Towards Safe Policy Learning under Partial Identifiability: A Causal Approach

Shalmali Joshi[*1], **Junzhe Zhang**[*2], **Elias Bareinboim**[2]

[1]Department of Biomedical Informatics
[2]Department of Computer Science
Columbia University
New York, NY

Abstract

Learning personalized treatment policies is a formative challenge in many real-world applications, including in healthcare, econometrics, artificial intelligence. However, the effectiveness of candidate policies is not always *identifiable*, i.e., it is not uniquely computable from the combination of the available data and assumptions about the generating mechanisms. This paper studies policy learning from data collected in various non-identifiable settings, i.e., (1) observational studies with unobserved confounding; (2) randomized experiments with partial observability; and (3) their combinations. We derive sharp, closed-formed bounds from observational and experimental data over the conditional treatment effects. Based on these novel bounds, we further characterize the problem of safe policy learning and develop an algorithm that trains a policy from data guaranteed to achieve, at least, the performance of the baseline policy currently deployed. Finally, we validate our proposed algorithm on synthetic data and a large clinical trial, demonstrating that it guarantees safe behaviors and robust performance.

Introduction

Learning optimal personalized treatment policies that maximize a primary outcome by drawing insights from a fixed dataset is a ubiquitous challenge in many real-world applications, including in healthcare, social science, robotics. Several conditions and algorithms have been proposed to solve this problem, including reinforcement learning (Strehl et al. 2010; Li, Munos, and Szepesvari 2015; Swaminathan and Joachims 2015; Levine et al. 2020) and causal inference (Murphy 2003, 2005; Chakraborty and Murphy 2014). Most of these algorithms require the critical assumption of *no unmeasured confounding* (NUC) (Robins 1997), also known as *unconfoundedness*, *ignorability* (Rubin 1974; Rosenbaum and Rubin 1983), or *backdoor* admissibility (Pearl 2009, Def. 3.3.1). This requires that the treatment allocation policy that generates the data considers only the observed covariates; no unobserved confounder affects the treatment and outcome simultaneously. However, the NUC assumption could be fragile and does not necessarily hold in consequential domains with human interactions. For example, when learning personalized medicine from electronic health records (EHR), the physician might unintentionally prescribe a new drug to patients with access to better healthcare, making the drug appear more effective than it is.

A common remedy for the presence of UCs is to perform direct experimentation. The NUC assumption could be made to hold by directly controlling the treatment assignment in specific environments, as sometimes done in randomized trials (Fisher 1926) and online reinforcement learning (Sutton and Barto 2018). Still, the challenge of policy evaluation could arise when experimental data are *partially observed*, i.e., it lacks critical measurements or has a mismatch in the measured covariates that will be used as input for candidate policies. For example, the Affordable Care Act required hospitals to collect demographic variables such as race that were not routinely collected before 2010 (Nuru-Jeter et al. 2018; of Sciences Engineering, Medicine et al. 2016). Consequently, even with randomized controlled trials being performed and the NUC holds, one could not learn a personalized policy to treat patients that accounts for race using past medical data.

Broadly, causal inference provides a collection of principles and tools for evaluating the effects of policies from the combination of data and structural assumptions about the environment (Pearl 2009; Spirtes et al. 2000; Bareinboim and Pearl 2016). There exist conditions and algorithms to infer the effect of a new intervention from observational studies by leveraging knowledge encoded in its causal models (Pearl 1995; Tian and Pearl 2002; Tian 2002; Shpitser and Pearl 2006; Huang and Valtorta 2006). Further, causal effects can also be inferred from randomized experiments with a mismatch in the intervened treatments (Bareinboim and Pearl 2012; Lee, Correa, and Bareinboim 2019) and the measured covariates (Lee and Bareinboim 2020). Recent advancements also lead to complete algorithmic solutions to combine observational and experimental data to identify causal effects (Lee, Correa, and Bareinboim 2019). However, when the unobserved confounders (UCs) generally exist, and critical covariates are partially observed, the effects of treatment policies are not necessarily always identifiable (Pearl 2009, Def. 3.2.3). The treatment effects may not be not uniquely computable from data, despite extensive synthesis and analysis of numerous samples collected across multiple regimes.

[*]These authors contributed equally.

Evaluating non-identifiable treatment effects from the combination of data and assumptions has been studied under the rubrics of *partial identification*. There is a growing body of work in causal inference (Manski 1990; Robins 1989; Balke and Pearl 1995; Chickering and Pearl 1996; Fan and Park 2010; Richardson et al. 2014; Zhang and Bareinboim 2021; Zhang, Tian, and Bareinboim 2022; Gresele et al. 2022), and more recently, in machine learning (Kallus, Puli, and Shalit 2018; Namkoong et al. 2020) tackling this challenge. Among these works, one of the following approaches is employed (not exclusively): (1) bounds on the treatment effects are estimated; (2) additional parametric assumptions are invoked, and sensitivity analysis is conducted to assess how treatment effects change as parametric assumptions are perturbed. While cases exist where the partial identification analysis lead to a particular treatment recommendation (Cornfield et al. 1959), there is no safety guarantee for the recommended treatment's performance. Our goal in AI is to build intelligent systems that can reason and act autonomously, which means we need to move from a heuristic understanding of the interplay between partial identification and policy learning to a more principled understanding of a robust decision-making process. There are still significant challenges in policy learning under non-identifiability.

This paper aims to overcome these challenges and develop a framework for safe policy learning through causal lenses. In particular, from a fixed dataset, train a policy that is guaranteed to perform as well as a baseline policy currently deployed in the environment (Thomas, Theocharous, and Ghavamzadeh 2015; Ghavamzadeh, Petrik, and Chow 2016; Kallus and Zhou 2018). This framework supports evidence-based medicine since the learner can validate whether the treatment policy significantly improves the standard of current care without direct interaction with the patients. Closet to our work, Kallus and Zhou (2018) studied the problem of confounding-robust policy improvement that optimizes a policy to achieve the best worst-case improvement relative to a baseline policy. This method computes a policy recommendation based only on observational data. Our work, instead, will account for additional data collected from controlled experiments and explore the nuanced and fundamental interplays between the observational and experimental data on policy evaluation in non-identifiable settings. For a more detailed survey of the related work, we refer readers to (Joshi, Zhang, and Bareinboim 2023, Sec. 9).

This paper departs from existing approaches and studies safe policy learning in several non-identifiable settings, including learning from observational studies with unobserved confounders, past randomized experiments with partial observability, and combining the two. Our contributions are summarized as follows. (1) We derive closed-form bounds on effect estimates conditioned on new features that combine observational and experimental data (collected with limited context). (2) We prove these bounds are sharp (i.e., cannot be improved without additional assumptions) and identify sufficient conditions when bounds will improve over the purely observational setting. (3) We formulate two objectives and propose a new notion of safe policy learning that leverages these bounds, deviating from worst-case approaches explored in literature. Finally, the proposed approach is evaluated in the synthetic dataset and a large clinical trial. Due to space constraints, all proofs and details on the experiment setup are in the complete technical report (Joshi, Zhang, and Bareinboim 2023, Secs. 10 & 12).

