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. Clarification note. Even though both authors contributed significantly to this work, only the first one may appear in some version of this manuscript due to a glitch during the submission. See (Zhang & Bareinboim, 2020), for the full report. 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 $\mathtt{Proj}(\mathcal{G}, \boldsymbol{S})$ 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, $\mathtt{Proj}(\mathcal{G}, \{X_2, Y\})$ 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 $X_1 \leftrightarrow S_1 \leftrightarrow X_2$.

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 $M^*_{\sigma_{\boldsymbol{X}}}$,

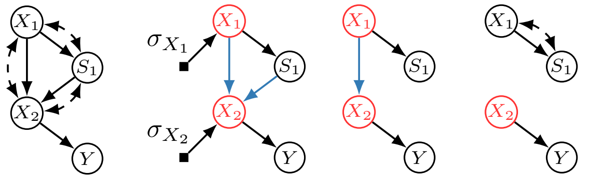
$$P_{\sigma_{\boldsymbol{X}}}(\boldsymbol{v}) = \sum_{\boldsymbol{u}} P(\boldsymbol{u}) \prod_{V \notin \boldsymbol{X}} P(v|pa_V, \boldsymbol{u}_V) \prod_{X \in \boldsymbol{X}} \sigma_X(x|h_X).$$

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 $[\![\mathcal{G}, \Pi, Y]\!]$.



(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 $\Pi_{\text{MIN}} = \{\mathfrak{D}_\emptyset \mapsto \mathfrak{D}_{X_2}\}$.

2.1. 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 $\llbracket \mathcal{G}, \Pi, Y \rrbracket$, 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* $\llbracket \mathcal{G}, \Pi, Y \rrbracket$, $\Pi \setminus \{\tilde{\boldsymbol{S}} \mapsto X\}$ *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 $\llbracket \mathcal{G}, \Pi, Y \rrbracket$, 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* $\llbracket \mathcal{G}, \Pi, Y \rrbracket$, *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 $\llbracket \mathcal{G}, \Pi, Y \rrbracket$, 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* $\llbracket \mathcal{G}, \Pi, Y \rrbracket$*, 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* $\llbracket \mathcal{G}, \Pi, Y \rrbracket$, `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 $\llbracket \mathcal{G}, \Pi, Y \rrbracket$.
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 $i = n, \ldots, 1$ 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

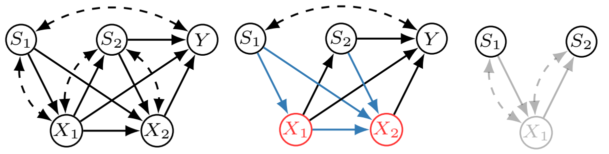
The goal of this section is to design online RL algorithms that find the optimal DTR $\sigma^*_{\boldsymbol{X}}$ in an unknown SCM M^* based solely on the information summarized in $\llbracket \mathcal{G}, \Pi, Y \rrbracket$.

An online learning algorithm learns the underlying system dynamics of M^* through repeated episodes of interactions $t = 1, \ldots, T$. 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

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

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
       $$\sigma^t_{\boldsymbol{X}} = \operatorname*{arg\,max}_{\sigma_{\boldsymbol{X}} \in \Pi} \max_{P^t_{\boldsymbol{x}}(\boldsymbol{s}) \in \mathcal{P}_t} V_{\sigma_{\boldsymbol{X}}}(P^t_{\boldsymbol{x}}(\boldsymbol{s})) \qquad (3)$$
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,

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

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 longitudinal settings. For instance, in Fig. 2a, $P_{x_1}(s_1, s_2)$ is not identifiable due to the presence of UCs (i.e., $X_1 \leftrightarrow S_2$). Let $\boldsymbol{S} = \{S_2\}$ and $\boldsymbol{C} = \{S_1, S_2, X_1\}$. Lem. 5 allows us to bound $P_{x_1}(s_2)$ from $P(s_1, s_2, x_1)$ as $P_{x_1}(s_2) \geq \max_{s_1} P(s_1, s_2, x_1)$ and $P_{x_1}(s_2) \leq \min_{s_1} P(s_1, s_2, x_1) - P(s_1, x_1) + 1$.

