ |\mathfrak{C}_{S_k}|T, 17\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}|T \log(|\boldsymbol{S}|T/\delta)} \right\}.$$

It follows immediately that the regret bound in Thm. 6 is smaller than the bound given by Thm. 3 if $T < 12^2|\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}|T \log(|\boldsymbol{S}|T/\delta)/|\mathfrak{C}_{S_k}|^2$ for some S_k. This means that the causal bounds $\mathfrak{C}$ give `OFU-DTR` a head start when bounds $\mathfrak{C}$ are informative, i.e., the dimension $|\mathfrak{C}_{S_k}|$ is small for some S_k. When $P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$ is identifiable, i.e., $|\mathfrak{C}_{S_k}| = 0$, no exploration is required.

Posterior Sampling We also provide an efficient method to account for the observational data through causal bounds



(a) Lung Cancer (b) Lung Cancer (c) Dyspnoea (d) Dyspnoea

Figure 3: Simulations comparing the sequential multiple assignment randomized trail (*rand*), OFU-DTR algorithm (*ofu-dtr*), PS-DTR algorithm (*ps-dtr*) and UC-DTR algorithm (*uc-dtr*). We use superscript + to indicate algorithms warm-started with causal bounds derived from the confounded observational data (*ofu-dtr*$^+$, *ps-dtr*$^+$, *uc-dtr*$^+$).

$\mathfrak{C}$ in PS-DTR. We will employ a rejection sampling procedure which repeatedly samples from ϕ until the sampled estimate $P^t_{\boldsymbol{x}}(\boldsymbol{s})$ is compatible with the parameter family $\mathcal{P}_c$. That is, we replace Step 4 in PS-DTR with the following:

$$\textbf{repeat } P^t_{\boldsymbol{x}}(\boldsymbol{s}) \sim \phi(\cdot|\mathcal{H}_t) \textbf{ until } P^t_{\boldsymbol{x}}(\boldsymbol{s}) \in \mathcal{P}_c$$

The remainder of PS-DTR proceeds accordingly, without any modification. We next show that the above procedure allows PS-DTR to achieve the similar performance as OFU-DTR provided with the causal bounds $\mathfrak{C}$.

Theorem 7. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, a prior ϕ and causal bounds $\mathfrak{C}$, if ϕ satisfies Eq. (5), it holds for any $T > 1$, the regret of PS-DTR is bounded by*

$$R(T, \phi) \leq \Delta(T, \mathfrak{C}, 1/T) + 1, \tag{9}$$

where function $\Delta(T, \mathfrak{C}, \delta)$ follows the definition in Thm. 6.

Thm. 7 implies that PS-DTR provided with causal bounds $\mathfrak{C}$ consistently dominate its counterpart without using any observational data in terms of the performance. The condition of improvements coincides with that of OFU-DTR, which we show in Thm. 6.

5. Experiments

We evaluate the new algorithms on several SCMs, including multi-stage treatment regimes for lung cancer (Nease Jr & Owens, 1997) and dyspnoea (Cowell et al., 2006). We found that the new algorithms consistently outperform the state-of-art methods in terms of both the online performance and the efficiency of utilizing the observational data.

Throughout all the experiments, we test OFU-DTR algorithm (*ofu-dtr*) with failure tolerance $\delta = 1/T$, OFU-DTR with causal bounds (*ofu-dtr*$^+$) with causal bounds derived from the observational data, PS-DTR algorithm (*ps-dtr*) using uninformative dirichlet priors, and PS-DTR incorporating causal bounds via rejection sampling (*ps-dtr*$^+$). As a baseline, we also include the sequential multiple assignment randomized trail (*rand*), UC-DTR algorithm (*uc-dtr*), and causal UC-DTR algorithm (*uc-dtr*$^+$) developed in (Zhang & Bareinboim, 2019). To emulate the unobserved confounding, we generate 2×10^6 observational samples using a behavior policy and hide some of the covariates (i.e., some columns). Each experiment lasts for $T = 5.5 \times 10^3$ episodes. For all algorithms, we measure their average regrets $R(T, M^*)/T$ over 100 repetitions. We refer readers to (Zhang & Bareinboim, 2020, Appendix E) for more details on the experiments.

Lung Cancer We test the model of treatment regimes for lung cancer described in (Nease Jr & Owens, 1997). Given the results of CT for mediastinal metastases, the physician could decide to perform an additional mediastinoscopy test. Finally, based on the test results and treatment histories, the physician could recommend a thoracotomy or a radio therapy. The average regret of all algorithms are reported in Fig. 3a. We find that our algorithms (*ofu-dtr*, *ofu-dtr*$^+$), leveraging the causal diagram, demonstrate faster convergence compared to the state-of-art methods (*uc-dtr*, *uc-dtr*$^+$). The causal bounds derived from the observational data generally improve the online performance (*ofu-dtr*$^+$, *uc-dtr*$^+$). By exploiting sharper causal bounds, *ofu-dtr*$^+$ finds the optimal treatment policy almost immediately while *uc-dtr*$^+$ still does not converge until 4×10^3 episodes. We also compare the performance of OFU-DTR and PS-DTR in Fig. 3b. In the pure online settings (without any previous observation), *ps-dtr* shows faster convergence than *ofu-dtr*. Provided with the same causal bounds, *ps-dtr*$^+$ rivals *ofu-dtr*$^+$ in terms of the performance and finds the optimal policy after only 500 episodes.

Dyspnoea We test the model of treatment regimes for dysponea (shortness of breath) described in (Cowell et al., 2006), called DEC-ASIA. Based on the patients' travel history, the physician could decide to perform a chest X-ray. If a test is carried out, the doctor has access to the results and the symptom of dysponea at the time she determining whether to hospitalize or not. We measure the average regrets for all algorithms, reported in Figs. 3c and 3d. As expected, `OFU-DTR` consistently outperforms the state-of-art methods `UC-DTR` in terms of both the online performance (*ofu-dtr*, *uc-dtr*) and the efficiency of extrapolating observational data (*ofu-dtr*$^+$, *uc-dtr*$^+$). Compared to `OFU-DTR`, `PS-DTR` demonstrates faster convergence in the pure online settings (*ps-dtr*) and achieves similar regrets when observational data are provided (*ps-dtr*$^+$). These results suggest that `PS-DTR` seems to be an attractive option in practice.

6. Conclusion

We present the first online algorithms with provable regret bounds for learning the optimal dynamic treatment regime in an unknown environment while leveraging the order relationships represented in the form of a causal diagram. These algorithms reduce the learning problem to finding an optimal policy for the most optimistic instance from a family of causal models whose interventional distributions are imprecise, bounded in a set of convex intervals. We believe that our results provide new opportunities for designing dynamic treatment regimes in unknown, and structured environments, even when the causal effects of candidate policies are not point-identifiable from the confounded observational data.

7. Acknowledgments

Zhang and Bareinboim are supported in parts by grants from NSF IIS-1704352 and IIS-1750807 (CAREER).

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Balke, A. and Pearl, J. Counterfactuals and policy analysis in structural models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 11–18, 1995.

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

Cabañas, R., Antonucci, A., Cano, A., and Gómez-Olmedo, M. Evaluating interval-valued influence diagrams. *International Journal of Approximate Reasoning*, 80, 2017.

Chakraborty, B. and Moodie, E. *Statistical methods for dynamic treatment regimes*. Springer, 2013.

Chickering, D. and Pearl, J. A clinician's apprentice for analyzing non-compliance. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume Volume II, pp. 1269–1276. MIT Press, Menlo Park, CA, 1996.

Cinelli, C., Kumor, D., Chen, B., Pearl, J., and Bareinboim, E. Sensitivity analysis of linear structural causal models. In *International Conference on Machine Learning*, pp. 1252–1261, 2019.

Cowell, R. G., Dawid, P., Lauritzen, S. L., and Spiegelhalter, D. J. *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer Science & Business Media, 2006.

Huang, Y. and Valtorta, M. Pearl's calculus of intervention is complete. In Dechter, R. and Richardson, T. (eds.), *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 217–224. AUAI Press, Corvallis, OR, 2006.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Kallus, N. and Zhou, A. Confounding-robust policy improvement. In *Advances in neural information processing systems*, pp. 9269–9279, 2018.

Kallus, N., Puli, A. M., and Shalit, U. Removing hidden confounding by experimental grounding. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 10911–10920. Curran Associates, Inc., 2018.

Kearns, M. and Koller, D. Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pp. 740–747, 1999.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Koller, D. and Milch, B. Multi-agent influence diagrams for representing and solving games. *Games and economic behavior*, 45(1):181–221, 2003.

Lauritzen, S. L. and Nilsson, D. Representing and solving decision problems with limited information. *Management Science*, 47(9):1235–1251, 2001.

Lavori, P. W. and Dawson, R. A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):29–38, 2000.

Lavori, P. W. and Dawson, R. Adaptive treatment strategies in chronic disease. *Annu. Rev. Med.*, 59:443–453, 2008.

Lunceford, J. K., Davidian, M., and Tsiatis, A. A. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 58(1):48–57, 2002.

Manski, C. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80: 319–323, 1990.

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

Murphy, S. A. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481, 2005a.

Murphy, S. A. A generalization error for q-learning. *Journal of Machine Learning Research*, 6(Jul), 2005b.

Murphy, S. A., van der Laan, M. J., and Robins, J. M. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Nease Jr, R. F. and Owens, D. K. Use of influence diagrams to structure medical decisions. *Medical Decision Making*, 17(3):263–275, 1997.

Nilsson, D. and Lauritzen, S. L. Evaluating influence diagrams using limids. In *Proceedings of the 16th conference on UAI*, pp. 436–445. Morgan Kaufmann Publishers Inc., 2000.

Osband, I. and Van Roy, B. Near-optimal reinforcement learning in factored mdps. In *Advances in Neural Information Processing Systems*, pp. 604–612, 2014.

Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. In *Advances in NeurIPS*, pp. 3003–3011, 2013.

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

Pearl, J. and Robins, J. Probabilistic evaluation of sequential plans from causal models with hidden variables. In Besnard, P. and Hanks, S. (eds.), *Uncertainty in Artificial Intelligence 11*. Morgan Kaufmann, San Francisco, 1995.

Richardson, A., Hudgens, M. G., Gilbert, P. B., and Fine, J. P. Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):596, 2014.

Robins, J., Orellana, L., and Rotnitzky, A. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in medicine*, 27(23):4678–4721, 2008.

Rosenbaum, P. and Rubin, D. The central role of propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

Rubin, D. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6(1):34–58, 1978.

Shpitser, I. and Pearl, J. Identification of conditional interventional distributions. In Dechter, R. and Richardson, T. (eds.), *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 437–444. AUAI Press, Corvallis, OR, 2006.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*, volume 81. MIT press, 2001.

Strens, M. A bayesian framework for reinforcement learning. In *ICML*, volume 2000, pp. 943–950, 2000.

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

Thall, P. F., Millikan, R. E., and Sung, H.-G. Evaluating multiple treatment courses in clinical trials. *Statistics in medicine*, 19(8):1011–1028, 2000.

Thall, P. F., Sung, H.-G., and Estey, E. H. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the American Statistical Association*, 97(457):29–39, 2002.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Tian, J. *Studies in Causal Reasoning and Learning*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, November 2002.

Tian, J. and Pearl, J. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pp. 567–573. AAAI Press/The MIT Press, Menlo Park, CA, 2002.

Wahed, A. S. and Tsiatis, A. A. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 60(1):124–133, 2004.

Wahed, A. S. and Tsiatis, A. A. Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. *Biometrika*, 93(1):163–177, 2006.

Wang, L., Rotnitzky, A., Lin, X., Millikan, R. E., and Thall, P. F. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association*, 107(498):493–508, 2012.

Zhang, J. and Bareinboim, E. Transfer learning in multi-armed bandits: a causal approach. In *Proceedings of the 26th IJCAI*, pp. 1340–1346, 2017.

Zhang, J. and Bareinboim, E. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, 2019.

Zhang, J. and Bareinboim, E. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. Technical Report R-57, Causal Artificial Intelligence Lab, Columbia University, 2020. URL https://causalai.net/r47-full.pdf.

"Designing Optimal Dynamic Treatment Regimes: A Causal Reinforcement Learning Approach" Supplemental Material

Anonymous Authors[1]

Appendix A. Proofs of Results in Section 2.1

In this section, we provide proofs for the results presented in Sec. 2.1. We first introduce some notations and lemmas that will be instrumental in the proofs. For a DAG G and a subset of nodes $\boldsymbol{X}$, we denote by $\mathcal{G}_{\overline{\boldsymbol{X}}}$ a subgraph of G by removing all incoming arrow into $\boldsymbol{X}$; $\mathcal{G}_{\underline{\boldsymbol{X}}}$ stands for a subgraph of G by removing all outgoing arrow of $\boldsymbol{X}$. For a signature $[\![\mathcal{G}, \Pi, Y]\!]$, we will consistently use $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ to represent the manipulated diagram of Π. For a subset $\boldsymbol{X}' \subseteq \boldsymbol{X}$, let $\mathcal{G}_{\sigma_{\boldsymbol{X}'}}$ be a manipulated diagram obtained from $\mathcal{G}$ and Π by changing parents to each treatment node $X \in \boldsymbol{X}'$ to nodes in H_X; arrows pointing to other treatments $\boldsymbol{X} \setminus \boldsymbol{X}'$ remain the same. For a reduction Π' of policy space Π, unless it is explicitly specified, the manipulated diagram of Π' is denoted by $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$. For any policy $\sigma_{\boldsymbol{X}} \in \Pi$ and subset of treatments $\boldsymbol{X}' \subseteq \boldsymbol{X}$, we denote by $\sigma_{\boldsymbol{X}'}$ a partial policy obtained from $\sigma_{\boldsymbol{X}}$ with restriction to treatments in the subset $\boldsymbol{X}'$.

Our proofs depend on the three inference rules of σ-calculus introduced in (Correa & Bareinboim, 2020, Thm. 1). The rules are derived based on the soundness of d-separation in DAGs. We first show that some basic causal constraints are preserved under the removal of irrelevant treatments.

Lemma 6. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, let subset $\tilde{\boldsymbol{X}} \subseteq \boldsymbol{X} \setminus (\boldsymbol{X} \cap An(Y))_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$. For any treatment $X \in \boldsymbol{X}$, $X \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{X}}}}}$ if and only if $X \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$.*

Proof. We first prove the "if" direction. For any treatment $X \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, suppose there exists a directed path g (called causal path) from X to Y in $\mathcal{G}_{\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{X}}}}$. Since $X \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, path g must contain incoming arrows $V_j \rightarrow X'$ for some $X' \in \tilde{\boldsymbol{X}}$ such that $X \neq X'$. Let X' denote the last treatment on l that are in $\tilde{\boldsymbol{X}}$. We could then obtain from g a subpath g' that is a causal path from X' to Y in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$. Since X' is the last treatment on g that is in $\tilde{\boldsymbol{X}}$, the subpath g' must also exists in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$, i.e., $X' \in An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, which is a contradiction.

We now prove the "only if" direction. Suppose there exists a treatment $X \in An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$ but $X \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{X}}}}}$. Let g denote a causal path from X to Y in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$. Since $X \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{X}}}}}$, path g must contain incoming arrows $V_j \rightarrow X'$ for some $X' \in \tilde{\boldsymbol{X}}$ such that $X \neq X'$. Let X' denote the last treatment on l that are in $\tilde{\boldsymbol{X}}$. We could thus obtain a causal path g' from X' to Y in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$. This means that $X' \in An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, which is a contradiction. □

Lem. 6 allows us to show that the acyclicity is preserved under reduction.

Lemma 7. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, let Π' be a reduction of Π. Let $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$ denote the manipulated diagram of Π'. $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$ is acyclic if G and $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ are acyclic.*

Proof. It suffices to prove that the acyclicity is preserved under the removal of irrelevant treatments and evidences. Suppose Π' is a reduction of Π obtained by removing irrelevant evidences $\tilde{\boldsymbol{S}} \mapsto X$. Since $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ is a DAG and removing arrows from a DAG does not create cycles, $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$ is acyclic.