Preliminaries

This section will introduce basic notation and definitions used in this paper and provide a short review of related work. We use capital letters X to indicate random variables and lowercase letters x to indicate their realizations. Bold-face capital letters indicate multivariate random variables. Domain of a random variable X is denoted by Ω_X and its cardinality by $|\Omega_X|$. $P(X)$ indicates the probability distribution of X and $P(x)$ the probability that $X = x$. Let $\mathbf{1}_{X=x}$ denote an indicator function that takes the value 1 if X realizes to x and is 0 otherwise. We use $[K]$ to denote the set $\{1, 2, \cdots, K\}$. We indicate the event $X \neq x$ with the shorthand notation $\neg x$.

We use Structural Causal Models (SCM) as the basic semantical framework to represent data-generating mechanisms (Pearl 2009). An SCM is a tuple $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P \rangle$ where V are the observed random variables in the system, and $\mathbf{U}$ are unobserved exogenous variables that introduce stochasticity in the system. Dependence across observed variables is governed by functional relationships $\mathcal{F}$. That is, for every $V \in \mathbf{V}$, $v \leftarrow f_V(\mathbf{pa}_V, \mathbf{u}_V)$ denotes the values of V will be determined by the function $f_V \in \mathcal{F}$ taking as input a set of observed parents $\mathbf{PA}_V \subseteq \mathbf{V}$ and unobserved parents $\mathbf{U}_V \subseteq \mathbf{U}$. Values of unobserved variable $\mathbf{U}$ are drawn from an exogenous distribution $P(\mathbf{U})$. Naturally, every SCM $\mathcal{M}$ defines an *observational distribution* $P(\mathbf{V})$ over endogenous variables $\mathbf{V}$ (Bareinboim et al. 2020, Def. 2). The SCM can be more coarsely represented as a causal diagram $\mathcal{G}$, which is a directed acyclic graph with solid nodes representing observed variables ($\mathbf{V}$), empty nodes for unobserved variables ($\mathbf{U}$), and directed edges codifying the causal dependencies to the parents.

An intervention on a set of observed nodes $\mathbf{X} \subseteq \mathbf{V}$, denoted by $\text{do}(\mathbf{x})$, is an operation that anchors realizations of $\mathbf{X}$ to constants $\mathbf{x}$, removing the dependence on the parents (and exogenous nodes). The $\text{do}(\cdot)$ operation mechanistically allows us to measure the causal effect of the intervened variables $\mathbf{X}$ on the other observed variables $\mathbf{V} \setminus \mathbf{X}$. We will denote the original SCM by $\mathcal{M}$ and the intervened SCM (after a do operation) as $\mathcal{M}_{\mathbf{x}}$. The interventional distribution $P_{\mathbf{x}}(\mathbf{Y})$ is defined as the distribution over $\mathbf{Y}$ in the submodel $\mathcal{M}_{\mathbf{x}}$, i.e., $P_{\mathbf{x}}(\mathbf{Y}) \triangleq P_{\mathcal{M}_{\mathbf{x}}}(\mathbf{Y})$ (Bareinboim et al. 2020, Def. 5). We denote by $P_{\mathbf{X}}(\mathbf{Y})$ a collection of interventional distributions $\{P_{\mathbf{x}}(\mathbf{Y}) \mid \forall \mathbf{x} \in \Omega_{\mathbf{X}}\}$. Potential outcomes $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ are solutions for a set of observed variables $\mathbf{Y} \subseteq \mathbf{V}$ evaluated in the intervened SCM $\mathcal{M}_{\mathbf{x}}$ after intervention on $\mathbf{x}$. Fix a value $\mathbf{y} \in \Omega_{\mathbf{Y}}$. Let $\mathbf{y}_{\mathbf{x}}$ denote an event $\mathbf{Y}_{\mathbf{x}} = \mathbf{y}$. For a set of variables $\mathbf{X}, \dots \mathbf{W}, \mathbf{Y}, \dots, \mathbf{Z}$, the counterfactual distribution $P(\mathbf{Y}_{\mathbf{x}}, \dots, , \mathbf{Z}_{\mathbf{w}})$ is a joint distribution over potential outcomes $\mathbf{Y}_{\mathbf{x}}, \dots, , \mathbf{Z}_{\mathbf{w}}$ in SCM $\mathcal{M}$, given by $P(\mathbf{y}_{\mathbf{x}}, \dots, \mathbf{z}_{\mathbf{w}}) = \sum_{\mathbf{u}} \mathbf{1}_{\mathbf{Y}_{\mathbf{x}}(\mathbf{u})=\mathbf{y}, \dots, \mathbf{Z}_{\mathbf{w}}(\mathbf{u})=\mathbf{z}} P(\mathbf{u})$ (Bareinboim et al. 2020, Def. 7).

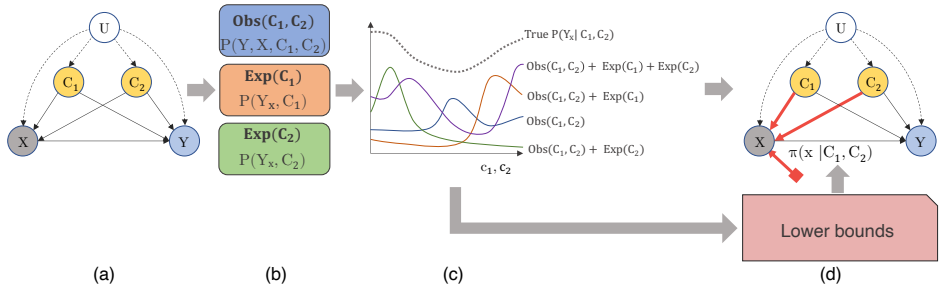

Figure 1: Pipeline of the proposed safe policy learning framework. (a) Causal diagram $\mathcal{G}$ for the underlying SCM $\mathcal{M}^*$ with confounding attributions $\mathbf{C}_1$ and $\mathbf{C}_2$. (b) Available data drawn from the observational $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$ and interventional $P_X(Y, \mathbf{C}_1), P_X(Y, \mathbf{C}_2)$ distributions. (c) *Lower bounds* for the true treatment effect of x on Y obtained from different combinations of data. (d) A safe policy $\pi(X \mid \mathbf{C}_1, \mathbf{C}_2)$ optimizing the worst-case treatment effect.

Safe Policy Learning under Partial Identifiability

We will study the problem of optimizing an action X based on values of observed covariates $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2\}$ to maximize a primary outcome (i.e., reward) Y in an SCM $\mathcal{M}^*$. Fig. 1 (a) shows the causal diagram $\mathcal{G}$ associated with this SCM, where unobserved confounders $\mathbf{U}$ exist affecting the action X, outcome Y, and covariates $\mathbf{C}_1, \mathbf{C}_2$, simultaneously. This class of environmental models is also referred to as contextual bandit (Langford and Zhang 2008) and is widely applied in reinforcement learning literature (Li et al. 2010). Throughout this paper, we will consistently assume domains of variables $X, Y, \mathbf{C}_1, \mathbf{C}_2$ are discrete and finite; both $\mathbf{C}_1$ and $\mathbf{C}_2$ can be high-dimensional, i.e., $|\Omega_{\mathbf{C}_i}| \gg |\Omega_X|, |\Omega_Y|$ for $i = 1, 2$.