However, the bounds in Lem. 5 could be improved by exploiting the independence relationships among variables in $\boldsymbol{S}$. Consider again the example in Fig. 2a. Variables S_1 and S_2 are independent under $do(x_1)$ (as shown in Fig. 2c). That is, $P_{x_1}(s_2) = P_{x_1}(s_2|s_1) = P_{x_1}(s_2, s_1)/P(s_1)$. This equation, together with Lem. 5, allows us to bound $P_{x_1}(s_2)$ from $P(s_1, s_2, x_1)$ as follows: $P_{x_1}(s_2) \geq \max_{s_1} P(x_1, s_2|s_1)$ and $P_{x_1}(s_2) \leq \min_{s_1} P(x_1, s_2|s_1) - P(x_1|s_1) + 1$. Since $P(s_1) \in [0, 1]$, it is immediate to see that such bounds are tighter than those derived from Lem. 5 alone, without utilizing the independence relationship $(S_1 \perp\!\!\!\perp S_2)_{\mathcal{G}[\{S_1, S_2\}]}$. Our next result applies this intuition to bound transitional probabilities $P_{\bar{\boldsymbol{x}}_k}(s_k|\bar{\boldsymbol{s}}_k \setminus \{s_k\})$ from the observational data $P(\boldsymbol{v})$ in general settings.

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 extends the DTR bounds in (Zhang & Bareinboim, 2019) to an arbitrary causal diagram.

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 $\mathfrak{P}_c$ 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 $\mathfrak{P}_c \cap \mathfrak{P}_t$. That is, we replace the optimization problem defined in Eq. (3) with the following:

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

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).

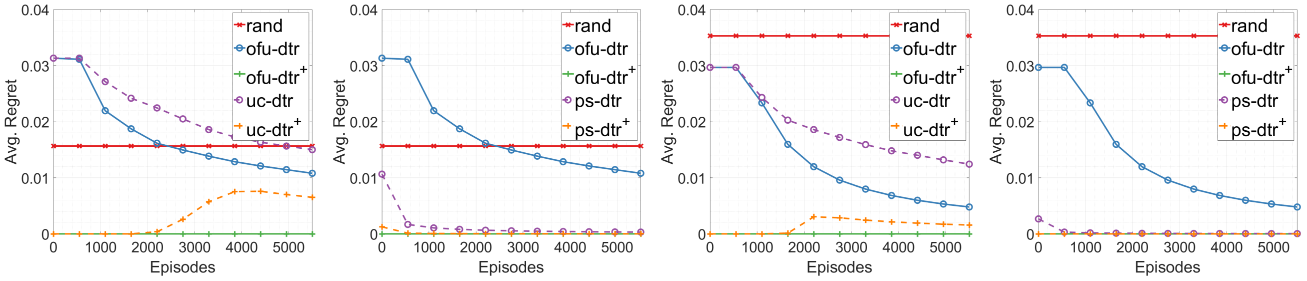
Theorem 6. *Given $[\![\mathcal{G}, \Pi, Y]\!]$ and causal bounds $\mathfrak{C}$, fix a $\delta \in (0, 1)$. W.p. at least $1 - \delta$, it holds for any $T > 1$, the regret of `OFU-DTR` is bounded by*

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

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 $T <$



(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*$^+$).

$12^2|\mathfrak{D}_{\bar{\boldsymbol{S}}_k\cup\bar{\boldsymbol{X}}_k}|T\log(|\boldsymbol{S}|T/\delta)/|\mathfrak{C}_{S_k}|^2$ 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 $\mathfrak{C}$ 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 $\llbracket\mathcal{G},\Pi,Y\rrbracket$, 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*

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

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 $T=5.5\times10^3$ 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`.