Consider now that Π' is a reduction of Π obtained by removing irrelevant treatments $\tilde{\boldsymbol{X}}$. Suppose there exists a cycle l in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$. Since both G and $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ are acyclic, there must exist a pair X_1, X_2 on l where $X_1 \in \tilde{\boldsymbol{X}}$ and $X_2 \in \boldsymbol{X} \setminus \tilde{\boldsymbol{X}}$. Lem. 6 implies that $X_1 \notin An(Y)_{\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}}$. By definitions, $X_2 \in An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, i.e., there exist a causal path g from X_2 to Y in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$. Since $\tilde{\boldsymbol{X}}$ are irrelevant in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$, g must not contain any incoming arrow $V_i \rightarrow X'$ where $X' \in \tilde{\boldsymbol{X}}$. That is, path g is preserved in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$. We could thus obtain a causal path from X_1 to Y by concatenating g with a subsequence in l from X_1 to X_2, which is a contradiction. □

We are now ready to prove the results presented in Sec. 2.2. By Lem. 7, any reduction Π' of the policy space Π will induce a DAG $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$. We could thus assume without loss of generality that for any signature $[\![\mathcal{G}, \Pi, Y]\!]$ of interest, the manipulated graph $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ must be a DAG. We will use this assumption throughout the proof.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Lemma 1. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, $\Pi \setminus \tilde{\boldsymbol{X}}$ is equivalent to Π if treatments $\tilde{\boldsymbol{X}}$ are irrelevant.*

Proof. Let Π' denote the reduction $\Pi \setminus \tilde{\boldsymbol{X}}$. By definitions, $\tilde{\boldsymbol{X}} = \boldsymbol{X} \setminus An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$. For any $\sigma_{\boldsymbol{X}} \in \Pi$, let $\sigma_{\boldsymbol{X} \setminus \tilde{\boldsymbol{X}}}$ denote its partial policy with restriction in $\boldsymbol{X} \setminus \tilde{\boldsymbol{X}}$; naturally, $\sigma_{\boldsymbol{X} \setminus \tilde{\boldsymbol{X}}} \in \Pi'$. Lem. 6 implies that Y is not a non-descendant of $\tilde{\boldsymbol{X}}$ in $\mathcal{G}_{\sigma_{\boldsymbol{X} \setminus \tilde{\boldsymbol{X}}}}$. We thus have

$$(Y \perp\!\!\!\perp \tilde{\boldsymbol{X}})_{\mathcal{G}_{\sigma_{\boldsymbol{X}} \overline{\tilde{\boldsymbol{X}}}}}, \qquad (Y \perp\!\!\!\perp \tilde{\boldsymbol{X}})_{\mathcal{G}_{\sigma_{\boldsymbol{X} \setminus \tilde{\boldsymbol{X}}} \overline{\tilde{\boldsymbol{X}}}}}.$$

Lem. 7 implies that $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$ is a DAG. The acyclicity guarantee, together with the above independence relationships, gives that

$$P_{\sigma_{\boldsymbol{X}}}(y) = P_{\sigma_{\boldsymbol{X} \setminus \tilde{\boldsymbol{X}}}}(y).$$

The above equality is ensured by (Correa & Bareinboim, 2020, Thm. 1), which proves the statement. □

Lemma 2. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, $\Pi \setminus \{\tilde{\boldsymbol{S}} \mapsto X\}$ is equivalent to Π if evidences $\tilde{\boldsymbol{S}} \mapsto X$ are irrelevant.*

Proof. Let Π' denote the reduction $\Pi \setminus \{\tilde{\boldsymbol{S}} \mapsto X\}$. If $X \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, we have

$$(Y \perp\!\!\!\perp X)_{\mathcal{G}_{\sigma_{\boldsymbol{X}} \overline{X}}}.$$

where $\mathcal{G}_{\sigma_{\boldsymbol{X}} \overline{X}}$ is a subgraph of $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ by removing incoming arrows into X. By Rule 3 of (Correa & Bareinboim, 2020, Thm. 1), the above independence relationship implies that, for any policy $\sigma_{\boldsymbol{X}} \in \Pi$ and any $\sigma'_X \in \{\mathfrak{D}_{H_X} \mapsto \mathfrak{D}_X\}$,

$$P_{\sigma_{\boldsymbol{X} \setminus X}, \sigma_X}(y) = P_{\sigma_{\boldsymbol{X} \setminus X}, \sigma'_X}(y).$$

Let the decision rule $\sigma'_X \in \{D_{H_X \setminus \tilde{\boldsymbol{S}}} \mapsto \mathfrak{D}_X\}$ and let $\sigma'_{\boldsymbol{X}} = \{\sigma_{\boldsymbol{X} \setminus X}, \sigma'_X\}$. We thus obtain a policy $\sigma'_{\boldsymbol{X}} \in \Pi'$ such that $E_{\sigma_{\boldsymbol{X}}}[Y] = E_{\boldsymbol{\pi}'}[Y]$.

We now consider the case where $X \in An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$. By basic probabilistic operations,

$$P_{\sigma_{\boldsymbol{X}}}(y) = \sum_{h_X, x} P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X) \sigma_X(x|h_X) P_{\sigma_{\boldsymbol{X} \setminus X}, x}(y|h_X). \tag{10}$$

Since $\tilde{\boldsymbol{S}} \mapsto X$ are irrelevant,

$$(Y \perp\!\!\!\perp \tilde{\boldsymbol{S}} | H_{X^+} \setminus \tilde{\boldsymbol{S}})_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}.$$

Since H_X are all parent nodes of X in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$, the above independence relationship is equivalent to

$$(Y \perp\!\!\!\perp \tilde{\boldsymbol{S}} | H_X \setminus \tilde{\boldsymbol{S}})_{\mathcal{G}_{\sigma_{\boldsymbol{X}} \overline{X}}}.$$

By Rule 1 of (Correa & Bareinboim, 2020, Thm. 1), this relationship implies that:

$$P_{\sigma_{\boldsymbol{X} \setminus X}, x}(y|h_X) = P_{\sigma_{\boldsymbol{X} \setminus X}, x}(y|h_X \setminus \tilde{\boldsymbol{s}}). \tag{11}$$

Eqs. (10) and (11) together gives

$$\begin{aligned} P_{\sigma_{\boldsymbol{X}}}(y) = & \sum_{h_X \setminus \tilde{\boldsymbol{s}}, x} P_{\sigma_{\boldsymbol{X} \setminus X}, x}(y|h_X \setminus \tilde{\boldsymbol{s}}) \\ & \cdot \sum_{\tilde{\boldsymbol{s}}} P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X) \sigma_X(x|h_X) \\ = & \sum_{h_X \setminus \tilde{\boldsymbol{s}}, x} P_{\sigma_{\boldsymbol{X} \setminus X}, x}(y|h_X \setminus \tilde{\boldsymbol{s}}) \\ & \cdot P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X \setminus \tilde{\boldsymbol{s}}) \sigma'_X(x|h_X \setminus \tilde{\boldsymbol{s}}). \end{aligned} \tag{12}$$

where $\sigma'_X(x|h_X \setminus \tilde{\boldsymbol{s}})$ is a function given by:

$$\sigma'_X(x|h_X \setminus \tilde{\boldsymbol{s}}) = \frac{\sum_{\tilde{\boldsymbol{s}}} P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X) \sigma_X(x|h_X)}{P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X \setminus \tilde{\boldsymbol{s}})}.$$

Since X is not an ancestor of H_X in $\mathcal{G}_{\sigma_{\boldsymbol{X}} \overline{X}}$, $P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X)$ is not a function of x. Therefore,

$$\begin{aligned} \sum_x \sigma'_X(x|h_X \setminus \tilde{\boldsymbol{s}}) &= \sum_x \frac{\sum_{\tilde{\boldsymbol{s}}} P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X) \sigma_X(x|h_X)}{P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X \setminus \tilde{\boldsymbol{s}})} \\ &= \frac{\sum_{\tilde{\boldsymbol{s}}} P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X) \sum_x \sigma_X(x|h_X)}{P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X \setminus \tilde{\boldsymbol{s}})} \\ &= \frac{\sum_{\tilde{\boldsymbol{s}}} P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X)}{P_{\sigma_{\boldsymbol{X} \setminus X}, x}(h_X \setminus \tilde{\boldsymbol{s}})} = 1. \end{aligned}$$

Therefore, σ'_X is a decision rule in the probabilistic space of $\{\mathfrak{D}_{H_X \setminus \tilde{\boldsymbol{S}}} \mapsto \mathfrak{D}_X\}$. Let $\sigma'_{\boldsymbol{X}} = \{\sigma_{\boldsymbol{X} \setminus X}, \sigma'_X\}$. Eq. (12) implies

$$P_{\sigma_{\boldsymbol{X}}}(y) = P_{\sigma'_{\boldsymbol{X}}}(y),$$

which completes the proof. □

Lemma 3. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, a reduction Π' of the policy space Π is soluble if Π is soluble.*

Proof. Let $\prec$ denote the a total ordering over $\boldsymbol{X}$ induced by the soluble ordering of Π. We first show that $\prec$ is preserved under reduction, and first the removal of irrelevant evidences $\tilde{\boldsymbol{S}} \mapsto X$. For any $X_j \in \boldsymbol{X}$, if $X_j \neq X$, since d-separation is preserved under edge removal, for any $X_i \prec X_j$,

$$(\sigma_{X_i} \perp\!\!\!\perp \{Y\} \cap De(X_j) | H_{X_j^+})_{\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}}.$$

Consider the case that $X_j = X$. Since $\tilde{\boldsymbol{S}}$ is irrelevant for X_j, by definitions, we have

$$(\{Y\} \cap De(X_j) \perp\!\!\!\perp \tilde{\boldsymbol{S}} | H_{X_j^+} \setminus \tilde{\boldsymbol{S}})_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}.$$

Since $\prec$ is a soluble ordering, for any $X_i \prec X_j$,

$$(\{Y\} \cap De(X_j) \perp\!\!\!\perp \sigma_{X_i} | H_{X_j^+})_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}.$$

By the contraction axiom (Pearl, 2000, Ch. 1.1.5),

$$(\{Y\} \cap De(X_j) \perp\!\!\!\perp \sigma_{X_i}, \tilde{\boldsymbol{S}} | H_{X_j^+} \setminus \tilde{\boldsymbol{S}})_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}.$$

which implies

$$(\{Y\} \cap De(X_j) \perp\!\!\!\perp \sigma_{X_i} | H_{X_j^+} \setminus \tilde{\boldsymbol{S}})_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}.$$

Since d-separation is preserved under edge removal, the above independence also holds in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$. That is, the total ordering $\prec$ is preserved.

We now consider the case where Π' is a reduction of Π obtained by removing irrelevant treatments $\tilde{\boldsymbol{X}} = \boldsymbol{X} \setminus An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$. That is, for any $\sigma_{\boldsymbol{X}} \in \Pi$, $\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{X}}} \in \Pi'$. By definitions, for a soluble ordering $\prec$, for any $X_i \prec X_j$,

$$(\{Y\} \cap De(X_j) \perp\!\!\!\perp \sigma_{X_i} | H_{X_j^+})_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}.$$

If $X_i \in \tilde{\boldsymbol{X}}$, by Lem. 6, $X_i \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{X}}}}}$. The above relationship is preserved in $\mathcal{G}_{\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{X}}}}$. It thus suffices to focus on the settings where $X_i \notin \tilde{\boldsymbol{X}}$.

For any $X_j \notin \tilde{\boldsymbol{X}}$, by definitions, $\tilde{\boldsymbol{X}}$ must not contain any ancestor of $\{H_{X_j}, X_j, Y\}$ in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$. That is,

$$(\{Y, H_{X_j}, X_j\} \perp\!\!\!\perp \tilde{\boldsymbol{X}})_{\mathcal{G}_{\sigma_{\boldsymbol{X}}\overline{\tilde{X}}}}. \tag{13}$$

Similarly, by Lem. 6, we have

$$(\{Y, H_{X_j}, X_j\} \perp\!\!\!\perp \tilde{\boldsymbol{X}})_{\mathcal{G}_{\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{X}}}\overline{\tilde{X}}}}. \tag{14}$$

By Rules 3 of (Correa & Bareinboim, 2020, Thm. 1), Eqs. (13) and (14) imply that

$$P_{\sigma_{\boldsymbol{X}}}(y|h_{X_j}, x_j) = P_{\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{X}}}}(y|h_{X_j}, x_j). \tag{15}$$

Since $\prec$ is a soluble ordering, for any $X_i \prec X_j$,

$$(\{Y\} \cap De(X_j) \perp\!\!\!\perp \sigma_{X_i} | H_{X_j^+})_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}.$$

By (Koller & Milch, 2003, Lem. 5.2) (which can be seen as the combination of Rules 2 and 3 in (Correa & Bareinboim, 2020, Thm. 1)), we have for any $\sigma_{\boldsymbol{X}} \in \Pi$ and any decision rule $\sigma'_{X_i} \in \{\mathcal{D}_{H_{X_i}} \mapsto \mathcal{D}_{X_i}\}$,

$$P_{\sigma_{\boldsymbol{X}}}(y|h_{X_j}, x_j) = P_{\sigma_{\boldsymbol{X}\setminus\{X_i\}}, \sigma'_{X_i}}(y|h_{X_j}, x_j). \tag{16}$$

Eqs. (15) and (16) imply that for any $\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{X}}} \in \Pi'$ and any $\sigma'_{X_i} \in \{\mathcal{D}_{H_{X_i}} \mapsto \mathcal{D}_{X_i}\}$,

$$P_{\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{x}}}}(y|h_{X_j}, x_j) = P_{\sigma_{\boldsymbol{X}\setminus(\tilde{\boldsymbol{x}}\cup\{X_i\})}, \sigma'_{X_i}}(y|h_{X_j}, x_j).$$

in any SCM M conforming to G. By the completeness of d-separation, for any treatment $X_i \prec X_j$ in $\mathcal{G}_{\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{x}}}}$

$$(\{Y\} \cap De(X_j) \perp\!\!\!\perp \sigma_{X_i} | H_{X_j^+})_{\mathcal{G}_{\sigma_{\boldsymbol{X}\setminus\tilde{\boldsymbol{x}}}}}.$$

It is now sufficient to show that $\prec$ does not violate the topological ordering in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$. If Π' is a reduction obtained from Π by removing irrelevant evidences, a topological ordering in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ is preserved under edge removal. Therefore, Π' is soluble.

Consider now Π' is a reduction obtained from Π by removing irrelevant treatments $\tilde{\boldsymbol{X}} = \boldsymbol{X} \setminus An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$. Suppose there exists a pair $X_i, X_j \in (\boldsymbol{X} \setminus \tilde{\boldsymbol{X}})$ such that $X_i \prec X_j$ and $X_j \in An(X_i)_{\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}}$. Let g be a causal path from X_j to X_i in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$. Since $\prec$ is a topological ordering in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$, $X_j \notin An(X_i)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$. Path g must contains an incoming edge $V_i \to X'$ for some $X' \in \tilde{\boldsymbol{X}}$. Let X' be the last such treatment node on g. By definitions, $X_i \in An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$. We could thus obtain from g a causal path g' from X' to Y. That is, $X' \in An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, which is a contradiction.

This means that $\prec$ respects the ancestral relationships among $\boldsymbol{X} \setminus \tilde{\boldsymbol{X}}$ in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$. Since $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$ is a DAG (Lem. 7), there must exist a topological ordering in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$ compatible with $\prec$, which proves the statement. □

Theorem 2. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, Reduce returns the minimal reduction Π_{MIN} of a soluble policy space Π.*

Proof. By the graphoid axioms of contraction and weak unions (Pearl, 2000, Ch. 1.1.5), it is verifiable that the reduction Π after Step 7 has no irrelevant evidences. By definitions, for any treatment $X \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, all of its evidences are irrelevant. That is, the manipulated graph $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$ coincides with the subgraph $\mathcal{G}_{\sigma_{\boldsymbol{X}}\overline{\tilde{\boldsymbol{X}}}}$ where $\tilde{\boldsymbol{X}} = \boldsymbol{X} \setminus (\boldsymbol{X} \cap An(Y))_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$. Therefore, removing irrelevant treatments $\tilde{\boldsymbol{X}}$ only adds arrows into $\tilde{\boldsymbol{X}}$ in the graph $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$. Since adding arrows into a DAG does not introduce independence, $\Pi \setminus \tilde{\boldsymbol{X}}$ has no irrelevant evidence. That is, Reduce(G, Π, Y) returns the minimal reduction Π_{MIN}. □

Proof of Theorem 1

In this section, we will provide proofs for the uniqueness of the minimal reduction. We first define the stepwise reduction, which searches through the space of reductions in a sequential, stepwise fashion.

Definition 7. Given $[\![\mathcal{G}, \Pi, Y]\!]$, a policy space Π' is a stepwise reduction of Π if it is obtainable from Π by successively applying the following operations:

1. $\Pi' = \Pi \setminus \{X\}$ where X is a treatment in $\boldsymbol{X}$ such that $X \notin \boldsymbol{X} \cap An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$.
2. $\Pi' = \Pi \setminus \{S \mapsto X\}$ where S is an evidence in H_X for a treatment X such that $(\{Y\} \cap De(X) \perp\!\!\!\perp \{S\} | H_{X^+} \setminus \{S\})_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$.

Similarly, unless it is explicitly specified, we denote by $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$ the manipulated diagram of a stepwise reduction Π'

obtained from $[\![\mathcal{G}, \Pi, Y]\!]$. We also define the minimal stepwise reduction as one that does not contain any irrelevant treatment and evidence.

Definition 8. Given $[\![\mathcal{G}, \Pi, Y]\!]$, a stepwise reduction $\Pi_{\text{S-MIN}}$ of Π is *minimal* if it has no stepwise reduction.

The operation of stepwise reduction have some interesting properties, and first, the preservation of irrelevant treatments and evidences.