A policy $\pi(X \mid \mathbf{C}_1, \mathbf{C}_2)$ is a function mapping from domains of covariates $\mathbf{C}_1, \mathbf{C}_2$ to the space of probability distribution over the action domain X. The collection of such policies defines a policy space Π. An intervention on action X following the policy π, denoted by $\text{do}(\pi)$, induces an interventional distribution $P_\pi(X, Y, \mathbf{C}_1, \mathbf{C}_2)$ given by

$$P_\pi(x, y, \mathbf{c}_1, \mathbf{c}_2) = P_x(y, \mathbf{c}_1, \mathbf{c}_2)\pi(x \mid \mathbf{c}_1, \mathbf{c}_2) \qquad (1)$$

The expected reward associated with a policy $\pi(X \mid \mathbf{C}_1, \mathbf{C}_2)$ is thus given by

$$\mathbb{E}_\pi[Y] = \sum_{x, y, \mathbf{c}_1, \mathbf{c}_2} y P_x(y, \mathbf{c}_1, \mathbf{c}_2)\pi(x \mid \mathbf{c}_1, \mathbf{c}_2) \qquad (2)$$

The agent is interested in learning a policy π that maximizes the expected reward $\mathbb{E}_\pi[Y]$ evaluated in SCM $\mathcal{M}^*$. When detailed parametrization of the SCM $\mathcal{M}^*$ is provided, efficient planning algorithms exist to solve for an optimal policy (Bellman 1957; Sutton and Barto 2018). In many practical applications, however, the environment's system dynamics $P_x(y, \mathbf{c}_1, \mathbf{c}_2)$ are assumed to be unknown. Instead, the learner can access data collected from the SCM $\mathcal{M}^*$ under different regimes (observational studies or randomized experiments). We assume access to the following data sources:

1. **Observational Data Obs$(\mathbf{C}_1, \mathbf{C}_2)$.** An observational study is performed to collected samples $\text{Obs}(\mathbf{C}_1, \mathbf{C}_2)$ drawn from the observational distribution $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$. For convenience, since observational data always uses both sets of covariates $\mathbf{C}_1, \mathbf{C}_2$, we denote $\text{Obs}(\mathbf{C}_1, \mathbf{C}_2) \equiv \text{Obs}$.

2. **Experimental Data Exp$(\mathbf{C}_1)$, Exp$(\mathbf{C}_2)$.** Randomized controlled trials (RCTs) are conducted on subjects (e.g., patients) from a population $\mathbf{C}_1 = \mathbf{c}_1$ or $\mathbf{C}_2 = \mathbf{c}_2$, but never the combination of the two. This means covariates $\mathbf{C}_1, \mathbf{C}_2$ are never observed simultaneously in experimental data. Consequently, experimental data can come in two forms: $\text{Exp}(\mathbf{C}_1) \sim P_X(Y, \mathbf{C}_1)$ or $\text{Exp}(\mathbf{C}_2) \sim P_X(Y, \mathbf{C}_2)$.

The key challenge in evaluating a policy $\pi(X \mid \mathbf{C}_1, \mathbf{C}_2)$ is to find a function that recovers the expected reward $\mathbb{E}_\pi[Y]$ from $\text{Obs}(\mathbf{C}_1, \mathbf{C}_2)$, $\text{Exp}(\mathbf{C}_1)$, $\text{Exp}(\mathbf{C}_2)$ in all possible SCMs $\mathcal{M}$ generating the data. However, classic results in the causal identification suggest this is infeasible (Lee, Correa, and Bareinboim 2019; Lee and Bareinboim 2020).

Corollary 1. *The interventional distribution $P_\pi(Y)$ is not identifiable from $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$, $P_X(Y, \mathbf{C}_1)$, and $P_X(Y, \mathbf{C}_2)$ in contextual bandits. That is, there exists SCMs $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ compatible with Fig. 1 (a) such that $P^{(1)}(X, Y, \mathbf{C}_1, \mathbf{C}_2) = P^{(2)}(X, Y, \mathbf{C}_1, \mathbf{C}_2)$, $P^{(1)}_X(Y, \mathbf{C}_1) = P^{(2)}_X(Y, \mathbf{C}_1)$, $P^{(1)}_X(Y, \mathbf{C}_2) = P^{(2)}_X(Y, \mathbf{C}_2)$ while $P^{(1)}_\pi(Y) \neq P^{(2)}_\pi(Y)$ for some policies $\pi(X \mid \mathbf{C}_1, \mathbf{C}_2)$.*

In words, one cannot uniquely determine the expected reward $\mathbb{E}_\pi[Y]$ from any combination of the observational and interventional distributions $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$, $P_X(Y, \mathbf{C}_1)$, $P_X(Y, \mathbf{C}_2)$, regardless of how many samples are collected. This result seems to suggest that when the expected reward is not identifiable from available data, it is impossible to learn a policy with satisfactory performance.

To address this challenge, we now formulate the safe policy learning problem. Instead of learning an optimal policy maximizing the expected reward, the agent attempts to obtain a robust policy to achieve a specific baseline performance τ. Let $\mathbb{M}$ be the set of SCMs $\mathcal{M}$ compatible with distributions $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$, $P_X(Y, \mathbf{C}_1)$, and $P_X(Y, \mathbf{C}_2)$. Formally, a robust policy π^* is given by

$$\pi^* = \arg\max_{\pi \in \Pi} \min_{\mathcal{M} \in \mathbb{M}} \underbrace{\mathbb{E}_\pi[Y; \mathcal{M}]}_{\text{Worst-case effect}} - \underbrace{\mathbb{E}[Y; \mathcal{M}^*]}_{\text{Baseline performance } \tau} \qquad (3)$$

Among quantities in the above maximin program, the inner minimization in the first term computes the worst-case treatment effect of a policy $\pi(X \mid \mathbf{C}_1, \mathbf{C}_2)$. Naturally, the solution $\min_{\mathcal{M}} \mathbb{E}_\pi[Y; \mathcal{M}] \leq \mathbb{E}_\pi[Y; \mathcal{M}^*]$ is a lower bound for the expected reward for policy π evaluated in the true SCM $\mathcal{M}^*$. The second term is the baseline performance $\tau = \mathbb{E}[Y; \mathcal{M}^*]$ achieved by the behavioral policy $X \leftarrow f_X(\mathbf{C}_1, \mathbf{C}_2, \mathbf{U})$ that generates the observational data in the underlying $\mathcal{M}^*$.[1] We show in Fig. 1 a graphical representation of our problem setup.

Partial Identification from a Single Distribution

Note that in the maximin program of Eq. (3), the baseline $\mathbb{E}[Y; \mathcal{M}^*] = \mathbb{E}[Y]$ is an observational quantity and is computable from distribution $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$. Following Eq. (2) and the convexity of a minimum function, the worst-case treatment effect could be further written as:

$$\min_{\mathcal{M} \in \mathbb{M}} \mathbb{E}_\pi[Y; \mathcal{M}] \tag{4}$$

$$\geq \sum_{x, y, \mathbf{c}_1, \mathbf{c}_2} \pi(x \mid \mathbf{c}_1, \mathbf{c}_2) y \min_{\mathcal{M} \in \mathbb{M}} P_x(y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}) \tag{5}$$

It is thus sufficient to consider the problem of bounding interventional probabilities $P_x(y, \mathbf{c}_1, \mathbf{c}_2)$ from distributions $P(X, Y, \mathbf{C}_1, \mathbf{C}_2), P_X(Y, \mathbf{C}_1)$, and $P_X(Y, \mathbf{C}_2)$. Formally,

Definition 1 (Lower Causal Bound)**.** Let $\mathcal{G}$ be a causal diagram over variables $\mathbf{V}$, $\mathcal{P}$ be a set of distributions (observational or interventional) over $\mathbf{V}$, and $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ be (disjoint) subsets of $\mathbf{V}$. A lower bound over the causal effects $P_{\mathbf{X}}(\mathbf{Y})$ is *an expression for a function* $l(\mathbf{x}, \mathbf{y})$ in terms of $\mathcal{P}$ such that for every SCM $\mathcal{M}$ compatible with $\mathcal{G}$, $P_{\mathbf{x}}(\mathbf{y}; \mathcal{M}) \geq l(\mathbf{x}, \mathbf{y}; \mathcal{M}), \forall (\mathbf{x}, \mathbf{y}) \in \Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}$.

First, a function of the observational distribution $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$ that consistently lower bounds $P_x(y, \mathbf{c}_1, \mathbf{c}_2)$ in all SCM $\mathcal{M}$ compatible with Fig. 1 (a), called the *natural bound* (Manski 1990),

$$P_x(y, \mathbf{c}_1, \mathbf{c}_2) \geq P(x, y, \mathbf{c}_1, \mathbf{c}_2) \tag{6}$$

Interestingly, it can be shown that the marginal interventional distribution $P_X(Y, \mathbf{C}_1)$ or $P_X(Y, \mathbf{C}_2)$ does not impose any informative constraint over the joint distribution $P_X(Y, \mathbf{C}_1, \mathbf{C}_2)$. Consider, for example, $\mathbf{C}_1 = \mathbf{c}_1$. One could always construct an SCM $\mathcal{M}$ compatible with Fig. 1 (a) such that $P_x(y, \mathbf{c}_1; \mathcal{M}) = P_x(y, \mathbf{c}_1, \neg \mathbf{c}_2; \mathcal{M})$ and $P_x(y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}) = 0$. This implies a lower bound

$$P_x(y, \mathbf{c}_1, \mathbf{c}_2) \geq 0 \tag{7}$$

So far, our analysis reveals that marginal interventional distributions $P_X(Y, \mathbf{C}_1)$ or $P_X(Y, \mathbf{C}_2)$ do not impose any meaningful constraint over the target effect $P_x(y, \mathbf{c}_1, \mathbf{c}_2)$. This seems to suggest when computing the worst-case treatment effect, it is sufficient to consider only the observational distribution $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$. For the remainder of this paper, we will show this is not the case by investigating nontrivial interactions between the observational and interventional distributions.

[1] More generally, the baseline performance $\tau \in \mathbb{R}$ could be an arbitrary real value based on the context. This paper focuses on finding a robust policy that improves over the policy f_X currently deployed in the environment.

Partial Identification from Multiple Distributions

This section will derive novel lower bounds over the target causal effects $P_X(Y, \mathbf{C}_1, \mathbf{C}_2)$ from the combination of observational and interventional distributions $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$, $P_X(Y, \mathbf{C}_1)$, $P_X(Y, \mathbf{C}_2)$ in the models compatible with the causal diagram of Fig. 1 (a). We start with a novel lower bound over the target effects by combining the observational distribution $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$ and a marginal interventional distribution $P_X(Y, \mathbf{C}_1)$ over partial covariates $\mathbf{C}_1$.

Lemma 1 (Obs + Exp($\mathbf{C}_1$))**.** *Given distributions $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$ and $P_X(Y, \mathbf{C}_1)$, the lower bound over $P_x(y, \mathbf{c}_1, \mathbf{c}_2)$ for all $(x, y, \mathbf{c}_1, \mathbf{c}_2) \in \Omega_X \times \Omega_Y \times \Omega_{\mathbf{C}_1} \times \Omega_{\mathbf{C}_2}$ is given by*

$$P_x(y, \mathbf{c}_1, \mathbf{c}_2) \geq \max\{l_1(x, y, \mathbf{c}_1, \mathbf{c}_2), l_2(x, y, \mathbf{c}_1, \mathbf{c}_2)\} \tag{8}$$

where l_1, l_2 are functions defined as

$$l_1(x, y, \mathbf{c}_1, \mathbf{c}_2) = P(x, y, \mathbf{c}_1, \mathbf{c}_2) \tag{9}$$

$$l_2(x, y, \mathbf{c}_1, \mathbf{c}_2) = P_x(y, \mathbf{c}_1) - P(x, y, \mathbf{c}_1, \neg \mathbf{c}_2) - P(\neg x, \mathbf{c}_1, \neg \mathbf{c}_2) \tag{10}$$

Among the quantities in Lem. 1, $l_1(x, y, \mathbf{c}_1, \mathbf{c}_2)$ is the natural bound, but $l_2(x, y, \mathbf{c}_1, \mathbf{c}_2)$ is a function of both the observational $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$ and interventional $P_X(Y, \mathbf{C}_1)$ distribution. It follows immediately that the bound in Lem. 1 is never inferior to the natural bound.

Definition 2. Let $\mathcal{G}$ be a causal diagram over variables $\mathbf{V}$, $\mathcal{P}$ be a set of distributions over $\mathbf{V}$, and $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$. For lower bounds $l_1(\mathbf{x}, \mathbf{y})$ and $l_2(\mathbf{x}, \mathbf{y})$ over the causal effects $P_{\mathbf{X}}(\mathbf{Y})$, $l_1(\mathbf{x}, \mathbf{y})$ is said to *consistently dominate* $l_2(\mathbf{x}, \mathbf{y})$ if the following conditions hold:

(i) For every SCM $\mathcal{M}$ compatible with $\mathcal{G}$, $l_1(\mathbf{x}, \mathbf{y}; \mathcal{M}) \geq l_2(\mathbf{x}, \mathbf{y}; \mathcal{M}), \forall (\mathbf{x}, \mathbf{y}) \in \Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}$.

(ii) There is an SCM $\mathcal{M}$ compatible with $\mathcal{G}$ s.t. $l_1(\mathbf{x}, \mathbf{y}; \mathcal{M}) > l_2(\mathbf{x}, \mathbf{y}; \mathcal{M}), \exists (\mathbf{x}, \mathbf{y}) \in \Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}$.

The more interesting question is whether instances exist where Lem. 1 is strictly tighter than the natural bound. For instance, consider an SCM $\mathcal{M}^*$ compatible with Fig. 1 (a) with exogenous variables $\mathbf{U} = \{U_1, U_2, U_3\}$ independently drawn over the binary domain $\{0, 1\}$ such that $P(U_1 = 0) = P(U_2 = 1) = 0.9$, $P(U_3 = 0) = 0.5$. Values of $X, Y, \mathbf{C}_1, \mathbf{C}_2$ in $\mathcal{M}^*$ are given by

$$X \leftarrow U_1 \oplus U_3, \qquad Y \leftarrow X \oplus U_1 \oplus U_2,$$
$$\mathbf{C}_1 \leftarrow U_1, \qquad \mathbf{C}_2 \leftarrow U_2 \tag{11}$$

where $\oplus$ is the "xor" operator. Evaluating the causal effect $P_x(y, \mathbf{c}_1, \mathbf{c}_2)$ in SCM $\mathcal{M}^*$ gives

$$P_{X=0}(Y = 1, \mathbf{C}_1 = 0, \mathbf{C}_2 = 1) = P(U_1 = 0, U_2 = 1) = 0.81 \tag{12}$$