Lemma 8. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, let Π' be a stepwise reduction of Π. For any treatment $X \in \boldsymbol{X}$, if $X \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, then $X \notin An(Y)_{\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}}$.*

Proof. Suppose Π' is a stepwise reduction obtained by removing irrelevant some treatments $\tilde{\boldsymbol{X}} \subseteq \boldsymbol{X} \setminus An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$. The proof follows immediately from Lem. 6.

If Π' is a stepwise reduction of Π obtained by removing irrelevant evidences. If $X \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, X is not an ancestor of Y in any subgraph of $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$, i.e., $X \notin An(Y)_{\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}}$. □

Lemma 9. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, let $\Pi' = \{\mathcal{D}_{H'_X} \mapsto \mathcal{D}_X : \forall X \in \boldsymbol{X}'\}$ be a stepwise reduction of Π. For any $X \in \boldsymbol{X}'$, any evidence $S \in H_X$, if $(\{Y\} \cap De(X) \perp\!\!\!\perp S | H_{X+} \setminus S)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, then $(\{Y\} \cap De(X) \perp\!\!\!\perp S | H'_{X+} \setminus S)_{\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}}$.*

Proof. Suppose that Π' is a stepwise reduction of Π obtained by removing irrelevant evidences. It follows from (Lauritzen & Nilsson, 2001, Lem. 7) that an irrelevant evidence is preserved by removing other irrelevant evidences.

We now consider the case where Π' is a stepwise reduction obtained by removing an irrelevant treatment $X' \in (\boldsymbol{X} \setminus An(Y))_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$; therefore, $\boldsymbol{X}' = \boldsymbol{X} \setminus \{X'\}$ and $H'_X = H_X$ for any $X \in \boldsymbol{X}'$. If $X \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, Lem. 6 implies that $X \notin An(Y)_{\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}}$. Therefore, the following independence relationship trivially holds.

$$(\{Y\} \cap De(X) \perp\!\!\!\perp H | H'_{X+} \setminus H)_{\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}}.$$

Suppose now $X \in An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$. Since $X' \notin An(Y)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, Lem. 6 implies that implies that $X \in An(Y)_{\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}}$ and $X' \notin An(Y)_{\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}}$. This implies that H_X and X are non-descendants of X in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$.

Since $(\{Y\} \cap De(X) \perp\!\!\!\perp H | H_{X+} \setminus H)_{\mathcal{G}_{\sigma_{\boldsymbol{X}}}}$, the path connecting H to Y given $H_{X+} \setminus H$ in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$ must be due to the change of incoming arrows into X'. If changing incoming arrows into X' opens a path containing $V_1 \rightarrow V \leftarrow V_2$ where $V \in An(X')_{\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}}$, there must exists a causal path from X' to a node in $H_{X+} \setminus H$. That is X' is an ancestor for a node in H_X, X, which is a contradiction.

Suppose now changing incoming arrows into X' opens a path containing $V_1 \leftarrow X' \leftarrow V_2$ in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$. By definitions of d-separation, there must exist a causal path from X' to a node in H_X, X, Y in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$. Since H_X, X, Y are non-descendants of X in $\mathcal{G}_{\sigma'_{\boldsymbol{X}'}}$, we have a contradiction, which completes the proof. □

Lems. 8 and 9 imply that for any reduction operation, one could simulate it through a series of stepwise reduction. Therefore, we could attain any reduction of the policy space Π through equivalent stepwise reductions.

Lemma 10. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, any reduction Π' of Π is a stepwise reduction of Π; any minimal reduction Π_{MIN} of Π is a minimal stepwise reduction of Π.*

Proof. Lems. 8 and 9 imply that any reduction of a policy space Π could be performed stepwise. That is, any reduction of Π is also a stepwise reduction. Since the minimal condition of reduction and stepwise reduction are equivalent, any minimal reduction Π_{MIN} of Π has no stepwise reduction. □

Since any minimal reduction of Π is also a minimal stepwise reduction, the set of all possible minimal stepwise reductions of Π must contain all minimal reductions of Π. If the minimal stepwise reduction is unique, then Π has at most one minimal reduction. For any two policy spaces $\Pi_1 = \{\mathcal{D}_{H^1_X} \mapsto \mathcal{D}_X : \forall X \in \boldsymbol{X}_1\}$ and $\Pi_2 = \{\mathcal{D}_{H^2_X} \mapsto \mathcal{D}_X : \forall X \in \boldsymbol{X}_2\}$, we define their intersection $\Pi_1 \cap \Pi_2$ as a policy space $\{\mathcal{D}_{H^1_X \cap H^2_X} \mapsto \mathcal{D}_X : \forall X \in \boldsymbol{X}_1 \cap \boldsymbol{X}_2\}$. The following results establishes the uniqueness of minimal stepwise reduction.

Lemma 11. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, let Π_1 and Π_2 be two stepwise reductions of Π. Then $\Pi_1 \cap \Pi_2$ is a stepwise reduction of both Π_1 and Π_2.*

Proof. Let m_1, m_2 be the number of reduction steps required to obtain Π_1 and Π_2 from Π respectively. We will show the results by induction after $m = m_1 + m_2$.

For $m = 2$, the result follows directly from Lems. 8 and 9. Suppose the result holds for $m \leq k$, where $k \geq 2$ and consider the case $m = k + 1$. So $\max\{m_1, m_2\} > 1$, say $m_2 > 1$. Thus Π_1 is obtained by successively removing m_2 irrelevant treatments or evidences from Π. Let Π'_2 be the stepwise reduction obtained by removing the first $m_2 - 1$ of these. By the induction assumption, $\Pi_1 \cap \Pi'_2$ is a stepwise reduction of Π'_2 obtained by moving at most m_1 steps from Π_2. Furthermore, Π_2 is also a stepwise reduction of Π'_2 obtained by removing exactly one irrelevant treatment or evidence. Since $(\Pi_1 \cap \Pi'_2) \cap \Pi_2 = \Pi_1 \cap \Pi_2$ and $m_1 + 1 \leq k$, the induction assumptions yields that $\Pi_1 \cap \Pi_2$ is a stepwise reduction of Π_2.

Similarly, the induction assumptions gives $\Pi_1 \cap \Pi_2$ is a stepwise reduction of $\Pi_1 \cap \Pi'_2$ and also that $\Pi_1 \cap \Pi'_2$ is a stepwise reduction of Π_1. By definitions, $\Pi_1 \cap \Pi_2$ is a stepwise reduction of Π and the proof is complete. □

Lemma 12. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, there exists a unique minimal stepwise reduction $\Pi_{\text{S-MIN}}$ of Π.*

Proof. Suppose there exists two different minimal stepwise reduction Π_1 and Π_2. Lem. 10 implies that $\Pi_1 \cap \Pi_2$ is reduction of both Π_1 and Π_2, which is a contradiction. □

Finally, we are ready to prove the uniqueness of the minimal reduction of a policy space.

Theorem 1. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, there exists a unique minimal reduction Π_{MIN} of policy space Π.*

Proof. By Lem. 10, any minimal reduction Π_{MIN} of Π is also a minimal stepwise reduction $\Pi_{\text{S-MIN}}$. Since $\Pi_{\text{S-MIN}}$ is unique (Lem. 12), there exists at most one minimal reduction Π_{MIN}. Since Π_{MIN} is well defined from $[\![\mathcal{G}, \Pi, Y]\!]$, Π must have a unique minimal reduction. □

Appendix B. Proofs of Results in Section 3

In this section, we provide proofs for the results presented in Sec. 3. We will use the notation in (Tian, 2002) and define function $Q[\boldsymbol{S}](\boldsymbol{v}) = P_{\boldsymbol{v} \setminus \boldsymbol{s}}(\boldsymbol{s})$ for an arbitrary subset $\boldsymbol{S} \subseteq \boldsymbol{V}$. Naturally, $Q[\boldsymbol{V}](\boldsymbol{v}) = P(\boldsymbol{v})$ and $Q[\emptyset](\boldsymbol{v}) = 1$. For convenience, we often omit input $\boldsymbol{v}$ and write $Q[\boldsymbol{S}]$.

Corollary 1. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, for any $S_k \in \boldsymbol{S}$, let $\bar{\boldsymbol{S}}_k$ denote a c-component in $\mathcal{G}_{[\boldsymbol{S}^{(k)}]}$ that contains S_k and let $\bar{\boldsymbol{X}}_k = Pa(\bar{\boldsymbol{S}}_k)_{\mathcal{G}} \setminus \bar{\boldsymbol{S}}_k$. $P_{\boldsymbol{x}}(\boldsymbol{s})$ could be written as:*

$$P_{\boldsymbol{x}}(\boldsymbol{s}) = \prod_{S_k \in \boldsymbol{S}} P_{\bar{\boldsymbol{x}}_k}(s_k | \bar{\boldsymbol{s}}_k \setminus \{s_k\}). \tag{17}$$

Proof. Since $\boldsymbol{S}$ are ordered following a topological ordering $\prec$, $S_k \notin An(\bar{\boldsymbol{S}}_{k-1})_{\mathcal{G}}$ for any S_k. By (Tian, 2002, Lemma 10), we have

$$Q[\boldsymbol{S}^{(k)}] = \sum_{s_k} Q[\boldsymbol{S}^{(k)}].$$

$P_{\boldsymbol{x}}(\boldsymbol{s})$ could thus be written as:

$$P_{\boldsymbol{x}}(\boldsymbol{s}) = \prod_{S_k \in \boldsymbol{S}} \frac{Q[\boldsymbol{S}^{(k)}]}{\sum_{s_k} Q[\boldsymbol{S}^{(k)}]}. \tag{18}$$

Let $\boldsymbol{C}^k_1, \dots, \boldsymbol{C}^k_l$ denote c-components in $\mathcal{G}_{[\boldsymbol{S}^{(k)}]}$ and let $\boldsymbol{C}^k_1$ be the c-component that contains S_k; therefore, $\bar{\boldsymbol{S}}_k = \boldsymbol{C}^k_1$. (Tian, 2002, Lem. 11) implies that

$$Q[\boldsymbol{S}^{(k)}] = \prod_{i=1,\dots,l} Q[\boldsymbol{C}^k_i]. \tag{19}$$

Sine $S_k \notin Pa(\boldsymbol{C}^k_i)_{\mathcal{G}}$ for any $i = 2, \dots, l$,

$$\sum_{s_k} Q[\boldsymbol{S}^{(k)}] = \sum_{s_k} Q[\boldsymbol{C}^k_1] \prod_{i=2,\dots,l} Q[\boldsymbol{C}^k_i].$$

The above equation, together with Eqs. (18) and (19), implies

$$P_{\boldsymbol{x}}(\boldsymbol{s}) = \prod_{S_k \in \boldsymbol{S}} \frac{Q[\boldsymbol{C}^k_1]}{\sum_{s_k} Q[\boldsymbol{C}^k_1]}.$$

By definitions, $\bar{\boldsymbol{S}}_k = \boldsymbol{C}^k_1$ and $Q[\boldsymbol{C}^k_1] = P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k)$, which complete the proof. □

Lemma 4. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, for any $S_k \in \boldsymbol{S}$ and any $\sigma_{\boldsymbol{X}} \in \Pi$, $P_{\sigma_{\boldsymbol{X}}}(s_k | \bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\}) = P_{\bar{\boldsymbol{x}}_k}(s_k | \bar{\boldsymbol{s}}_k \setminus \{s_k\})$.*

Proof. By Corol. 1 and basic probabilistic properties,

$$P_{\sigma_{\boldsymbol{X}}}(\boldsymbol{s}, \boldsymbol{x}) = \sum_{\boldsymbol{s}, \boldsymbol{x}} \prod_{S_k \in \boldsymbol{S}} P_{\bar{\boldsymbol{x}}_k}(s_k | \bar{\boldsymbol{s}}_k \setminus \{s_k\}) \prod_{X \in \boldsymbol{X}} \sigma_X(x | h_X).$$

Let $\prec$ be a solution ordering in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$. Marginalizing variables in $(\boldsymbol{S} \cup \boldsymbol{X}) \setminus (\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k)$ according to a reverse ordering relative to $\prec$ gives:

$$P_{\sigma_{\boldsymbol{X}}}(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k) = P_{\bar{\boldsymbol{x}}_k}(s_k | \bar{\boldsymbol{s}}_k \setminus \{s_k\}) P_{\sigma_{\boldsymbol{X}}}(\bar{\boldsymbol{s}}_k \setminus \{s_k\}, \bar{\boldsymbol{x}}_k).$$

The above equation implies that

$$P_{\sigma_{\boldsymbol{X}}}(s_k | \bar{\boldsymbol{s}}_k \setminus \{s_k\}, \bar{\boldsymbol{x}}_k) = P_{\bar{\boldsymbol{x}}_k}(s_k | \bar{\boldsymbol{s}}_k \setminus \{s_k\})$$

for any $\sigma_{\boldsymbol{X}} \in \Pi$, which completes the proof. □

Proof of Theorem 3

We begin by introducing some necessary lemmas. We first show that the confidence set $\mathcal{P}_t$ contains the actual interventional distribution $P_{\boldsymbol{x}}(\boldsymbol{s})$ with high probabilities.

Lemma 13. *Fix $\delta \in (0, 1)$, for any $t \geq 1$, with probability (w.p.) at least $1 - \delta/(4t^2)$, $P_{\boldsymbol{x}}(\boldsymbol{s}) \in \mathcal{P}_t$.*

Proof. Fix $n^t(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\})$ in $\{1, \dots, t-1\}$. Since

$$\sqrt{\frac{2 \log(2^{|\mathcal{D}_{S_k}|} 4t^3 |\boldsymbol{S}| |\mathcal{D}_{(\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k) \setminus \{S_k\}}| / \delta)}{\max\left\{n^t(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\}), 1\right\}}} \leq f_{S_k}(t, \delta)$$

where $f_{S_k}(t, \delta)$ is a function defined as

$$f_{S_k}(t, \delta) = \sqrt{\frac{6 |\mathcal{D}_{S_k}| \log(2 |\boldsymbol{S}| |\mathcal{D}_{(\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k) \setminus \{S_k\}}| t / \delta)}{\max\left\{n^t(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\}), 1\right\}}}$$

By the concentration inequality of (Jaksch et al., 2010, C.1), we have for any $S_k \in \boldsymbol{S}$, any $\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\}$,

$$\left\| P_{\bar{\boldsymbol{x}}_k}(\cdot | \bar{\boldsymbol{s}}_k \setminus \{s_k\}) - \hat{P}^t_{\bar{\boldsymbol{x}}_k}(\cdot | \bar{\boldsymbol{s}}_k \setminus \{s_k\}) \right\|_1 > f_{S_k}(t, \delta). \tag{20}$$

with probability at most $\delta/(4t^3|\boldsymbol{S}||D_{(\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|)$.

Hence a union bound over all possible values of $n^t(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\}) = 1, \dots, t-1$ implies that Eq. (20) holds for any $n^t(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\})$ with probability at most

$$\sum_{n=1}^{t-1} \frac{\delta}{4t^3|\boldsymbol{S}||D_{(\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|} = \frac{\delta}{4t^2|\boldsymbol{S}||D_{(\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|}.$$

Summing these error probabilities over state-action pairs $D_{(\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k)\setminus\{S_k\}}$ for all $S_k \in \boldsymbol{S}$ gives:

$$P(P_{\boldsymbol{x}}(\boldsymbol{s}) \notin \mathfrak{P}_t) \leq \frac{\delta}{4t^2}. \qquad \square$$

Lemma 14. *Fix $\delta \in (0,1)$. With probabilities (w.p.) at least $1 - \frac{\delta}{2}$, for all $t = 1, 2, \dots$, $V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) \geq E_{\sigma^*_{\boldsymbol{X}}}[Y]$.*

Proof. Since

$$\sum_{t=1}^{\infty} \frac{1}{4t^2} \leq \frac{\pi^2}{24}\delta < \frac{\delta}{2},$$

it follows from Lem. 13 that with probability at least $1 - \frac{\delta}{2}$, $P_{\boldsymbol{x}}(\boldsymbol{s}) \in \mathfrak{P}_t$ for all episodes $t = 1, 2, \dots$.