Evaluating the natural bound $l_1(x, y, \mathbf{c}_1, \mathbf{c}_2)$ gives

$$l_1(X = 0, Y = 1, \mathbf{C}_1 = 0, \mathbf{C}_2 = 1) = P(U_1 = 0, U_2 = 1, U_3 = 0) = 0.405 \tag{13}$$

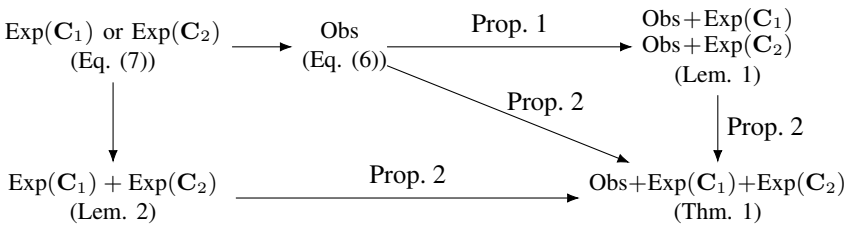

Figure 2: Hierarchy of lower bounds derived from different data sources. A directed path from $\mathcal{D}_1$ to $\mathcal{D}_2$ indicates the bound derived from dataset $\mathcal{D}_2$ consistently dominates the one from dataset $\mathcal{D}_1$.

On the other hand, evaluating the new bound $l_2(x, y, \mathbf{c}_1, \mathbf{c}_2)$ from Eq. (15) in SCM $\mathcal{M}^*$ gives

$$l_2(X = 0, Y = 1, \mathbf{C}_1 = 0, \mathbf{C}_2 = 1) \quad (14)$$
$$= P(U_1 = 0, U_2 = 1) - P(U_1 = 0, U_2 = 0, U_3 = 1)$$

Computing Eq. (14) gives. $l_2(X = 0, Y = 1, \mathbf{C}_1 = 0, \mathbf{C}_2 = 1) = 0.765$, which is larger than the natural bound. Recall that a single marginal distribution $P_X(Y, \mathbf{C}_1)$ does not impose any (lower) constraint on $P_x(y, \mathbf{c}_1, \mathbf{c}_2)$. Lem. 1 thus improves over the natural bound by exploring interactions between observational and interventional distributions.

Proposition 1. *Given distributions $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$ and $P_X(Y, \mathbf{C}_1)$, the lower bound given in Lem. 1 consistently dominates the natural bound (Eq.* (6)*).*

We also provide a lower bound computed from marginal distributions $P_X(Y, \mathbf{C}_1), P_X(Y, \mathbf{C}_2)$, i.e., the observational distribution $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$ is not available.

Lemma 2 (Exp($\mathbf{C}_1$) + Exp($\mathbf{C}_2$))**.** *Given distributions $P_X(Y, \mathbf{C}_1)$ and $P_X(Y, \mathbf{C}_2)$, the lower bound over $P_x(y, \mathbf{c}_1, \mathbf{c}_2)$ for all $(x, y, \mathbf{c}_1, \mathbf{c}_2) \in \Omega_X \times \Omega_Y \times \Omega_{\mathbf{C}_1} \times \Omega_{\mathbf{C}_2}$ is given by*

$$P_x(y, \mathbf{c}_1, \mathbf{c}_2) \geq P_x(y, \mathbf{c}_1) - P_x(y, \neg\mathbf{c}_2) \quad (15)$$

The above bound is informative if $P_x(y, \mathbf{c}_1) > P_x(y, \neg\mathbf{c}_2)$. However, there is no clear preference between the interventional bound in Lem. 2 and other bounds computed using the observational distribution, including the one in Lem. 1. Finally, we provide a novel bound utilizing all available data, including the observational and marginal interventional distributions over covariates $\mathbf{C}_1$, $\mathbf{C}_2$.

Theorem 1 (Obs + Exp($\mathbf{C}_1$) + Exp($\mathbf{C}_2$))**.** *Given distributions $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$, $P_X(Y, \mathbf{C}_1)$, and $P_X(Y, \mathbf{C}_2)$, the lower bound over $P_x(y, \mathbf{c}_1, \mathbf{c}_2)$ for all $(x, y, \mathbf{c}_1, \mathbf{c}_2) \in \Omega_X \times \Omega_Y \times \Omega_{\mathbf{C}_1} \times \Omega_{\mathbf{C}_2}$ is*

$$P_x(y, \mathbf{c}_1, \mathbf{c}_2) \geq \max\{l_1(x, y, \mathbf{c}_1, \mathbf{c}_2), l_2(x, y, \mathbf{c}_1, \mathbf{c}_2), l_3(x, y, \mathbf{c}_1, \mathbf{c}_2), l_4(x, y, \mathbf{c}_1, \mathbf{c}_2)\} \quad (16)$$

where l_1, l_2 are given by Eqs. (9) *and* (10)*, respectively; l_3, l_4 are functions defined as*

$$l_3(x, y, \mathbf{c}_1, \mathbf{c}_2) = P_x(y, \mathbf{c}_2) - P(x, y, \neg\mathbf{c}_1, \mathbf{c}_2) - P(\neg x, \neg\mathbf{c}_1, \mathbf{c}_2) \quad (17)$$

$$l_4(x, y, \mathbf{c}_1, \mathbf{c}_2) = P_x(y, \mathbf{c}_1) - P_x(y, \neg\mathbf{c}_2) + P(x, y, \neg\mathbf{c}_1, \neg\mathbf{c}_2) \quad (18)$$

Among quantities in Thm. 1, $l_3(x, y, \mathbf{c}_1, \mathbf{c}_2)$ is symmetric to $l_2(x, y, \mathbf{c}_1, \mathbf{c}_2)$, and follows from applying Lem. 1 with input $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$, $P_X(Y, \mathbf{C}_2)$. The constraint in $l_4(x, y, \mathbf{c}_1, \mathbf{c}_2)$ is a function of all available distributions $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$, $P_X(Y, \mathbf{C}_1)$, $P_X(Y, \mathbf{C}_2)$. One could see by inspection that Thm. 1 improves over the interventional bound in Lem. 2 if $P(x, y, \neg\mathbf{c}_1, \neg\mathbf{c}_2) > 0$. A more interesting question is how it compares with the bound given by Lem. 1. Consider again the SCM $\mathcal{M}^*$ described in Eq. (11). Evaluating the lower bound $l_4(x, y, \mathbf{c}_1, \mathbf{c}_2)$ gives

$$l_4(X = 0, Y = 1, \mathbf{C}_1 = 0, \mathbf{C}_2 = 1) \quad (19)$$
$$= P(U_1 = 0, U_2 = 1) - P(U_1 = 1, U_2 = 0, U_3 = 0)$$

Computing the above equation gives, $l_4(X = 0, Y = 1, \mathbf{C}_1 = 0, \mathbf{C}_2 = 1) = 0.805$, which consistently dominates lower bounds l_1, l_2 evaluated in Eqs. (13) and (14).