By definitions, $\sigma^t_{\boldsymbol{X}}$ is the optimal policy for the instance $P^t_{\boldsymbol{x}}(\boldsymbol{s})$ in $\mathfrak{P}_t$ that has the maximal optimal expected outcome. This implies that

$$V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) \geq V_{\sigma^*_{\boldsymbol{X}}}(P_{\boldsymbol{x}}(\boldsymbol{s})) = E_{\sigma^*_{\boldsymbol{X}}}[Y]. \qquad \square$$

Lemma 15. *Fix $\delta \in (0,1)$. W.p. at least $1 - \frac{\delta}{2}$, for any $T > 1$,*

$$\sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t \leq 2|\boldsymbol{S}|\sqrt{T\log(2|\boldsymbol{S}|T/\delta)} + \sum_{S_k\in\boldsymbol{S}} 12\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k}|T\log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|T/\delta)}.$$

Proof. For simplicity, let $\boldsymbol{V} = \boldsymbol{S} \cup \boldsymbol{X}$. For a solution ordering $\prec$ in $\mathcal{G}_{\sigma_{\boldsymbol{X}}}$, let variables in $\boldsymbol{V}$ be ordered by $V_1 \prec \cdots \prec V_{n+m}$. For any policy $\sigma_{\boldsymbol{X}} \in \Pi$ and any $i = 0, 1, \dots, m+n$, we define function $V_{\sigma_{\boldsymbol{X}}}(\boldsymbol{v}^{(i)}; P_{\boldsymbol{x}}(\boldsymbol{s}))$ as following:

$$V_{\sigma_{\boldsymbol{X}}}(\boldsymbol{v}^{(i)}; P_{\boldsymbol{x}}(\boldsymbol{s})) = \frac{\sum_{v\notin\boldsymbol{v}^{(i)}} E_{\boldsymbol{x}}[Y|\boldsymbol{s}]P_{\boldsymbol{x}}(\boldsymbol{s})\prod_{X\in\boldsymbol{X}}\sigma_X(x|h_X)}{\sum_{v\notin\boldsymbol{v}^{(i)}} P_{\boldsymbol{x}}(\boldsymbol{s})\prod_{X\in\boldsymbol{X}}\sigma_X(x|h_X)}$$

Naturally, we have

$$V_{\sigma_{\boldsymbol{X}}}(\boldsymbol{v}; P_{\boldsymbol{x}}(\boldsymbol{s})) = E_{\boldsymbol{x}}[Y|\boldsymbol{s}].$$

We can decompose $V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t$ as a telescoping sum:

$$V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t = \sum_{V_i\in\boldsymbol{V}} V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})). \tag{21}$$

It is a well-known fact in decision theory that no stochastic policy can improve on the utility of the best deterministic policy (see, e.g., (Liu & Ihler, 2012, Lem. 2.1)). This means that the policy $\sigma^t_{\boldsymbol{X}}$ must be deterministic. We have for any $V_i \in \boldsymbol{X}$,

$$V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) = 0$$

The above equation allows to write Eq. (21) as:

$$V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t = \sum_{V_i\in\boldsymbol{S}} V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})). \tag{22}$$

By Corol. 1,

$$P^t_{\boldsymbol{x}}(\boldsymbol{s}) = \prod_{S_k\in\boldsymbol{S}} P^t_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\}).$$

For any $V_i \in \boldsymbol{S}$, we denote by $V_i = S_k$. Let $P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})$ denote a distribution obtained from $P^t_{\boldsymbol{x}}(\boldsymbol{s})$ such that its associated distribution $P^t_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$ is replaced with the actual $P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$, i.e.,

$$P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s}) = P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) \cdot \prod_{S_j\neq V_i} P^t_{\bar{\boldsymbol{x}}_j}(s_j|\bar{\boldsymbol{s}}_j \setminus \{s_j\}). \tag{23}$$

We could further decompose $V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s}))$ as follows:

$$\begin{aligned} &V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) \\ &= V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) \qquad (24) \\ &+ V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})). \end{aligned}$$

Eqs. (21), (22) and (27) together imply:

$$\sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t$$

$$= \sum_{V_i\in\boldsymbol{S}}\sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) \tag{25}$$

$$+ \sum_{V_i\in\boldsymbol{S}}\sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})). \tag{26}$$

Bounding Eq. (25) For $V_i \in \boldsymbol{S}$, we denote by $V_i = S_k$. By basic probabilistic operations,

$$\begin{aligned}
&V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) \\
&\leq \left\| P^t_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) - P_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) \right\|_1 \\
&\cdot \max_{s_k} \left\{ V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) \right\} \\
&\leq 2\sqrt{\frac{6|\mathcal{D}_{S_k}| \log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k) \setminus \{S_k\}}|t/\delta)}{\max\left\{n^t(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\}), 1\right\}}}
\end{aligned}$$

Following the result in (Jaksch et al., 2010, C.3),

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{1}{\sqrt{\max\left\{n^t(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\}), 1\right\}}} \\
&\leq \sum_{\bar{\boldsymbol{x}}_k} \sum_{\bar{\boldsymbol{s}}_k \setminus \{s_k\}} (\sqrt{2}+1)\sqrt{n^t(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\})}
\end{aligned}$$

By Jensen's inequality we thus have

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{1}{\sqrt{\max\left\{n^t(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\}), 1\right\}}} \\
&\leq (\sqrt{2}+1)\sqrt{|\mathcal{D}_{(\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k) \setminus \{S_k\}}|T},
\end{aligned} \tag{27}$$

which gives

$$\begin{aligned}
&\sum_{V_i \in \boldsymbol{S}} \sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) \\
&\leq \sum_{S_k \in \boldsymbol{S}} 12\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}|T \log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k) \setminus \{S_k\}}|T/\delta)}
\end{aligned} \tag{28}$$

Bounding Eq. (26) For any $V_i \in \boldsymbol{S}$, we define

$$Z_t(V_i) = V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})).$$

Let the sampling history up to episode t be denoted by $\mathcal{H}_t = \{\boldsymbol{X}^i, \boldsymbol{S}^i\}_{i=1}^{t-1}$. Since $|Z_t(V_i)| \leq 1$ and $E[Z_{t+1}(V_i)|\mathcal{H}_t] = 0$, $\{Z_t(V_i) : t = 1, \dots, T\}$ is thus a sequence of martingale differences. By Azuma-Hoeffding inequality, we have that for all $V_i \in \boldsymbol{S}$, with probability at least $\frac{\delta}{4T^2}$,

$$\sum_{t=1}^{T} Z_t(V_i) \leq 2\sqrt{T \log(2|\boldsymbol{S}|T/\delta)}.$$

Since $\sum_{T=1}^{\infty} \frac{1}{4T^2} \leq \frac{\pi^2}{24}\delta < \frac{\delta}{2}$, it follows that with probability at least $1 - \frac{\delta}{2}$,

$$\begin{aligned}
&\sum_{V_i \in \boldsymbol{S}} \sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) \\
&\leq 2|\boldsymbol{S}|\sqrt{T \log(2|\boldsymbol{S}|T/\delta)}.
\end{aligned} \tag{29}$$

Bounding Eqs. (25) and (26) with Eqs. (28) and (29) proves the statement. □

Theorem 3. *Given $[\![G, \Pi, Y]\!]$, fix a $\delta \in (0, 1)$. With probability (w.p.) at least $1 - \delta$, it holds for any $T > 1$, the regret of* `OFU-DTR` *is bounded by*

$$R(T, M^*) \leq \Delta(T, \delta) + 2|\boldsymbol{S}|\sqrt{T \log(2|\boldsymbol{S}|T/\delta)},$$

where $\Delta(T, \delta)$ is a function defined as

$$\Delta(T, \delta) = \sum_{S_k \in \boldsymbol{S}} 17\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}|T \log(|\boldsymbol{S}|T/\delta)}.$$

Proof. Suppose

$$T \leq \sum_{S_k \in \boldsymbol{S}} 17^2 |\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}| \log(|\boldsymbol{S}|T/\delta).$$

Since $R(T, M^*) \leq T = (\sqrt{T})^2$, the above equation implies that

$$\begin{aligned}
R(T, M^*) &\leq 17\sqrt{\sum_{S_k \in \boldsymbol{S}} |\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}|T \log(|\boldsymbol{S}|T/\delta)} \\
&\leq \sum_{S_k \in \boldsymbol{S}} 17\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}|T \log(|\boldsymbol{S}|T/\delta)} \\
&= \Delta(T, \delta).
\end{aligned}$$

We now consider the case where

$$T > \sum_{S_k \in \boldsymbol{S}} 17^2 |\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}| \log(|\boldsymbol{S}|T/\delta). \tag{30}$$

Lems. 14 and 15 together imply that with probability at least $1 - \frac{\delta}{2} - \frac{\delta}{2} = 1 - \delta$, for any $T > 1$,

$$\begin{aligned}
&R(T, M^*) \leq \sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t \\
&\leq 2|\boldsymbol{S}|\sqrt{T \log(2|\boldsymbol{S}|T/\delta)} \\
&+ \sum_{S_k \in \boldsymbol{S}} 12\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}|T \log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k) \setminus \{S_k\}}|T/\delta)}.
\end{aligned}$$

Whenever Eq. (30) holds,

$$\log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k) \setminus \{S_k\}}|T/\delta) \leq 2\log(|\boldsymbol{S}|T/\delta).$$

We thus have

$$\begin{aligned}
R(T, M^*) &\leq 2|\boldsymbol{S}|\sqrt{T \log(2|\boldsymbol{S}|T/\delta)} \\
&+ \sum_{S_k \in \boldsymbol{S}} 12\sqrt{2|\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}|T \log(|\boldsymbol{S}|T/\delta)} \\
&\leq \Delta(T, \delta) + 2|\boldsymbol{S}|\sqrt{T \log(2|\boldsymbol{S}|T/\delta)}.
\end{aligned}$$

□

Proof of Theorem 4

Note that in the Bayesian setting, the actual SCM M^* is drawn from a distribution $\phi^*(M)$ over candidate models in $\mathcal{M}$. We say that ϕ is the prior of $P_{\boldsymbol{x}}(\boldsymbol{s})$ if

$$\phi(\boldsymbol{\theta}) = \sum_{M\in\mathcal{M}} I_{\{P_{M_{\boldsymbol{x}}}(\boldsymbol{s})=\boldsymbol{\theta}\}}\phi^*(M). \tag{31}$$

Before we prove Theorem 4, we first introduce some necessary lemmas.

Lemma 16. *If ϕ satisfies Eq. (31), it holds for any $T > 1$,*

$$\sum_{t=1}^{T} E_{\sigma^*_{\boldsymbol{X}}}[Y] = \sum_{t=1}^{T} E[V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s}))]. \tag{32}$$

Proof. Let the sampling history $\mathcal{H}_t = \{\boldsymbol{X}^i, \boldsymbol{S}^i\}_{i=1}^{t-1}$. Since ϕ satisfies Eq. (31), the actual $P_{\boldsymbol{x}}(\boldsymbol{s})$ and the sampled instance $P^t_{\boldsymbol{x}}(\boldsymbol{s})$ are identically distributed (Osband et al., 2013, Lem. 1). We thus have for any t,

$$\begin{aligned}&E_{\sigma^*_{\boldsymbol{X}}}[Y] - E[V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s}))]\\ &= E[V_{\sigma^*_{\boldsymbol{X}}}(P_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s}))]\\ &= E[E[V_{\sigma^*_{\boldsymbol{X}}}(P_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s}))|\mathcal{H}_t]] = 0,\end{aligned}$$

which proves the statement. □

Lemma 17. *If ϕ satisfies Eq. (31), it holds for any $T > 1$,*

$$\begin{aligned}&\sum_{t=1}^{T} E[V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t] \leq \delta T\\ &+ \sum_{S_k\in\boldsymbol{S}} 12\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k}|T\log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|T/\delta)}\end{aligned}$$

Proof. Since $P_{\boldsymbol{x}}(\boldsymbol{s})$ and $P^t_{\boldsymbol{x}}(\boldsymbol{s})$ are identically distributed given any history $\mathcal{H}_t$, following a similar argument in Lem. 14, we have

$$P(P_{\boldsymbol{x}}(\boldsymbol{s}), P^t_{\boldsymbol{x}}(\boldsymbol{s}) \in \mathcal{P}_t) \geq 1 - \delta.$$

$\sum_{t=1}^{T} E[V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t]$ could thus be written as:

$$\begin{aligned}&\sum_{t=1}^{T} E[V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t] \leq \delta T\\ &+ \sum_{t=1}^{T} E[V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t|P_{\boldsymbol{x}}(\boldsymbol{s}), P^t_{\boldsymbol{x}}(\boldsymbol{s}) \in \mathcal{P}_t].\end{aligned} \tag{33}$$

It thus suffices to bound $\sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t$ under the condition that $P_{\boldsymbol{x}}(\boldsymbol{s}), P^t_{\boldsymbol{x}}(\boldsymbol{s}) \in \mathcal{P}_t$. By Eqs. (25) and (26),

$$\begin{aligned}&\sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t\\ &= \sum_{V_i\in\boldsymbol{S}}\sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s}))\\ &+ \sum_{V_i\in\boldsymbol{S}}\sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})).\end{aligned}$$

By the construction Eq. (23) of $P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})$, we have that for any history $\mathcal{H}_t = \{\boldsymbol{S}^i, \boldsymbol{X}^i\}_{i=1}^{t-1}$,

$$E[V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s}))|\mathcal{H}_t] = 0.$$

By Eq. (28), we also have

$$\begin{aligned}&\sum_{V_i\in\boldsymbol{S}}\sum_{t=1}^{T} V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\sigma^t_{\boldsymbol{X}}}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s}))\\ &\leq \sum_{S_k\in\boldsymbol{S}} 12\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k}|T\log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|T/\delta)}\end{aligned}$$

The above equation, together with Eq. (33), gives

$$\begin{aligned}&\sum_{t=1}^{T} E[V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t] \leq \delta T\\ &+ \sum_{S_k\in\boldsymbol{S}} 12\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k}|T\log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|T/\delta)}.\end{aligned}$$

which proves the statement. □

Theorem 4. *Given $[\![G, \Pi, Y]\!]$ and a prior ϕ, if ϕ satisfies Eq. (31), it holds for any $T > 1$, the regret of* PS-DTR *is bounded by*

$$R(T, \phi^*) \leq \Delta(T, 1/T) + 1,$$

where function $\Delta(T, \delta)$ follows the definition in Thm. 3.

Proof. Lems. 16 and 17 together imply that

$$\begin{aligned}&R(T, \phi^*) = \sum_{t=1}^{T} E[V_{\sigma^t_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t] \leq \delta T\\ &+ \sum_{S_k\in\boldsymbol{S}} 12\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k}|T\log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|T/\delta)}\end{aligned}$$

Following a simplification procedure similar to Thm. 3,

$$R(T, \phi^*) \leq \Delta(T, \delta) + \delta T$$

Fix $\delta = 1/T$, which completes proof. □

Appendix C. Proofs of Results in Section 4

In this section, we provide proofs for causal bounds on transition probabilities. Our proofs build on the notion of counterfactual variables (Pearl, 2000, Ch. 7.1) and axioms of “composition, effectiveness and reversibility” defined in (Pearl, 2000, Ch. 7.3.1).

For a SCM M, arbitrary subsets of endogenous variables $\boldsymbol{X}, \boldsymbol{Y}$, the potential outcome of $\boldsymbol{Y}$ to intervention $do(\boldsymbol{x})$, denoted by $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})$, is the solution for Y with $\boldsymbol{U} = \boldsymbol{u}$ in the sub-model $M_{\boldsymbol{x}}$. It can be read as the counterfactual sentence “the value that $\boldsymbol{Y}$ would have obtained in situation $\boldsymbol{U} = \boldsymbol{u}$, had $\boldsymbol{X}$ been $\boldsymbol{x}$.” Statistically, averaging $\boldsymbol{u}$ over the distribution $P(\boldsymbol{u})$ leads to the counterfactual variables $\boldsymbol{Y}_{\boldsymbol{x}}$. We denote $P(\boldsymbol{Y}_{\boldsymbol{x}})$ a distribution over counterfactual variables $\boldsymbol{Y}_{\boldsymbol{x}}$. We use $P(\boldsymbol{y}_{\boldsymbol{x}})$ as a shorthand for probabilities $P(\boldsymbol{Y}_{\boldsymbol{x}} = \boldsymbol{y})$ when the identify of the counterfactual variables is clear. By definitions, $P_{\boldsymbol{x}}(\boldsymbol{y}) = P(\boldsymbol{y}_{\boldsymbol{x}})$.

Lemma 5. *For a SCM $\langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, P(\boldsymbol{u})\rangle$, let subsets $\boldsymbol{S} \subseteq \boldsymbol{C} \subseteq \boldsymbol{V}$. For a topological ordering $\prec$ in $\mathcal{G}$, let $\boldsymbol{S}$ be ordered by $S_1 \prec \cdots \prec S_k$. $Q[\boldsymbol{S}]$ is bounded from $Q[\boldsymbol{C}]$ as:*

$$Q[\boldsymbol{S}] \in \left[A(\boldsymbol{S}, Q[\boldsymbol{C}]), B(\boldsymbol{S}, Q[\boldsymbol{C}])\right],$$

where $A(\boldsymbol{S}, Q[\boldsymbol{C}]), B(\boldsymbol{S}, Q[\boldsymbol{C}])$ are functions defined as follows. Let $\boldsymbol{W} = An(\boldsymbol{S})_{\mathcal{G}_{[\boldsymbol{C}]}}$. If $\boldsymbol{W} = \boldsymbol{S}$,

$$A(\boldsymbol{S}, Q[\boldsymbol{C}]) = B(\boldsymbol{S}, Q[\boldsymbol{C}]) = Q[\boldsymbol{W}],$$

where $Q[\boldsymbol{W}] = \sum_{\boldsymbol{c}\setminus\boldsymbol{w}} Q[\boldsymbol{C}]$; otherwise,

$$\begin{aligned} A(\boldsymbol{S}, Q[\boldsymbol{C}]) &= \max_{\boldsymbol{z}} Q[\boldsymbol{W}], \\ B(\boldsymbol{S}, Q[\boldsymbol{C}]) &= \min_{\boldsymbol{z}} \left\{ Q[\boldsymbol{W}] - \sum_{s_k} Q[\boldsymbol{W}] \right\} \\ &\quad + B(\boldsymbol{S} \setminus \{S_k\}, Q[\boldsymbol{C}]), \end{aligned}$$

where $\boldsymbol{Z} = Pa(\boldsymbol{W})_{\mathcal{G}} \setminus Pa(\boldsymbol{S})_{\mathcal{G}}$.