Proposition 2. *Given distributions $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$, $P_X(Y, \mathbf{C}_1)$, and $P_X(Y, \mathbf{C}_2)$, the lower bound given in Thm. 1 consistently dominates Lems. 1 and 2 and the natural bound (Eq.* (6)*).*

In Fig. 2 we summarize all the bounds derived in this paper and their relationships. A single marginal distribution $P_X(Y, \mathbf{C}_1)$ or $P_X(Y, \mathbf{C}_2)$ does not impose any constraint on the target effect $P_X(Y, \mathbf{C}_1, \mathbf{C}_2)$. One can derive meaningful bounds by incorporating the observational distribution $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$ or multiple interventional distributions $P_X(Y, \mathbf{C}_1)$ and $P_X(Y, \mathbf{C}_2)$. Finally, Thm. 1 presents the most informative bounds using all available data sources.

A natural question at this point is whether the bound provided in Thm. 1 is sharp, i.e., a tighter bound can be derived through a more refined analysis. Fortunately, we will show this is not the case.

Definition 3 (Sharp Lower Bound)**.** Let $\mathcal{G}$ be a causal diagram over variables $\mathbf{V}$, $\mathcal{P}$ be a set of distributions over $\mathbf{V}$, and $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$. A lower bound $l(\mathbf{x}, \mathbf{y})$ over the causal effects $P_{\mathbf{X}}(\mathbf{Y})$ from $\mathcal{P}$ is said to be *sharp* if there is no other lower bound $l^*(\mathbf{x}, \mathbf{y})$ that consistently dominates $l(\mathbf{x}, \mathbf{y})$.

Suppose the bound in Thm. 1, denoted by $l = \max\{l_1, l_2, l_3, l_4\}$, is not sharp, and there is a different lower bound l^* consistently dominates l. There must exist an SCM $\mathcal{M}$ compatible with Fig. 1 (a) and a realization $(x, y, \mathbf{c}_1, \mathbf{c}_2)$ such that $l^*(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}) > l(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M})$. The key challenge is to construct an alternative SCM $\mathcal{M}^*$ from $\mathcal{M}$ so that the lower bound l^* no longer applies.

Theorem 2. *Given distributions $P(X, Y, \mathbf{C}_1, \mathbf{C}_2)$, $P_X(Y, \mathbf{C}_1)$, and $P_X(Y, \mathbf{C}_2)$, Thm. 1 is a sharp lower bound over the causal effects $P_X(Y, \mathbf{C}_1, \mathbf{C}_2)$ in the causal diagram of Fig. 1 (a).*

Proof (sketch).. Suppose there is an SCM $\mathcal{M}$ where $l^*(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}) > l(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M})$ for some $(x, y, \mathbf{c}_1, \mathbf{c}_2)$. Construct an alternative SCM $\mathcal{M}^*$ such that (1) $\mathcal{M}^*$ and $\mathcal{M}$ share the same $P(X, Y, \mathbf{C}_1, \mathbf{C}_1)$, $P_X(Y, \mathbf{C}_1)$, $P_X(Y, \mathbf{C}_2)$; and (2) the counterfactual distribution $P(X, Y_x, \mathbf{C}_1, \mathbf{C}_2)$ in $\mathcal{M}^*$ satisfy the following, based on the evaluation of lower bound l in $\mathcal{M}$:

1. $P(y_x \mid \neg x, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}^*) = 0$ if $l(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}) = l_1(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M})$;
2. $P(y_x \mid \neg x, \mathbf{c}_1, \neg \mathbf{c}_2; \mathcal{M}^*) = 1$ if $l(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}) = l_2(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M})$;
3. $P(y_x \mid \neg x, \neg \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}^*) = 1$ if $l(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}) = l_3(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M})$;
4. $P(y_x \mid \neg x, \neg \mathbf{c}_1, \neg \mathbf{c}_2; \mathcal{M}^*) = 0$ if $l(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}) = l_4(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M})$.

This construction is feasible since observational and interventional distributions are under-determined by counterfactual distributions in SCMs (Bareinboim et al. 2020). It is verifiable in this modified SCM $\mathcal{M}^*$, the target effect $P_x(y, \mathbf{c}_1, \mathbf{c}_2)$ matches the lower bound l given by Thm. 1,

$$P_x(y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}^*) = l(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}^*) \tag{20}$$

Since lower bounds l and l^* are functions of distributions $P(X, Y, \mathbf{C}_1, \mathbf{C}_1)$, $P_X(Y, \mathbf{C}_1)$, $P_X(Y, \mathbf{C}_2)$ which are shared across $\mathcal{M}$ and $\mathcal{M}^*$, we must have

$$l^*(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}^*) = l^*(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}), \tag{21}$$

$$l(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}^*) = l(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}) \tag{22}$$

Since l^* consistently dominates l in SCM $\mathcal{M}$, the above equations imply

$$l^*(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}^*) > l(x, y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}^*) \tag{23}$$

$$> P_x(y, \mathbf{c}_1, \mathbf{c}_2; \mathcal{M}^*) \tag{24}$$

This means that l^* is not a valid lower bound for $P_x(y, \mathbf{c}_1, \mathbf{c}_2)$ in $\mathcal{M}^*$, which is a contradiction. □

Safe Policy Learning with Partial Effects

We now apply the closed-form bounds derived so far to solve for a safe policy that outperforms the baseline policy. Without loss of generality, assume the reward $Y \in \{0, 1\}$; let y_1 denote the event $Y = 1$. By replacing the inner minimization in Eq. (5) with the lower bound given in Thm. 1, the worst-case treatment effect of policy π could be written as:

$$\min_{\mathcal{M} \in \mathbb{M}} \mathbb{E}_\pi[Y; \mathcal{M}]$$

$$\geq \sum_{x, y, \mathbf{c}_1, \mathbf{c}_2} \pi(x \mid \mathbf{c}_1, \mathbf{c}_2) y \max_{j=1,\ldots,4} l_j(x, y_1, \mathbf{c}_1, \mathbf{c}_2) \tag{25}$$

The maximin objective proposed in Eq. (3) could thus be written as

$$\arg\max_{\pi \in \Pi} \underbrace{\mathbb{E}_{x \sim \pi}\left[w(x, y_1, \mathbf{c}_1, \mathbf{c}_2)\right]}_{\text{Value function } \mathrm{V}(\pi)} - \underbrace{\mathbb{E}[Y]}_{\text{Baseline performance } \tau} \tag{26}$$

Algorithm 1: Safe Policy Learning

Input: Samples $(y_i, x_i, \mathbf{c}_{1,i}, \mathbf{c}_{2,i})_{i=1}^{N_{\text{obs}}}$, $(y_i, x_i, \mathbf{c}_{1,i})_{i=1}^{N_{\text{exp1}}}$, $(y_i, x_i, \mathbf{c}_{2,i})_{i=1}^{N_{\text{exp2}}}$ and learning rate $\lambda > 0$

1: Estimate lower bounds $l_j(x, y_1, \mathbf{c}_1, \mathbf{c}_2)$, $j = 1, \ldots, 4$, given by Thm. 1 from the observational and experimental data $(y_i, x_i, \mathbf{c}_{1,i}, \mathbf{c}_{2,i})_{i=1}^{N_{\text{obs}}}$, $(y_i, x_i, \mathbf{c}_{1,i})_{i=1}^{N_{\text{exp1}}}$, $(y_i, x_i, \mathbf{c}_{2,i})_{i=1}^{N_{\text{exp2}}}$
2: **for** $x, i \in \Omega_X \times [N_{obs}]$ **do**
3: $\quad w_i(x, y_1, \mathbf{c}_{1,i}, \mathbf{c}_{2,i}) \leftarrow \max_{j=1,\ldots,4} l_j(x, y_1, \mathbf{c}_{1,i}, \mathbf{c}_{2,i})$
4: **end for**
5: Initialize parameters of π_0 randomly.
6: **for** $e \in \{1, 2, \cdots, N_{\text{epochs}}\}$ **do**
7: $\quad \pi_{e+1} \leftarrow \pi_e + \lambda \nabla_\pi \mathrm{V}_{N_{obs}}(\pi)$ such that $\pi_e \in \Pi$
8: **end for**
9: **return** $\pi_{N_{\text{epochs}}+1}$

Among the above quantities, the performance baseline $\mathbb{E}[Y]$ is estimable by computing the empirical mean of reward in the observational data; the weight w is a function defined as

$$w(x, y, \mathbf{c}_1, \mathbf{c}_2) \triangleq \max_{j=1,\ldots,4} l_j(x, y, \mathbf{c}_1, \mathbf{c}_2) \tag{27}$$

Given access to multiple distributions, the worst-case treatment effect is the best estimate of the lower bound leveraging all data sources/multiple distributions. Thus, the minimax problem in Eq. (26) reduces to weighted maximization that corresponds to the best worst-case treatment effects.

We propose a three-step algorithm to learn a safe policy π, in Alg. 1 for the case when the bounds in Thm. 1 can be estimated reliably from data. In Step 1, we use plug-in estimates of the bounds given by Thm. 1 by first deriving the bounds as a function of conditional distributions $P(y|x, \mathbf{c}_1, \mathbf{c}_2), P(x|\mathbf{c}_1, \mathbf{c}_2), P_x(y|\mathbf{c}_1), P_x(y|\mathbf{c}_2), P(x|\mathbf{c}_1)$ and $P(x|\mathbf{c}_2)$ (see (Joshi, Zhang, and Bareinboim 2023, Sec. 11) for a detailed discussion), and then estimating a plug-in bound by estimating these conditional using standard supervised learning methods, such as regression and/or supervised classification, as appropriate. Steps 2 - 4 estimate the worst-case treatment effects for all instances of observed covariates $(\mathbf{c}_{1,i}, \mathbf{c}_{2,i})$ by computing the weight function w_i. Here we only consider the observational data since covariates $\mathbf{C}_1, \mathbf{C}_2$ are observed simultaneously. Steps 5 - 8 optimize for a safe policy, using estimates w_i, effectively resulting in a differentiable weighted loss function (Eq. (26)) that is maximized over a function family Π. We use policy ascent for the maximization. Alg. 1 is effectively an Oracle-based algorithm since we do not consider the statistical challenges of estimating the bounds.

Experiments

We evaluate the proposed method on 1) Synthetic data, and 2) the International Stroke Trial (IST) data (Group et al. 1997; Sandercock, Niewada, and Członkowska 2011) and learn four policies. Each policy uses one or more derived bounds in the maximin framework: i) Alg 1 $(\mathbf{C}_1, \mathbf{C}_2)$ - l_1: Uses only Obs bound, ii) Alg 1 $(\mathbf{C}_1, \mathbf{C}_2)$ - l_1, l_2: Uses Obs

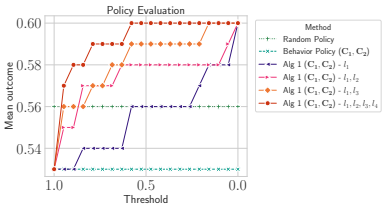

Figure 3: Synthetic Data policy evaluation at varying the threshold on policy scores, 1 (zero treated) $\rightarrow$ 0 (all treated). Higher is better.

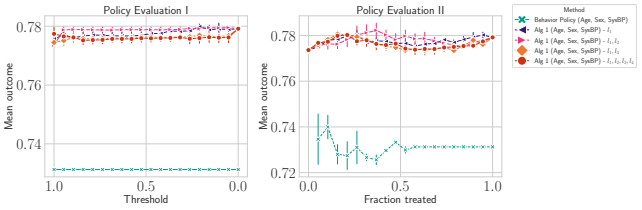

Figure 4: IST Policy Evaluation. Left: Policy evaluation with varying threshold 1 (zero treated) $\rightarrow$ 0 (all treated) on policy scores. Right: Mean outcome when x-fraction of the population is targeted for treatment using sorted policy scores.

and Obs+Exp($\mathbf{C}_1$) bound, iii) Alg 1 ($\mathbf{C}_1, \mathbf{C}_2$) - l_1, l_3: Uses Obs and Obs+Exp($\mathbf{C}_2$) bound, and iv) Alg 1 ($\mathbf{C}_1, \mathbf{C}_2$) - l_1, l_2, l_3, l_4: Uses all bounds. We compare to the following baselines: i) Random policy, ii) Behavior policy ($\mathbf{C}_1, \mathbf{C}_2$): policy used to collect the observational data.

Synthetic Data. The generative process of the data is,

$$
\begin{aligned}
&U \sim \mathcal{N}(0,1); \quad \mathbf{C}_1 = 0.1U + 2.05e-04; \\
&\mathbf{C}_2 = 0.43U + 8.97e-05 \\
&x = \mathbf{1}\left(\sigma(\mathbf{v}_x \cdot [U, \mathbf{C}_1, \mathbf{C}_2]^T + \mathcal{N}(0,1)) > 0.5\right) \\
&y = \mathbf{1}\left(\sigma(\mathbf{v}_y \cdot [U, \mathbf{C}_1, \mathbf{C}_2, x]^T + \mathcal{N}(0,1)) > 0.5\right)
\end{aligned} \tag{28}
$$

where vectors $\mathbf{v}_x = [-0.06, 0.39, 0.46]$ and $\mathbf{v}_y = [-0.16, 0.41, 0.04, 0.1]$. Three data sources are generated, each consisting of $1,000$ samples corresponding to Obs, Exp($\mathbf{C}_1$), and Exp($\mathbf{C}_2$). For experimental data, the treatment is sampled using: $x \sim$ Bernoulli(0.5). To implement Alg. 1, we first estimate all lower bounds using their simplified form derived in the plugin bounds. These require estimating intermediate conditional distributions such as $P(y|x, \mathbf{c}_1, \mathbf{c}_2), P(x|\mathbf{c}_1, \mathbf{c}_2), P(x|\mathbf{c}_1), P(x|\mathbf{c}_2)$ which require marginalization over U and/or $\mathbf{C}_1, \mathbf{C}_2$ for which numerical integration was used. These estimates are then used in Alg. 1 to learn a treatment policy π. The function family Π (see Eq. (26)) corresponds to a two-layer Multi-layer Perceptron (MLP) with 5 hidden units and the GELU activations (Hendrycks and Gimpel 2016).

Each policy returns a score between 0 and 1. All samples above a threshold can be chosen for treatment. We evaluate mean outcome over the data, of varying the threshold between 1 (treat no one) to 0 (treat everyone). Fig. 3 shows the mean outcome for varying thresholds averaged over 5-fold cross-validation (standard errors not visible due to low variability). The learned policy Alg 1 ($\mathbf{C}_1, \mathbf{C}_2$) - l_1, l_2, l_3, l_4 clearly outperforms all baselines suggesting our estimates of treatment effect indeed improve using multiple data-sources and can be leveraged to learn not only safe, but improved policies. Further, the bounds Obs+ Exp($\mathbf{C}_1$), and Obs+ Exp($\mathbf{C}_2$) improve over the natural bounds Obs for some covariate values (see Fig. 5 in (Joshi, Zhang, and Bareinboim 2023, Sec. 12)) providing better policies compared to Obs. Finally, the analysis suggests that bounds using Obs+ Exp($\mathbf{C}_1$)+Exp($\mathbf{C}_2$) are not informative. Nonetheless, using l_1, l_2, l_3 in conjunction provides significant improvement over behavior policy, and other variants.