Proof. If $\boldsymbol{W} = \boldsymbol{S}$, (Tian, 2002, Lemma 10) implies that $Q[\boldsymbol{S}] = Q[\boldsymbol{W}] = \sum_{\boldsymbol{c}\setminus\boldsymbol{w}} Q[\boldsymbol{C}]$. Therefore, we have

$$A(\boldsymbol{S}, Q[\boldsymbol{C}]) = B(\boldsymbol{S}, Q[\boldsymbol{C}]) = Q[\boldsymbol{W}].$$

If $\boldsymbol{W} \neq \boldsymbol{S}$, or equivalently, $\boldsymbol{S} \subset \boldsymbol{W}$, by definitions,

$$Q[\boldsymbol{S}] = P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{s}}), Q[\boldsymbol{S}] = P(\boldsymbol{w}_{\boldsymbol{v}\setminus\boldsymbol{w}}).$$

Let $\boldsymbol{R} = \boldsymbol{W} \setminus \boldsymbol{S}$. By basic probabilistic operations,

$$\begin{aligned} P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{s}}) &= \sum_{\boldsymbol{r}'} P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{s}}, \boldsymbol{r}'_{\boldsymbol{v}\setminus\boldsymbol{w}}) = \sum_{\boldsymbol{r}'} P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{w},\boldsymbol{r}}, \boldsymbol{r}'_{\boldsymbol{v}\setminus\boldsymbol{w}}) \\ &\geq P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{w},\boldsymbol{r}}, \boldsymbol{r}_{\boldsymbol{v}\setminus\boldsymbol{w}}) \end{aligned}$$

By the composition axiom,

$$P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{w},\boldsymbol{r}}, \boldsymbol{r}_{\boldsymbol{v}\setminus\boldsymbol{w}}) = P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{w}}, \boldsymbol{r}_{\boldsymbol{v}\setminus\boldsymbol{w}}) = P(\boldsymbol{w}_{\boldsymbol{v}\setminus\boldsymbol{w}}).$$

We thus have

$$Q[\boldsymbol{S}] \geq Q[\boldsymbol{W}].$$

Since $Q[\boldsymbol{S}]$ is a function of $Pa(\boldsymbol{S})_{\mathcal{G}}$, it does not depends on values of $\boldsymbol{Z} = Pa(\boldsymbol{W})_{\mathcal{G}} \setminus Pa(\boldsymbol{S})_{\mathcal{G}}$. Taking a maximum over $\boldsymbol{Z}$ gives

$$A(\boldsymbol{S}, Q[\boldsymbol{C}]) = \max_{\boldsymbol{z}} Q[\boldsymbol{W}].$$

We now prove $Q[\boldsymbol{S}] \leq B(\boldsymbol{S}, Q[\boldsymbol{C}])$ by induction. The base case $\boldsymbol{W} = \boldsymbol{S}$ is implied by (Tian, 2002, Lemma 10). For $\boldsymbol{W} \neq \boldsymbol{S}$, we assume that

$$Q[\boldsymbol{S} \setminus \{S_k\}] \leq B(\boldsymbol{S} \setminus \{S_k\}, Q[\boldsymbol{C}])$$

By basic probabilistic operations,

$$\begin{aligned} &P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{s}}) \\ &= \sum_{\boldsymbol{r}'} P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{s}}, \boldsymbol{r}'_{\boldsymbol{v}\setminus\boldsymbol{w}}) \\ &= P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{w},\boldsymbol{r}}, \boldsymbol{r}_{\boldsymbol{v}\setminus\boldsymbol{w}}) + \sum_{\boldsymbol{r}'\neq\boldsymbol{r}} P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{w},\boldsymbol{r}}, \boldsymbol{r}'_{\boldsymbol{v}\setminus\boldsymbol{w}}) \\ &\leq P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{w},\boldsymbol{r}}, \boldsymbol{r}_{\boldsymbol{v}\setminus\boldsymbol{w}}) + \sum_{\boldsymbol{r}'\neq\boldsymbol{r}} P((\boldsymbol{s} \setminus \{s_k\})_{\boldsymbol{v}\setminus\boldsymbol{w},\boldsymbol{r}}, \boldsymbol{r}'_{\boldsymbol{v}\setminus\boldsymbol{w}}) \\ &= P(\boldsymbol{s}_{\boldsymbol{v}\setminus\boldsymbol{w}}, \boldsymbol{r}_{\boldsymbol{v}\setminus\boldsymbol{w}}) - P((\boldsymbol{s} \setminus \{s_k\})_{\boldsymbol{v}\setminus\boldsymbol{w},\boldsymbol{r}}, \boldsymbol{r}_{\boldsymbol{v}\setminus\boldsymbol{w}}) \\ &+ P((\boldsymbol{s} \setminus \{s_k\})_{\boldsymbol{v}\setminus\boldsymbol{w},\boldsymbol{r}}) \\ &= Q[\boldsymbol{W}] - \sum_{s_k} Q[\boldsymbol{W}] + Q[\boldsymbol{S} \setminus \{S_k\}]. \end{aligned}$$

Since $Q[\boldsymbol{S}]$ and $Q[\boldsymbol{S} \setminus \{S_k\}]$ are not functions of $\boldsymbol{Z}$, taking a minimum over $\boldsymbol{Z}$ gives

$$Q[\boldsymbol{S}] \leq \min_{\boldsymbol{z}} \left\{ Q[\boldsymbol{W}] - \sum_{s_k} Q[\boldsymbol{W}] \right\} + Q[\boldsymbol{S} \setminus \{S_k\}]. \tag{34}$$

Replacing $Q[\boldsymbol{S} \setminus \{S_k\}]$ with $B(\boldsymbol{S} \setminus \{S_k\}, Q[\boldsymbol{C}])$ proves the statement. □

Theorem 5. *Given $[\![\mathcal{G}, \Pi, Y]\!]$, for any $S_k \in \boldsymbol{S}$, let $\boldsymbol{C}$ be a c-component in $\mathcal{G}$ that contains $\bar{\boldsymbol{S}}_k$. Let $\boldsymbol{C}_k = \boldsymbol{C} \cap \boldsymbol{S}^{(k)}$ and let $\boldsymbol{Z} = Pa(\boldsymbol{C}_k)_{\mathcal{G}} \setminus Pa(\bar{\boldsymbol{S}}_k)_{\mathcal{G}}$. $P_{\bar{\boldsymbol{x}}_k}(s_k | \bar{\boldsymbol{s}}_k \setminus \{s_k\})$ is bounded in $\left[a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}, b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}\right]$ where*

$$\begin{aligned} a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k} &= \max_{\boldsymbol{z}} \left\{ A(\boldsymbol{C}_k, Q[\boldsymbol{C}]) / B(\boldsymbol{C}_k \setminus \{S_k\}, Q[\boldsymbol{C}]) \right\}, \\ b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k} &= \min_{\boldsymbol{z}} \left\{ B(\boldsymbol{C}_k, Q[\boldsymbol{C}]) / B(\boldsymbol{C}_k \setminus \{S_k\}, Q[\boldsymbol{C}]) \right\}. \end{aligned}$$

Proof. Since $\boldsymbol{C}$ contains $\bar{\boldsymbol{S}}_k$ and $\boldsymbol{C}_k = \boldsymbol{C} \cap \boldsymbol{S}^{(k)}$, by the factorization of Eq. (19),

$$P_{\bar{\boldsymbol{x}}_k}(s_k | \bar{\boldsymbol{s}}_k \setminus \{s_k\}) = Q[\boldsymbol{C}_k] / Q[\boldsymbol{C}_k \setminus \{S_k\}].$$

It immediately follows from Lem. 5 that

$$\frac{Q[\boldsymbol{C}_k]}{Q[\boldsymbol{C}_k \setminus \{S_k\}]} \geq \frac{A(\boldsymbol{C}_k, Q[\boldsymbol{C}])}{B(\boldsymbol{C}_k \setminus \{S_k\}, Q[\boldsymbol{C}])}$$

Since $P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$ is not a function of $\boldsymbol{Z} = Pa(\boldsymbol{C}_k)_{\mathcal{G}} \setminus Pa(\bar{\boldsymbol{S}}_k)_{\mathcal{G}}$,

$$P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) \geq \max_{\boldsymbol{z}} \left\{ \frac{A(\boldsymbol{C}_k, Q[\boldsymbol{C}])}{B(\boldsymbol{C}_k \setminus \{S_k\}, Q[\boldsymbol{C}])} \right\}$$

To prove the upper bound, we first write

$$\frac{Q[\boldsymbol{C}_k]}{Q[\boldsymbol{C}_k \setminus \{S_k\}]} = 1 + \frac{Q[\boldsymbol{C}_k] - Q[\boldsymbol{C}_k \setminus \{S_k\}]}{Q[\boldsymbol{C}_k \setminus \{S_k\}]}$$

By (Tian, 2002, Lemma 10), $Q[\boldsymbol{C}_k \setminus \{S_k\}] = \sum_{s_k} Q[\boldsymbol{C}_k]$. This implies

$$Q[\boldsymbol{C}_k] - Q[\boldsymbol{C}_k \setminus \{S_k\}] \leq 0.$$

This means that $Q[\boldsymbol{C}_k]/Q[\boldsymbol{C}_k \setminus \{S_k\}]$ is upper bounded when $Q[\boldsymbol{C}_k \setminus \{S_k\}]$ is taking the maximum values, i.e.,

$$\frac{Q[\boldsymbol{C}_k]}{Q[\boldsymbol{C}_k \setminus \{S_k\}]} \leq 1 + \frac{Q[\boldsymbol{C}_k] - Q[\boldsymbol{C}_k \setminus \{S_k\}]}{B(\boldsymbol{C}_k \setminus \{S_k\}, Q[\boldsymbol{C}])}$$

Let $\boldsymbol{W} = An(\boldsymbol{C}_k)_{\mathcal{G}}$ and let $\tilde{Z} = Pa(\boldsymbol{C}_k)_{\mathcal{G}} \setminus Pa(\boldsymbol{W})_{\mathcal{G}}$. By Eq. (34),

$$\begin{aligned}\frac{Q[\boldsymbol{C}_k]}{Q[\boldsymbol{C}_k \setminus \{S_k\}]} &\leq 1 + \frac{\min_{\tilde{z}} \left\{ Q[\boldsymbol{W}] - \sum_{s_k} Q[\boldsymbol{W}] \right\}}{B(\boldsymbol{C}_k \setminus \{S_k\}, Q[\boldsymbol{C}])} \\ &= \frac{B(\boldsymbol{C}_k, Q[\boldsymbol{C}])}{B(\boldsymbol{C}_k \setminus \{S_k\}, Q[\boldsymbol{C}])}\end{aligned}$$

Since $P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$ is not a function of $\boldsymbol{Z} = Pa(\boldsymbol{C}_k)_{\mathcal{G}} \setminus Pa(\bar{\boldsymbol{S}}_k)_{\mathcal{G}}$, taking minimum over $\boldsymbol{z}$ gives

$$P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) \leq \max_{\boldsymbol{z}} \left\{ \frac{B(\boldsymbol{C}_k, Q[\boldsymbol{C}])}{B(\boldsymbol{C}_k \setminus \{S_k\}, Q[\boldsymbol{C}])} \right\}.$$

Finally, the interventional quantities $Q[\boldsymbol{C}]$ is identifiable from the observational distribution $P(\boldsymbol{v})$ following (Tian, 2002, Lem. 7), which completes the proof. □

Theorem 6. *Given $[\![\mathcal{G}, \Pi, Y]\!]$ and causal bounds $\mathfrak{C}$, fix a $\delta \in (0, 1)$. W.p. at least $1 - \delta$, it holds for any $T > 1$, the regret of* `OFU-DTR` *is bounded by*

$$R(T, M^*) \leq \Delta(T, \mathfrak{C}, \delta) + 2|\boldsymbol{S}|\sqrt{T \log(2|\boldsymbol{S}|T/\delta)},$$

where function $\Delta(T, \mathfrak{C}, \delta)$ is defined as

$$\sum_{S_k \in \boldsymbol{S}} \min \left\{ |\mathfrak{C}_{S_k}| T, 17\sqrt{|\mathfrak{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}| T \log(|\boldsymbol{S}|T/\delta)} \right\}.$$

Proof. Let $\mathfrak{P}_c$ denote the family of parameters $P_{\boldsymbol{x}}(\boldsymbol{s})$ defined by causal bounds $\mathfrak{C}$. Since $P(P_{\boldsymbol{x}}(\boldsymbol{s}) \in \mathfrak{P}_c) = 1$,

$$\begin{aligned}&P(P_{\boldsymbol{x}}(\boldsymbol{s}) \notin (\mathfrak{P}_c \cap \mathfrak{P}_t)) \\ &\leq P(P_{\boldsymbol{x}}(\boldsymbol{s}) \notin \mathfrak{P}_c) + P(P_{\boldsymbol{x}}(\boldsymbol{s}) \notin \mathfrak{P}_t) \\ &= P(P_{\boldsymbol{x}}(\boldsymbol{s}) \neq \mathfrak{P}_t) \leq \delta/(4t^2).\end{aligned}$$

The last step follows from Lem. 13. By similar arguments of Lem. 14, we have

$$R(T, M^*) \leq \sum_{t=1}^{T} V_{\boldsymbol{\pi}_t}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t,$$

for all $t = 1, 2, \dots$ with probabilities $1 - \delta/2$. By Eqs. (25), (26) and (29),

$$\begin{aligned}&\sum_{t=1}^{T} V_{\boldsymbol{\pi}_t}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t \leq 2|\boldsymbol{S}|\sqrt{T \log(2|\boldsymbol{S}|T/\delta)} \\ &+ \sum_{V_i \in \boldsymbol{S}} \sum_{t=1}^{T} V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})).\end{aligned}$$

It is thus sufficient to show that

$$\begin{aligned}&\sum_{t=1}^{T} V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) \\ &\leq \min \left\{ |\mathfrak{C}_{S_k}| T, 17\sqrt{|\mathfrak{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}| T \log(|\boldsymbol{S}|T/\delta)} \right\}. \quad (35)\end{aligned}$$

Suppose

$$T \leq 17^2 |\mathfrak{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}| \log(|\boldsymbol{S}|T/\delta). \quad (36)$$

By the causal bounds $\mathfrak{C}_{S_k}$,

$$\begin{aligned}&V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) \\ &\leq \left\| P^t_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) - P_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) \right\|_1 \\ &\cdot \max_{s_k} \left\{ V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) \right\} \\ &\leq \min \left\{ |\mathfrak{C}_{S_k}|, 1 \right\},\end{aligned}$$

which implies

$$\begin{aligned}&\sum_{t=1}^{T} V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) \\ &\leq \min \left\{ |\mathfrak{C}_{S_k}| T, T \right\} = \min \left\{ |\mathfrak{C}_{S_k}| T, (\sqrt{T})^2 \right\}.\end{aligned}$$

By Eq. (36), we have

$$\begin{aligned}&\sum_{t=1}^{T} V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) \\ &\leq \min \left\{ |\mathfrak{C}_{S_k}| T, \sqrt{T} \cdot \sqrt{17^2 |\mathfrak{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}| \log(|\boldsymbol{S}|T/\delta)} \right\} \\ &= \min \left\{ |\mathfrak{C}_{S_k}| T, 17\sqrt{|\mathfrak{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}| T \log(|\boldsymbol{S}|T/\delta)} \right\},\end{aligned}$$

which proves Eq. (35). We now consider the case where

$$T > 17^2 |\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}| \log(|\boldsymbol{S}|T/\delta). \tag{37}$$

The definitions of parameter families $\mathfrak{P}_c \cap \mathfrak{P}_t$ imply that

$$\begin{aligned}
&V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) \\
&\leq \left\| P^t_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) - P_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) \right\|_1 \\
&\cdot \max_{s_k} \left\{ V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) \right\} \\
&\leq \min \left\{ |\mathfrak{C}_{S_k}|, 2\sqrt{\frac{6|\mathcal{D}_{S_k}| \log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|t/\delta)}{\max\left\{n^t(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k \setminus \{s_k\}), 1\right\}}} \right\}
\end{aligned}$$

By Eq. (27), we have

$$\begin{aligned}
&\sum_{V_i \in \boldsymbol{S}} \sum_{t=1}^{T} V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) \\
&\leq \sum_{S_k \in \boldsymbol{S}} \min \Bigg\{ |\mathfrak{C}_{S_k}|T, \\
&12\sqrt{|\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}|T \log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|T/\delta)} \Bigg\}
\end{aligned} \tag{38}$$

Whenever Eq. (37) holds,

$$\log(2|\boldsymbol{S}||\mathcal{D}_{(\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k)\setminus\{S_k\}}|T/\delta) \leq 2\log(|\boldsymbol{S}|T/\delta).$$

We thus write Eq. (38) as

$$\begin{aligned}
&\sum_{V_i \in \boldsymbol{S}} \sum_{t=1}^{T} V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s})) \\
&\leq \sum_{S_k \in \boldsymbol{S}} \min \left\{ |\mathfrak{C}_{S_k}|T, 12\sqrt{2|\mathcal{D}_{\bar{\boldsymbol{S}}_k \cup \bar{\boldsymbol{X}}_k}|T \log(|\boldsymbol{S}|T/\delta)} \right\}
\end{aligned}$$

which implies Eq. (35). This completes the proof. □

Theorem 7. *Given* $[\![\mathcal{G}, \Pi, Y]\!]$*, a prior* ϕ *and causal bounds* $\mathfrak{C}$*, if* ϕ *satisfies Eq.* (31)*, it holds for any* $T > 1$*, the regret of* PS-DTR *is bounded by*

$$R(T, \phi) \leq \Delta(T, \mathfrak{C}, 1/T) + 1,$$

where function $\Delta(T, \mathfrak{C}, \delta)$ *follows the definition in Thm.* 6.

Proof. Since ϕ satisfies Eq. (31), the rejection sampling ensures that $P_{\boldsymbol{x}}(\boldsymbol{s})$ and $P^t_{\boldsymbol{x}}(\boldsymbol{s})$ are identically distributed given any history $\mathfrak{H}_t$ and causal bounds $\mathfrak{C}$. Following a similar procedure as the proofs for Lems. 16 and 17,

$$\begin{aligned}
R(T, \phi^*) &= \sum_{t=1}^{T} E[V_{\boldsymbol{\pi}^t}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) - Y^t] \leq \delta T \\
&+ \sum_{V_i \in \boldsymbol{S}} \sum_{t=1}^{T} E[V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^t_{\boldsymbol{x}}(\boldsymbol{s})) - V_{\boldsymbol{\pi}_t}(\boldsymbol{V}^{(i-1)}; P^{(i)}_{\boldsymbol{x}}(\boldsymbol{s}))].
\end{aligned}$$

Following a simplification procedure similar to Thm. 6,

$$R(T, \phi^*) \leq \Delta(T, \mathfrak{C}, \delta) + \delta T$$

Fix $\delta = 1/T$, which completes the proof. □

Appendix D. Optimistic Single Policy Update

In OFU-DTR, the agent needs to find a near-optimal policy $\sigma^t_{\boldsymbol{X}}$ for an optimistic $P^t_{\boldsymbol{x}}(\boldsymbol{s}) \in \mathfrak{P}_t$. We can formulate this as an general problem as follows. For any $S_k \in \boldsymbol{S}$, let $\mathfrak{P}_{\bar{\boldsymbol{s}}_k \setminus \{s_k\}, \bar{\boldsymbol{x}}_k}$ denote a convex polytope over $P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$. We are searching for a policy $\sigma_{\boldsymbol{X}}$ and a distribution $P_{\boldsymbol{x}}(\boldsymbol{s})$ solving the optimization problem defined as:

$$\begin{aligned}
\max_{\sigma_{\boldsymbol{X}} \in \Pi, P_{\boldsymbol{x}}(\boldsymbol{s})} & V_{\sigma_{\boldsymbol{X}}}(P_{\boldsymbol{x}}(\boldsymbol{s})) \\
\text{s.t. } & P_{\boldsymbol{x}}(\boldsymbol{s}) = \prod_{S_k \in \boldsymbol{S}} P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) \\
& \forall S_k \in \boldsymbol{S},\ P_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) \in \mathfrak{P}_{\bar{\boldsymbol{s}}_k \setminus \{s_k\}, \bar{\boldsymbol{x}}_k} \\
& \forall S_k \in \boldsymbol{S},\ \sum_{s_k} P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) = 1, \\
& \forall S_k \in \boldsymbol{S},\ P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\}) \in [0, 1].
\end{aligned} \tag{39}$$

In general, solving the above polynomial program could be NP-hard (Håstad, 2001). We will next introduce an alternative factorization of $P_{\boldsymbol{x}}(\boldsymbol{s})$ that allows us to solve the optimization program in Eq. (39) through a series of local optimization. Consider a soluble ordering $\prec$ in $G_{\sigma_{\boldsymbol{X}}}$ defined as follows. Let $\boldsymbol{X}$ be ordered by $X_1 \prec \cdots \prec X_n$. We define $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n$ be a partition over $\boldsymbol{S}$ as:

$$\boldsymbol{C}_i = H_{X_i} \setminus (\cup_{j=1}^{i-1} H_{X_j^+}).$$

We assume that $\boldsymbol{S} \cup \boldsymbol{X}$ are ordered by $\prec$ as follows:

$$\boldsymbol{C}_1 \prec X_1 \prec \boldsymbol{C}_2 \prec X_2 \prec \cdots \prec \boldsymbol{C}_n \prec X_n.$$

Since Π is soluble and minimal, $P_{\boldsymbol{x}}(\boldsymbol{s})$ could be factorized over $\prec$ as follows:

$$P_{\boldsymbol{x}}(\boldsymbol{s}) = \prod_{i=2}^{n+1} \prod_{S_k \in \boldsymbol{C}_{i-1}} \tilde{P}(s_k|\hat{pa}_k \setminus \{s_k\}),$$

where $\hat{Pa}_k = (H_{X^+_{i-1}} \cup \{S_k\}) \cup \{S_j \in \boldsymbol{C}_{i-1} : S_j \prec S_k\}$; $\tilde{P}(s_k|\hat{pa}_k \setminus \{s_k\})$ is a mapping from domains of $\hat{Pa}_k \setminus \{S_k\}$ to the probabilistic domains over values of S_k. It is verifiable that $\boldsymbol{S}_k \subseteq Pa_k$. We reformulate the optimization program in Eq. (39) using the above factorization as follows:

$$\begin{aligned}
\max_{\sigma_{\boldsymbol{X}} \in \Pi, P_{\boldsymbol{x}}(\boldsymbol{s})} & V_{\sigma_{\boldsymbol{X}}}(P_{\boldsymbol{x}}(\boldsymbol{s})) \\
\text{s.t. } & P_{\boldsymbol{x}}(\boldsymbol{s}) = \prod_{S_k \in \boldsymbol{S}} \tilde{P}(s_k|\hat{pa}_k \setminus \{s_k\}) \\
& \forall S_k \in \boldsymbol{S},\ \tilde{P}_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_{k-1}) \in \mathfrak{P}_{\bar{\boldsymbol{s}}_k \setminus \{s_k\}, \bar{\boldsymbol{x}}_k} \\
& \forall S_k \in \boldsymbol{S},\ \sum_{s_k} \tilde{P}(s_k|\hat{pa}_k \setminus \{s_k\}) = 1, \\
& \forall S_k \in \boldsymbol{S},\ \tilde{P}(s_k|\hat{pa}_k \setminus \{s_k\}) \in [0, 1].
\end{aligned} \tag{40}$$

By constructions, Eq. (40) provides an upper bound for the solution of Eq. (39). However, since it still considers the confidence set $\mathcal{P}_{\bar{\boldsymbol{s}}_k\setminus\{s_k\},\bar{\boldsymbol{x}}_k}$, the approximate given by Eq. (40) is still reasonably close to the actual optimal $E_{\sigma^*_{\boldsymbol{X}}}[Y]$.

Since Π is soluble, one could solve Eq. (40) through a series of local optimization following a reverse ordering relative to $\prec$. For any $X_i \in \boldsymbol{X}$, we define function $\tilde{V}(x_i, h_{X_i})$ as:

$$\tilde{V}(x_i, h_{X_i}) = \sum_{v\setminus\{h_{X_i}, x_i\}} E_{\boldsymbol{x}}[Y|\boldsymbol{s}] \prod_{S_i\in\boldsymbol{S}} \tilde{P}(s_i|\hat{pa}_i \setminus \{s_i\}) \prod_{X\in\boldsymbol{X}\setminus\{X_i\}} \sigma_X(x|h_X)$$

The optimal decision rule $\sigma_{X_i}(x_i|h_{X_i})$ is given by

$$\sigma_{X_i}(x_i|h_{X_i}) = \arg\max_{x_i} \tilde{V}(x_i, h_{X_i}).$$

For any $S_k \in \boldsymbol{S}$, we define function $\tilde{V}(\hat{pa}_k)$ as:

$$\tilde{V}(\hat{pa}_k) = \sum_{v\setminus pa_k} E_{\boldsymbol{x}}[Y|\boldsymbol{s}] \prod_{S_i\in\boldsymbol{S}\setminus\{S_k\}} \tilde{P}(s_i|\hat{pa}_i \setminus \{s_i\}) \prod_{X\in\boldsymbol{X}} \sigma_X(x|h_X)$$

The solution $\tilde{P}(s_k|\hat{pa}_k \setminus \{s_k\})$ is given by

$$\tilde{P}(s_k|\hat{pa}_k \setminus \{s_k\}) = \arg\max_{p\in\mathcal{P}_{\bar{\boldsymbol{s}}_k\setminus\{s_k\},\bar{\boldsymbol{x}}_k}} \sum_{s_k} p(s_k)\tilde{V}(\hat{pa}_k).$$

In the above equations, $p(s_k)$ is a vector in the convex polytope $\mathcal{P}_{\bar{\boldsymbol{s}}_k\setminus\{s_k\},\bar{\boldsymbol{x}}_k}$. The maximization of $p(s_k)$ is a linear program over $\mathcal{P}_{\bar{\boldsymbol{s}}_k\setminus\{s_k\},\bar{\boldsymbol{x}}_k}$, which is solvable using the standard linear programming algorithms.

Appendix E. Experimental Setup

In this section, we provide details of the setup for experiments presented in Sec. 5. We demonstrate our algorithms on several SCMs, including multi-stage treatment regimes for lung cancer (Nease Jr & Owens, 1997) and dyspnoea (Cowell et al., 2006). In all experiments, we test OFU-DTR algorithm (*ofu-dtr*) with failure tolerance $\delta = 1/T$, OFU-DTR with causal bounds (*ofu-dtr*$^+$) with causal bounds derived from observational data, PS-DTR algorithm (*ps-dtr*) using uninformative dirichlet priors, and PS-DTR incorporating causal bounds via rejection sampling (*ps-dtr*$^+$). As a baseline, we also include the sequential multiple assignment randomized trail (*rand*), UC-DTR algorithm (*uc-dtr*) and causal UC-DTR algorithm (*uc-dtr*$^+$) developed in (Zhang & Bareinboim, 2019). To emulate the unobserved confounding, we generate 2×10^6 observational samples using a behavior policy and hide some columns of covariates. Each experiment lasts for $T = 5.5 \times 10^3$ episodes. For all algorithms, we measure their average regrets $R(T, M^*)/T$ over 100 repetitions.

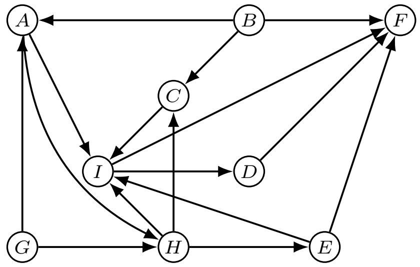


Figure 4: There causal diagram $\mathcal{G}_{\text{LUNG}}$ of the lung cancer staging example.

Variable	Description	Domain
A	CT Result	$0, 1, 2$
B	Mediastinal Metastases	$0, 1$
C	Mediastinoscopy Result	$0, 1, 2$
D	Treatment Death	$0, 1$
E	Mediastinoscopy Death	$0, 1$
F	Life Expectancy	$0, 1$
G	CT?	$0, 1$
H	Mediastinoscopy?	$0, 1$
I	Treatment	$0, 1$

Table 1: Summary of variables in the Lung cancer staging example described in Fig. 4.

Lung Cancer Staging

We consider a multi-staged treatment regime for the lung cancer introduced in (Nease Jr & Owens, 1997), which we shall refer to as M_{LUNG}.

> Consider the case of a patient with a known non-small-cell carcinoma of the lung. The primary tumor is 1cm in diameter; a chest x-ray examination suggests that the tumor does not abut the chest wall or mediastinum. Additional workup reveals no evidence of distance metastases. The preferred treatment in such a situation is thoracotomy, followed by lobectomy or pneumonectomy, depending on whether the primary tumor has metastasized to the hilar lymph nodes.
>
> Of fundamental importance in the decision to perform thoracotomy is the likelihood of mediastinal metastases. If mediastinal metastases are known to be present, most clinicians would deem thoracotomy to be contraindicated: thoracotomy subjects the patient to a risk of death but confers no health benefit. (Some surgeons attempt to resect mediastinal metastases that are ipsilateral to the primary tumor, but this approach remains controversial.) If mediastinal metastases are known to be absent, thoracotomy offers a substantial survival advantage, so long as the primary tumor has not

> metastasized to distant organs. There are several diagnostic tests available to assess any involvement of the mediastinum. For this example, we shall focus on computed tomography (CT) of the chest and mediastinoscopy. Our problem involves three decisions. First, should the patient undergo a CT scan? Second, given our decision about CT and any CT results obtained, should the patient undergo mediastinoscopy? Third, given the results of any tests that we have decided to perform, should the patient undergo thoracotomy?

The graphical representation $\mathcal{G}_{\text{LUNG}}$ of this environment is shown in Fig. 4. The detailed description of each node is shown in Table 1. We will consistently use 0 for “Yes”, 1 for “No” and 2 for “N/A”. We will next provide the numerical specification of this environment. For any variable X, we will use x_0, x_1, x_2 to represent realizations $X = 0, X = 1, X = 2$ respectively. The values of the conditional probabilities are given in Table 3; they are for illustrative purposes only.

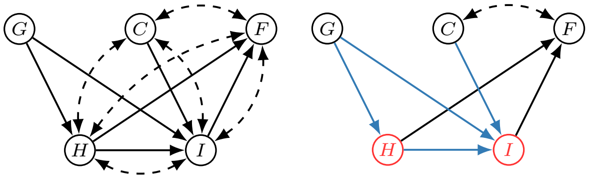
To generate the observational data, we sample from M_{LUNG} following the behavior policies described in Table 3 (i.e., the conditional probability distributions of G, H, I) and collect observed outcomes. To emulate the unobserved confounding, we hide columns of variables A, B, D, E, inducing an observational distribution $P(c, f, g, h, i)$. The causal diagram $\mathcal{G}$ compatible with $P(c, f, g, h, i)$ is thus the projection of $\mathcal{G}_{\text{LUNG}}$ onto variables C, F, G, H, I, which we show in Fig. 5a. Hypothetically, the “actual” SCM M^* conforming to $\mathcal{G}$ is the projection of SCM M_{LUNG} onto variables C, F, G, H, I, following an algorithm described in (Lee & Bareinboim, 2019). We will use the lift expectancy F as the primary outcome. The candidate policy space Π is given by $\{\mathcal{D}_G \mapsto \mathcal{D}_H, \mathcal{D}_{\{G,H,C\}} \mapsto \mathcal{D}_I\}$. We summarize this learning problem as the signature $[\![\mathcal{G}, \Pi, F]\!]$; Fig. 5b describes its associated manipulated diagram $\mathcal{G}_{\sigma_{H,I}}$.

The optimal policy $\sigma^*_{H,I}$ is described as follows:

$$\begin{aligned} I: \quad & \sigma^*_I(i_1|g, h_0, c_1) = 0, \\ & \text{otherwise } \sigma^*_I(i_1|g, h, c) = 1. \\ H: \quad & \sigma^*_H(i_1|g) = 1. \end{aligned}$$

The expected outcome $E_{\sigma^*_{H,I}}[F]$ of the optimal policy is equal to 0.5891. The procedure $\texttt{Reduce}(\mathcal{G}, \Pi, Y)$ finds the minimal reduction $\Pi_{\text{MIN}} = \{\mathcal{D}_\emptyset \mapsto \mathcal{D}_H, \mathcal{D}_{\{H,C\}} \mapsto \mathcal{D}_I\}$. OFU-DTR and PS-DTR thus focus on the transition distributions $P_h(c)$. For completeness, we provide parameters for transition probabilities $P(g)$ and $P_h(c)$ and the immediate outcome $E_{h,i}[F|c]$ in Table 4.

Following the analysis in the main draft, we assume that parameters of the immediate outcome $E_{h,i}[F|c]$ are provided. In all experiments, our proposed algorithms *ofu-dtr*, *ofu-dtr*$^+$, *ps-dtr*, *ps-dtr*$^+$ have access to the causal diagram $\mathcal{G}$; while other baseline algorithms *rand*, *uc-dtr*, *uc-dtr*$^+$ do not. Oblivious of the independence between G and C under $do(h)$, UC-DTR learns parameters of transition probabilities $P_h(c)$ using the empirical mean of distribution $P_h(c|g)$.



(a) $\mathcal{G}$ (b) $\mathcal{G}_{\sigma_{H,I}}$

Figure 5: (a) A causal diagram $\mathcal{G}$ induced by the projection of $\mathcal{G}_{\text{LUNG}}$ onto C, F, G, H, I; (b) the manipulated diagram $\mathcal{G}_{\sigma_{H,I}}$ with $\Pi = \{\mathcal{D}_G \mapsto \mathcal{D}_H, \mathcal{D}_{\{G,H,C\}} \mapsto \mathcal{D}_I\}$.

Among these algorithms, *rand*, *uc-dtr*, *ofu-dtr* and *ps-dtr* learn from the scratch. Other procedures including *ofu-dtr*$^+$ and *ps-dtr*$^+$ derive causal bounds $[a_{h,c}, b_{h,c}]$ over $P_h(c)$ from $P(g, c, f, h, i)$ and $\mathcal{G}$ using the method introduced in Thm. 5. Oblivious of the causal diagram $\mathcal{G}$, *uc-dtr*$^+$ derive bounds $P_h(c|g) \in [a_{h,g,c}, b_{h,g,c}]$. The details of these causal bounds are given in Table 5.

Dyspnoea

We consider a multi-staged treatment regime for the dyspnoea introduced in (Cowell et al., 2006), which we shall refer to as M_{DYSPNOEA}.

> Shortness of breath (dyspnoea) may be due to tuberculosis, lung cancer, bronchitis, none of them or more than one of them but its presence or absence does not discriminate between the diseases. A recent visit to Asia increases the chances of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. Suppose a doctor must decide whether a patient arriving at a clinic is to be hospitalized or not. Before taking the decision the doctor can obtain information as to whether the patient has gone to Asia or suffers from dyspnoea, but other relevant factors like smoking history or the presence of any diseases are not known. It has also been suggested that it may be worthwhile to screen the patient by taking chest X-rays. The results of a chest X-ray do not discriminate between lung cancer or tuberculosis. Proponents of the test say that it should be carried out at least for the people that have visited Asia. If a test is carried out, the doctor has access to the results at the time he determines whether to hospitalize or not. If the patient suffers

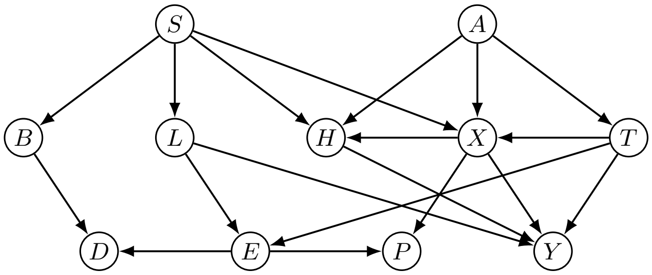
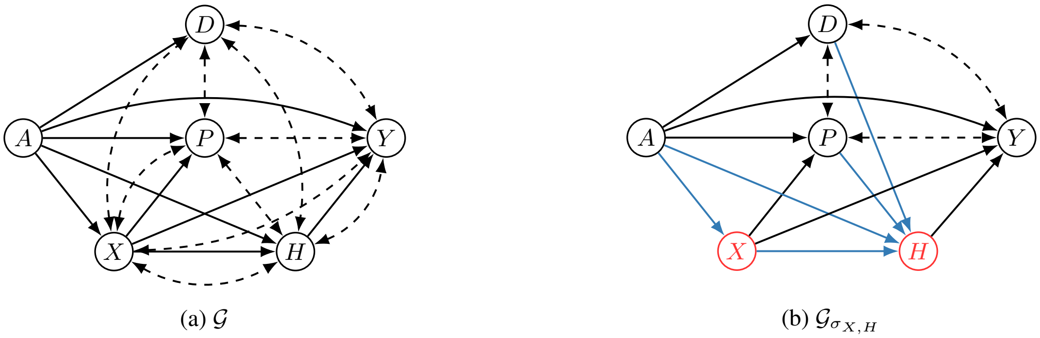


Figure 6: There causal diagram $\mathcal{G}_{\text{DYSPNOEA}}$ of the dyspnoea treatment regime example.



(a) $\mathcal{G}$

(b) $\mathcal{G}_{\sigma_{X,H}}$

Figure 7: (a) A causal diagram $\mathcal{G}$ induced by the projection of $\mathcal{G}_{\text{DYSPNOEA}}$ onto A, X, D, P, H, Y; (b) the manipulated diagram $\mathcal{G}_{\sigma_{X,H}}$ with policy space $\Pi = \{\mathcal{D}_A \mapsto \mathcal{D}_X, \mathcal{D}_{\{A,X,D,P\}} \mapsto \mathcal{D}_H\}$.

Variable	Description	Domain
S	Smoking	$0, 1$
A	Visit to Asia?	$0, 1$
T	Tuberculosis?	$0, 1$
B	Bronchitis?	$0, 1$
L	Lung cancer?	$0, 1$
E	Either tub. or cancer?	$0, 1$
X	X-ray?	$0, 1$
D	Dyspnoea?	$0, 1$
P	Positive X-ray?	$0, 1$
H	Hospitalize?	$0, 1$

Table 2: Summary of variables in the dyspnoea treatment regime example described in Fig. 6.

> from tuberculosis or lung cancer, he can be treated better in hospital, but hospitalization of healthy individuals should be avoided. Taking X-rays is harmful in itself and the adverse effects are more severe if the patient suffers from tuberculosis.

The graphical representation $\mathcal{G}_{\text{DYSPNOEA}}$ of this environment is shown in Fig. 6. The detailed description of each node is shown in Table 2. We will consistently use 0 for "Yes" and 1 for "No". We will next provide the numerical specification of this environment. For any variable X, we will use x_0, x_1 to represent realizations $X = 0, X = 1$ respectively. The values of the conditional probabilities are given in Table 6; they are for illustrative purposes only.

To generate the observational data, we sample from M_{DYSPNOEA} following the behavior policies described in Table 2 (i.e., the conditional probability distributions of X, H) and collect observed outcomes. To emulate the unobserved confounding, we hide columns of variables S, B, L, T, E, inducing an observational distribution $P(a, x, h, d, p, y)$. The causal diagram $\mathcal{G}$ compatible with $P(a, x, h, d, p, y)$ is thus the projection of $\mathcal{G}_{\text{DYSPNOEA}}$ onto variables A, X, H, D, P, Y, which we show in Fig. 7a. Hypothetically, the "actual" SCM M^* conforming to $\mathcal{G}$ is the projection of model M_{DYSPNOEA} onto variables A, X, H, D, P, Y, following an algorithm described in (Lee & Bareinboim, 2019). We will use the utility Y as the primary outcome. The candidate policy space Π is given by $\{\mathcal{D}_V \mapsto \mathcal{D}_X, \mathcal{D}_{\{A,X,D,P\}} \mapsto \mathcal{D}_H\}$. We summarize this learning problem as the signature $[\![\mathcal{G}, \Pi, F]\!]$; Fig. 7b describes its associated manipulated diagram $\mathcal{G}_{\sigma_{X,H}}$.

The optimal policy $\sigma^*_{X,H}$ is described as follows:

$$\begin{aligned} H: \quad & \sigma^*_H(h_1|a_1, x_0, d_0, p_1) = 1, \\ & \text{otherwise } \sigma^*_H(h_1|a, x, d, p) = 0. \\ X: \quad & \sigma^*_X(x_1|a) = 0. \end{aligned}$$

The expected outcome $E_{\sigma^*_{X,H}}[Y]$ of the optimal policy is 0.789. For completeness, we also provide probabilities for the transition distribution $P(v)$, $P(d|a)$ and $P_x(p|d,a)$ and the immediate outcome $E_{x,h}[Y|a,d,p]$ in Table 7.

Following the analysis in the main draft, we assume that parameters of the immediate outcome $E_{x,h}[Y|a,d,p]$ are provided. We also simplify the optimization procedure and do not require the learning of $P(v)$, since its parameters do not affect the optimal policy $\sigma^*_{X,H}$. In all experiments, our proposed algorithms *ofu-dtr*, *ofu-dtr*$^+$, *ps-dtr*, *ps-dtr*$^+$ have access to the causal diagram $\mathcal{G}$; while other baseline algorithms *rand*, *uc-dtr*, *uc-dtr*$^+$ do not. Oblivious of the causal relationships encoded in $\mathcal{G}$, `UC-DTR` treat variables D, P *en bloc* and focuses on learning the transition probabilities $P_x(d,p|v)$. On the other hand, *ofu-dtr*, *ofu-dtr*$^+$, *ps-dtr*, *ps-dtr*$^+$ utilize the factorization

$$P_x(d,p|v) = P(d|a)P_x(p|d,a),$$

and learn parameters of $P(d|a)$ and $P_x(p|d,a)$ separately.

Among these algorithms, *rand*, *uc-dtr*, *ofu-dtr* and *ps-dtr* learn from the scratch; while *ofu-dtr*$^+$, *ps-dtr*$^+$ and *uc-dtr*$^+$ also utilize the observational data. Since $P(d|a)$ is identifiable from $P(a,x,h,d,p,y)$, *ofu-dtr*$^+$ and *ps-dtr*$^+$ estimate parameters of $P(d|a)$ from the observational data using its empirical means. Furthermore, *ofu-dtr*$^+$ and *ps-dtr*$^+$ compute the causal bounds $[a_{x,a,d}(p), b_{x,a,d}(p)]$ over $P_x(p|d,a)$ from the empirical estimates of $P(a,x,h,d,p,y)$. Oblivious of the causal diagram $\mathcal{G}$, *uc-dtr*$^+$ derive bounds $P_x(d,p|a) \in [a_{x,a}(d,p), b_{x,a}(d,p)]$. The details of these causal bounds are given in Table 8.

References

Correa, J. and Bareinboim, E. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

Cowell, R. G., Dawid, P., Lauritzen, S. L., and Spiegelhalter, D. J. *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer Science & Business Media, 2006.

Håstad, J. Some optimal inapproximability results. *Journal of the ACM (JACM)*, 48(4):798–859, 2001.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Koller, D. and Milch, B. Multi-agent influence diagrams for representing and solving games. *Games and economic behavior*, 45(1):181–221, 2003.

Lauritzen, S. L. and Nilsson, D. Representing and solving decision problems with limited information. *Management Science*, 47(9):1235–1251, 2001.

Lee, S. and Bareinboim, E. Structural causal bandits with non-manipulable variables. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 4164–4172, Honolulu, Hawaii, 2019. AAAI Press.

Liu, Q. and Ihler, A. Belief propagation for structured decision making. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 523–532. AUAI Press, 2012.

Nease Jr, R. F. and Owens, D. K. Use of influence diagrams to structure medical decisions. *Medical Decision Making*, 17(3):263–275, 1997.

Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. In *Advances in NeurIPS*, pp. 3003–3011, 2013.

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

Tian, J. *Studies in Causal Reasoning and Learning*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, November 2002.

Zhang, J. and Bareinboim, E. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, 2019.

A:	$P(a_0 \| b_0, g_0)$	$=$	0.2841	$P(a_1 \| b_0, g_0)$	$=$	0.5005
	$P(a_0 \| b_0, g_1)$	$=$	0.4862	$P(a_1 \| b_0, g_1)$	$=$	0.4792
	$P(a_0 \| b_1, g_0)$	$=$	0.4680	$P(a_1 \| b_1, g_0)$	$=$	0.4077
	$P(a_0 \| b_1, g_1)$	$=$	0.0330	$P(a_1 \| b_1, g_1)$	$=$	0.6757
B:	$P(b_0)$	$=$	0.5417	$P(b_1)$	$=$	0.4583
C:	$P(c_0 \| b_0, h_0)$	$=$	0.4103	$P(c_1 \| b_0, h_0)$	$=$	0.1062
	$P(c_0 \| b_0, h_1)$	$=$	0.3080	$P(c_1 \| b_0, h_1)$	$=$	0.4666
	$P(c_0 \| b_1, h_0)$	$=$	0.3997	$P(c_1 \| b_1, h_0)$	$=$	0.5083
	$P(c_0 \| b_1, h_1)$	$=$	0.3017	$P(c_1 \| b_1, h_1)$	$=$	0.3389
D:	$P(d_0 \| i_0)$	$=$	0.4328	$P(d_0 \| i_1)$	$=$	0.2731
E:	$P(e_1 \| h_0)$	$=$	0.1473	$P(e_1 \| h_1)$	$=$	0.8849
F:	$P(f_1 \| b_0, d_0, e_0, i_0)$	$=$	0.1491	$P(f_1 \| b_0, d_0, e_0, i_1)$	$=$	0.9693
	$P(f_1 \| b_0, d_0, e_1, i_0)$	$=$	0.0177	$P(f_1 \| b_0, d_0, e_1, i_1)$	$=$	0.2382
	$P(f_1 \| b_0, d_1, e_0, i_0)$	$=$	0.8229	$P(f_1 \| b_0, d_1, e_0, i_1)$	$=$	0.9601
	$P(f_1 \| b_0, d_1, e_1, i_0)$	$=$	0.2460	$P(f_1 \| b_0, d_1, e_1, i_1)$	$=$	0.8257
	$P(f_1 \| b_1, d_0, e_0, i_0)$	$=$	0.0937	$P(f_1 \| b_1, d_0, e_0, i_1)$	$=$	0.2567
	$P(f_1 \| b_1, d_0, e_1, i_0)$	$=$	0.5303	$P(f_1 \| b_1, d_0, e_1, i_1)$	$=$	0.1900
	$P(f_1 \| b_1, d_1, e_0, i_0)$	$=$	0.4400	$P(f_1 \| b_1, d_1, e_0, i_1)$	$=$	0.3264
	$P(f_1 \| b_1, d_1, e_1, i_0)$	$=$	0.6326	$P(f_1 \| b_1, d_1, e_1, i_1)$	$=$	0.3320
G:	$P(g_0)$	$=$	0.2546	$P(g_1)$	$=$	0.7454
H:	$P(h_1 \| a_0, g_0)$	$=$	0.9456	$P(h_1 \| a_0, g_1)$	$=$	0.4239
	$P(h_1 \| a_1, g_0)$	$=$	0.7273	$P(h_1 \| a_1, g_1)$	$=$	0.6931
	$P(h_1 \| a_2, g_0)$	$=$	0.4035	$P(h_1 \| a_2, g_1)$	$=$	0.4228
I:	$P(i_0 \| a, c_0, e_0, g_0, h_0)$	$=$	0.1576	$P(i_0 \| a, c_0, e_0, g_0, h_1)$	$=$	0.8491
	$P(i_0 \| a, c_0, e_0, g_1, h_0)$	$=$	0.4218	$P(i_0 \| a, c_0, e_0, g_1, h_1)$	$=$	0.6555
	$P(i_0 \| a, c_0, e_1, g_0, h_0)$	$=$	0.4854	$P(i_0 \| a, c_0, e_1, g_0, h_1)$	$=$	0.7577
	$P(i_0 \| a, c_0, e_1, g_1, h_0)$	$=$	0.9595	$P(i_0 \| a, c_0, e_1, g_1, h_1)$	$=$	0.0318
	$P(i_0 \| a, c_1, e_0, g_0, h_0)$	$=$	0.9706	$P(i_0 \| a, c_1, e_0, g_0, h_1)$	$=$	0.9340
	$P(i_0 \| a, c_1, e_0, g_1, h_0)$	$=$	0.9157	$P(i_0 \| a, c_1, e_0, g_1, h_1)$	$=$	0.1712
	$P(i_0 \| a, c_1, e_1, g_0, h_0)$	$=$	0.8003	$P(i_0 \| a, c_1, e_1, g_0, h_1)$	$=$	0.7431
	$P(i_0 \| a, c_1, e_1, g_1, h_0)$	$=$	0.6557	$P(i_0 \| a, c_1, e_1, g_1, h_1)$	$=$	0.2769
	$P(i_0 \| a, c_2, e_0, g_0, h_0)$	$=$	0.9572	$P(i_0 \| a, c_2, e_0, g_0, h_1)$	$=$	0.6787
	$P(i_0 \| a, c_2, e_0, g_1, h_0)$	$=$	0.7922	$P(i_0 \| a, c_2, e_0, g_1, h_1)$	$=$	0.7060
	$P(i_0 \| a, c_2, e_1, g_0, h_0)$	$=$	0.1419	$P(i_0 \| a, c_2, e_1, g_0, h_1)$	$=$	0.3922
	$P(i_0 \| a, c_2, e_1, g_1, h_0)$	$=$	0.0357	$P(i_0 \| a, c_2, e_1, g_1, h_1)$	$=$	0.0462

Table 3: Conditional probability distributions for the Lung cancer staging example described in Fig. 4.

G:	$P(g_0)$	$=$	0.2546	$P(g_1)$	$=$	0.7454
C:	$P_{h_0}(c_0)$	$=$	0.4055	$P_{h_1}(c_0)$	$=$	0.3051
	$P_{h_0}(c_1)$	$=$	0.2904	$P_{h_1}(c_1)$	$=$	0.4081
	$P_{h_0}(c_2)$	$=$	0.3041	$P_{h_1}(c_2)$	$=$	0.2868
F:	$E_{h_0, i_0}[F \| c_0]$	$=$	0.3559	$E_{h_1, i_0}[F \| c_0]$	$=$	0.3759
	$E_{h_0, i_0}[F \| c_1]$	$=$	0.4546	$E_{h_1, i_0}[F \| c_1]$	$=$	0.3707
	$E_{h_0, i_0}[F \| c_2]$	$=$	0.2677	$E_{h_1, i_0}[F \| c_2]$	$=$	0.3845
	$E_{h_0, i_1}[F \| c_0]$	$=$	0.5406	$E_{h_1, i_1}[F \| c_0]$	$=$	0.5919
	$E_{h_0, i_1}[F \| c_1]$	$=$	0.3854	$E_{h_1, i_1}[F \| c_1]$	$=$	0.6303
	$E_{h_0, i_1}[F \| c_2]$	$=$	0.6794	$E_{h_1, i_1}[F \| c_2]$	$=$	0.5276

Table 4: Transition distributions and the immediate outcome for the learning problem of the Lung cancer staging example.

$P_h(c)$:	$a_{h_0}(c_0)$	$=$	0.3045	$b_{h_0}(c_0)$	$=$	0.5530
	$a_{h_1}(c_0)$	$=$	0.1292	$b_{h_1}(c_0)$	$=$	0.7061
	$a_{h_0}(c_1)$	$=$	0.2252	$b_{h_0}(c_1)$	$=$	0.4737
	$a_{h_1}(c_1)$	$=$	0.1743	$b_{h_1}(c_1)$	$=$	0.7513
	$a_{h_0}(c_2)$	$=$	0.2218	$b_{h_0}(c_2)$	$=$	0.4703
	$a_{h_1}(c_2)$	$=$	0.1196	$b_{h_1}(c_2)$	$=$	0.6965
$P_h(c\|g)$:	$a_{h_0,g_0}(c_0)$	$=$	0.3045	$b_{h_0,g_0}(c_0)$	$=$	0.5530
	$a_{h_0,g_1}(c_0)$	$=$	0.2338	$b_{h_0,g_1}(c_0)$	$=$	0.6568
	$a_{h_1,g_0}(c_0)$	$=$	0.0759	$b_{h_1,g_0}(c_0)$	$=$	0.8274
	$a_{h_1,g_1}(c_0)$	$=$	0.1292	$b_{h_1,g_1}(c_0)$	$=$	0.7061
	$a_{h_0,g_0}(c_1)$	$=$	0.2252	$b_{h_0,g_0}(c_1)$	$=$	0.4737
	$a_{h_0,g_1}(c_1)$	$=$	0.1728	$b_{h_0,g_1}(c_1)$	$=$	0.5959
	$a_{h_1,g_0}(c_1)$	$=$	0.1036	$b_{h_1,g_0}(c_1)$	$=$	0.8551
	$a_{h_1,g_1}(c_1)$	$=$	0.1743	$b_{h_1,g_1}(c_1)$	$=$	0.7513
	$a_{h_0,g_0}(c_2)$	$=$	0.2218	$b_{h_0,g_0}(c_2)$	$=$	0.4703
	$a_{h_0,g_1}(c_2)$	$=$	0.1703	$b_{h_0,g_1}(c_2)$	$=$	0.5934
	$a_{h_1,g_0}(c_2)$	$=$	0.0690	$b_{h_1,g_0}(c_2)$	$=$	0.8205
	$a_{h_1,g_1}(c_2)$	$=$	0.1196	$b_{h_1,g_1}(c_2)$	$=$	0.6965

Table 5: Causal bounds for the transition probabilities $P_h(c) \in [a_h(c), b_h(c)]$ and $P_h(c|g) \in [a_{h,g}(c), b_{h,g}(c)]$ in the Lung cancer staging example.

A:	$P(a_0)$	$=$	0.8147	$P(a_1)$	$=$	0.1853
B:	$P(b_0\|s_0)$	$=$	0.1270	$P(b_0\|s_1)$	$=$	0.9134
D:	$P(d_0\|b_0, e_0)$	$=$	0.6324	$P(d_0\|b_0, e_1)$	$=$	0.2785
	$P(d_0\|b_1, e_0)$	$=$	0.0975	$P(d_0\|b_1, e_1)$	$=$	0.5469
E:	$P(e_0\|l_0, i_0)$	$=$	0.9575	$P(e_0\|l_0, i_1)$	$=$	0.1576
	$P(e_0\|l_1, i_0)$	$=$	0.9649	$P(e_0\|l_1, i_1)$	$=$	0.9706
L:	$P(l_0\|s_0)$	$=$	0.9572	$P(l_0\|s_1)$	$=$	0.4854
S:	$P(s_0)$	$=$	0.9058	$P(s_1)$	$=$	0.0942
T:	$P(t_0\|a_0)$	$=$	0.8003	$P(t_0\|a_1)$	$=$	0.1419
P:	$P(p_0\|e_0, x_0)$	$=$	0.4218	$P(p_0\|e_0, x_1)$	$=$	0.7922
	$P(p_0\|e_1, x_0)$	$=$	0.9157	$P(p_0\|e_1, x_1)$	$=$	0.9595
X:	$P(x_0\|s_0, a_0)$	$=$	0.6557	$P(x_0\|s_0, a_1)$	$=$	0.0357
	$P(x_0\|s_1, a_0)$	$=$	0.6557	$P(x_0\|s_1, a_1)$	$=$	0.0357
H:	$P(h_0\|s_0, a_0)$	$=$	0.0971	$P(h_0\|s_0, a_1)$	$=$	0.6948
	$P(h_0\|s_1, a_0)$	$=$	0.8235	$P(h_0\|s_1, a_1)$	$=$	0.3171
Y:	$P(y_0\|l_0, t_0, x_0, h_0)$	$=$	0.8491	$P(y_0\|l_0, t_0, x_0, h_1)$	$=$	0.6787
	$P(y_0\|l_0, t_0, x_1, h_0)$	$=$	0.9340	$P(y_0\|l_0, t_0, x_1, h_1)$	$=$	0.7577
	$P(y_0\|l_0, t_1, x_0, h_0)$	$=$	0.7431	$P(y_0\|l_0, t_1, x_0, h_1)$	$=$	0.6555
	$P(y_0\|l_0, t_1, x_1, h_0)$	$=$	0.3922	$P(y_0\|l_0, t_1, x_1, h_1)$	$=$	0.1712
	$P(y_0\|l_1, t_0, x_0, h_0)$	$=$	0.7060	$P(y_0\|l_1, t_0, x_0, h_1)$	$=$	0.2769
	$P(y_0\|l_1, t_0, x_1, h_0)$	$=$	0.0318	$P(y_0\|l_1, t_0, x_1, h_1)$	$=$	0.0462
	$P(y_0\|l_1, t_1, x_0, h_0)$	$=$	0.0971	$P(y_0\|l_1, t_1, x_0, h_1)$	$=$	0.6948
	$P(y_0\|l_1, t_1, x_1, h_0)$	$=$	0.8235	$P(y_0\|l_1, t_1, x_1, h_1)$	$=$	0.3171

Table 6: Conditional probability distributions for the dyspnoea treatment example described in Fig. 4.

D:	$P(d_0 \vert a_0)$	$=$	0.2633	$P(d_0 \vert a_1)$	$=$	0.4151
P:	$P_{x_0}(p_0 \vert d_0, a_0)$	$=$	0.5979	$P_{x_0}(p_0 \vert d_0, a_1)$	$=$	0.8206
	$P_{x_1}(p_0 \vert d_0, a_0)$	$=$	0.8518	$P_{x_1}(p_0 \vert d_0, a_1)$	$=$	0.9273
	$P_{x_0}(p_0 \vert d_1, a_0)$	$=$	0.4846	$P_{x_0}(p_0 \vert d_1, a_1)$	$=$	0.7028
	$P_{x_1}(p_0 \vert d_1, a_0)$	$=$	0.8135	$P_{x_1}(p_0 \vert d_1, a_1)$	$=$	0.8874
Y:	$E_{x_0,h_0}[Y \vert a_0, d_0, p_0]$	$=$	0.7745	$E_{x_1,h_0}[Y \vert a_0, d_0, p_0]$	$=$	0.6529
	$E_{x_0,h_0}[Y \vert a_1, d_0, p_0]$	$=$	0.7220	$E_{x_1,h_0}[Y \vert a_1, d_0, p_0]$	$=$	0.4447
	$E_{x_0,h_0}[Y \vert a_0, d_1, p_0]$	$=$	0.8084	$E_{x_1,h_0}[Y \vert a_0, d_1, p_0]$	$=$	0.7990
	$E_{x_0,h_0}[Y \vert a_1, d_1, p_0]$	$=$	0.7236	$E_{x_1,h_0}[Y \vert a_1, d_1, p_0]$	$=$	0.5041
	$E_{x_0,h_0}[Y \vert a_0, d_0, p_1]$	$=$	0.7906	$E_{x_1,h_0}[Y \vert a_0, d_0, p_0]$	$=$	0.7410
	$E_{x_0,h_0}[Y \vert a_1, d_0, p_1]$	$=$	0.6150	$E_{x_1,h_0}[Y \vert a_1, d_0, p_1]$	$=$	0.5552
	$E_{x_0,h_0}[Y \vert a_0, d_1, p_1]$	$=$	0.8230	$E_{x_1,h_0}[Y \vert a_0, d_1, p_1]$	$=$	0.8453
	$E_{x_0,h_0}[Y \vert a_1, d_1, p_1]$	$=$	0.6837	$E_{x_1,h_0}[Y \vert a_1, d_1, p_1]$	$=$	0.6171
	$E_{x_0,h_1}[Y \vert a_0, d_0, p_0]$	$=$	0.6371	$E_{x_1,h_1}[Y \vert a_0, d_0, p_0]$	$=$	0.4717
	$E_{x_0,h_1}[Y \vert a_1, d_0, p_0]$	$=$	0.6554	$E_{x_1,h_1}[Y \vert a_1, d_0, p_0]$	$=$	0.2109
	$E_{x_0,h_1}[Y \vert a_0, d_1, p_0]$	$=$	0.6530	$E_{x_1,h_1}[Y \vert a_0, d_1, p_0]$	$=$	0.6219
	$E_{x_0,h_1}[Y \vert a_1, d_1, p_0]$	$=$	0.6569	$E_{x_1,h_1}[Y \vert a_1, d_1, p_0]$	$=$	0.2731
	$E_{x_0,h_1}[Y \vert a_0, d_0, p_1]$	$=$	0.6179	$E_{x_1,h_1}[Y \vert a_0, d_0, p_0]$	$=$	0.5755
	$E_{x_0,h_1}[Y \vert a_1, d_0, p_1]$	$=$	0.6545	$E_{x_1,h_1}[Y \vert a_1, d_0, p_1]$	$=$	0.2919
	$E_{x_0,h_1}[Y \vert a_0, d_1, p_1]$	$=$	0.6528	$E_{x_1,h_1}[Y \vert a_0, d_1, p_1]$	$=$	0.6731
	$E_{x_0,h_1}[Y \vert a_1, d_1, p_1]$	$=$	0.6600	$E_{x_1,h_1}[Y \vert a_1, d_1, p_1]$	$=$	0.3762

Table 7: Transition distributions and the immediate outcome for the learning problem of the dyspnoea treatment example.

$P_x(p\|d,a)$:	$a_{x_0,a_0,d_0}(p_0)$	$=$	0.3920	$b_{x_0,a_0,d_0}(p_0)$	$=$	0.7363
	$a_{x_1,a_0,d_0}(p_0)$	$=$	0.2933	$b_{x_1,a_0,d_0}(p_0)$	$=$	0.9490
	$a_{x_0,a_1,d_0}(p_0)$	$=$	0.0293	$b_{x_0,a_1,d_0}(p_0)$	$=$	0.9936
	$a_{x_1,a_1,d_0}(p_0)$	$=$	0.8942	$b_{x_1,a_1,d_0}(p_0)$	$=$	0.9299
	$a_{x_0,a_0,d_1}(p_0)$	$=$	0.3178	$b_{x_0,a_0,d_1}(p_0)$	$=$	0.6620
	$a_{x_1,a_0,d_1}(p_0)$	$=$	0.2800	$b_{x_1,a_0,d_1}(p_0)$	$=$	0.9358
	$a_{x_0,a_1,d_1}(p_0)$	$=$	0.0251	$b_{x_0,a_1,d_1}(p_0)$	$=$	0.9894
	$a_{x_1,a_1,d_1}(p_0)$	$=$	0.8557	$b_{x_1,a_1,d_1}(p_0)$	$=$	0.8914
	$a_{x_0,a_0,d_0}(p_1)$	$=$	0.2637	$b_{x_0,a_0,d_0}(p_1)$	$=$	0.6080
	$a_{x_1,a_0,d_0}(p_1)$	$=$	0.0510	$b_{x_1,a_0,d_0}(p_1)$	$=$	0.7067
	$a_{x_0,a_1,d_0}(p_1)$	$=$	0.0064	$b_{x_0,a_1,d_0}(p_1)$	$=$	0.9707
	$a_{x_1,a_1,d_0}(p_1)$	$=$	0.0701	$b_{x_1,a_1,d_0}(p_1)$	$=$	0.1058
	$a_{x_0,a_0,d_1}(p_1)$	$=$	0.3380	$b_{x_0,a_0,d_1}(p_1)$	$=$	0.6822
	$a_{x_1,a_0,d_1}(p_1)$	$=$	0.0642	$b_{x_1,a_0,d_1}(p_1)$	$=$	0.7200
	$a_{x_0,a_1,d_1}(p_1)$	$=$	0.0106	$b_{x_0,a_1,d_1}(p_1)$	$=$	0.9749
	$a_{x_1,a_1,d_1}(p_1)$	$=$	0.1086	$b_{x_1,a_1,d_1}(p_1)$	$=$	0.1443
$P_x(d,p\|a)$:	$a_{x_0,a_0}(d_0,p_0)$	$=$	0.1032	$b_{x_0,a_0}(d_0,p_0)$	$=$	0.4475
	$a_{x_1,a_0}(d_0,p_0)$	$=$	0.0772	$b_{x_1,a_0}(d_0,p_0)$	$=$	0.7330
	$a_{x_0,a_1}(d_0,p_0)$	$=$	0.0122	$b_{x_0,a_1}(d_0,p_0)$	$=$	0.9765
	$a_{x_1,a_1}(d_0,p_0)$	$=$	0.3712	$b_{x_1,a_1}(d_0,p_0)$	$=$	0.4069
	$a_{x_0,a_0}(d_1,p_0)$	$=$	0.2341	$b_{x_0,a_0}(d_1,p_0)$	$=$	0.5783
	$a_{x_1,a_0}(d_1,p_0)$	$=$	0.2063	$b_{x_1,a_0}(d_1,p_0)$	$=$	0.8620
	$a_{x_0,a_1}(d_1,p_0)$	$=$	0.0147	$b_{x_0,a_1}(d_1,p_0)$	$=$	0.9790
	$a_{x_1,a_1}(d_1,p_0)$	$=$	0.5005	$b_{x_1,a_1}(d_1,p_0)$	$=$	0.5362
	$a_{x_0,a_0}(d_0,p_1)$	$=$	0.0694	$b_{x_0,a_0}(d_0,p_1)$	$=$	0.4137
	$a_{x_1,a_0}(d_0,p_1)$	$=$	0.0134	$b_{x_1,a_0}(d_0,p_1)$	$=$	0.6692
	$a_{x_0,a_1}(d_0,p_1)$	$=$	0.0027	$b_{x_0,a_1}(d_0,p_1)$	$=$	0.9669
	$a_{x_1,a_1}(d_0,p_1)$	$=$	0.0027	$b_{x_1,a_1}(d_0,p_1)$	$=$	0.0648
	$a_{x_0,a_0}(d_1,p_1)$	$=$	0.2490	$b_{x_0,a_0}(d_1,p_1)$	$=$	0.5932
	$a_{x_1,a_0}(d_1,p_1)$	$=$	0.0473	$b_{x_1,a_0}(d_1,p_1)$	$=$	0.7030
	$a_{x_0,a_1}(d_1,p_1)$	$=$	0.0062	$b_{x_0,a_1}(d_1,p_1)$	$=$	0.9705
	$a_{x_1,a_1}(d_1,p_1)$	$=$	0.0635	$b_{x_1,a_1}(d_1,p_1)$	$=$	0.0992

Table 8: Causal bounds for the transition probabilities $P_x(p|d,a) \in [a_{x,a,d}(p), b_{x,a,d}(p)]$ and $P_x(d,p|a) \in [a_{x,a}(d,p), b_{x,a}(d,p)]$ in the dyspnoea treatment example.