International Stroke Trial (IST). The goal of this trial was to estimate the effect of Aspirin on the clinical course of Ischemic Stroke. The database consists of patient age, gender, conscious state at randomization, and systolic blood pressure at randomization, among others. We study the outcome at 6 months post-treatment ($Y = 1$ if a patient survives, 0 otherwise). From this trial data, we create an observational dataset by inducing selection bias as a function of age, gender, conscious state at randomization, and systolic blood pressure (see (Joshi, Zhang, and Bareinboim 2023, Sec. 12) for more details). We treat conscious state as unobserved confounding. $\mathbf{C}_1 = \{\text{Age, Sex}\}$ and $\mathbf{C}_2 = \{\text{Systolic Blood Pressure (SysBP)}\}$ are observed confounding attributes. We set aside 30% data as a held-out test set.

Fig. 4 demonstrates the results. On the left, we show the mean outcome for varying thresholds on the policy score (a higher threshold implies fewer patients selected for treatment). Thus higher mean outcome while selecting fewer patients is desirable. The learned policies clearly dominate the behavior policy. On the right, we show the outcome when x-fraction of the population is selected for treatment based on policy scores. Again, while the improvement over behavior policy is significant, improvement using different bounds is limited. Fig. 8 in (Joshi, Zhang, and Bareinboim 2023, Sec. 12) shows the bounds obtained in each case. Notice that the effect of treatment is relatively small and only occurs for a small region of the SysBP space. Second, our bounds require parametric assumptions on $P(\text{Age, Sex, SysBP})$. We model SysBP as a Gaussian, and Age and Sex as independent Bernoulli variables. Misspecification in our parametrization may result in biased estimates in our bounds. Additional covariates may further improve the bounds and in turn the treatment policy.

Conclusions

We propose a safe policy learning framework in non-identifiable settings using observational studies with unobserved confounding, experimental studies with partial observability, and combinations thereof. We derive closed-form bounds over conditional treatment effects. We propose a robust policy improvement framework to train policies that maximize the worst-case treatment effect by using our lower bounds that is guaranteed to improve over a baseline policy that generates the observational data. We demonstrate utility in synthetic and real-world experimental data.

Acknowledgements

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

References

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

Bareinboim, E.; Correa, J.; Ibeling, D.; and Icard, T. 2020. On pearl's hierarchy and the foundations of causal inference. *ACM Special Volume in Honor of Judea Pearl (provisional title)*.

Bareinboim, E.; and Pearl, J. 2012. Causal inference by surrogate experiments: z-identifiability. In de Freitas, N.; and Murphy, K., eds., *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 113–120. Corvallis, OR: AUAI Press.

Bareinboim, E.; and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27): 7345–7352.

Bellman, R. 1957. *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Chakraborty, B.; and Murphy, S. A. 2014. Dynamic treatment regimes. *Annual review of statistics and its application*, 1: 447–464.

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

Cornfield, J.; Haenszel, W.; Hammond, E. C.; Lilienfeld, A. M.; Shimkin, M. B.; and Wynder, E. L. 1959. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1): 173–203.

Fan, Y.; and Park, S. S. 2010. SHARP BOUNDS ON THE DISTRIBUTION OF TREATMENT EFFECTS AND THEIR STATISTICAL INFERENCE. *Econometric Theory*, 26(3): 931–951.

Fisher, R. 1926. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33: 503–513.

Ghavamzadeh, M.; Petrik, M.; and Chow, Y. 2016. Safe policy improvement by minimizing robust baseline regret. *Advances in Neural Information Processing Systems*, 29.

Gresele, L.; Von Kügelgen, J.; Kübler, J.; Kirschbaum, E.; Schölkopf, B.; and Janzing, D. 2022. Causal inference through the structural causal marginal problem. In *International Conference on Machine Learning*, 7793–7824. PMLR.

Group, I. S. T. C.; et al. 1997. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. *The Lancet*, 349(9065): 1569–1581.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Huang, Y.; and Valtorta, M. 2006. 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*, 217–224. Corvallis, OR: AUAI Press.

Joshi, J.; Zhang, J.; and Bareinboim, E. 2023. Towards Safe Policy Learning under Partial Identifiability: A Causal Approach. Technical Report R-96, Causal Artificial Intelligence Lab, Columbia University.

Kallus, N.; Puli, A. M.; and Shalit, U. 2018. Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 31.

Kallus, N.; and Zhou, A. 2018. Confounding-robust policy improvement. *arXiv preprint arXiv:1805.08593*.

Langford, J.; and Zhang, T. 2008. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In Platt, J.; Koller, D.; Singer, Y.; and Roweis, S., eds., *Advances in Neural Information Processing Systems 20*, 817–824. Curran Associates, Inc.

Lee, S.; and Bareinboim, E. 2020. Causal effect identifiability under partial-observability. In *International Conference on Machine Learning*, 5692–5701. PMLR.

Lee, S.; Correa, J.; and Bareinboim, E. 2019. General Identifiability with Arbitrary Surrogate Experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. Tel Aviv, Israel: AUAI Press.

Levine, S.; Kumar, A.; Tucker, G.; and Fu, J. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 661–670. ACM.

Li, L.; Munos, R.; and Szepesvari, C. 2015. Toward Minimax Off-policy Value Estimation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*.

Manski, C. F. 1990. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2): 319–323.

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

Murphy, S. A. 2005. A Generalization Error for Q-Learning. *Journal of machine learning research: JMLR*, 6: 1073–1097.

Namkoong, H.; Keramati, R.; Yadlowsky, S.; and Brunskill, E. 2020. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33: 18819–18831.

Nuru-Jeter, A. M.; Michaels, E. K.; Thomas, M. D.; Reeves, A. N.; Thorpe Jr, R. J.; and LaVeist, T. A. 2018. Relative roles of race versus socioeconomic position in studies of health inequalities: a matter of interpretation. *Annual review of public health*, 39: 169–188.

of Sciences Engineering, N. A.; Medicine; et al. 2016. Metrics that matter for population health action: workshop summary.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688.

Pearl, J. 2009. *Causality*. Cambridge university press.

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

Robins, J. 1989. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In Sechrest, L.; Freeman, H.; and Mulley, A., eds., *Health Service Research Methodology: A Focus on AIDS*, 113–159. Washington, D.C.: NCHSR, U.S. Public Health Service.

Robins, J. M. 1997. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, 69–117. Springer.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.

Sandercock, P. A.; Niewada, M.; and Członkowska, A. 2011. The international stroke trial database. *Trials*, 12(1): 1–7.

Shpitser, I.; and Pearl, J. 2006. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Strehl, A.; Langford, J.; Li, L.; and Kakade, S. M. 2010. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, 2217–2225.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Swaminathan, A.; and Joachims, T. 2015. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, 814–823. PMLR.

Thomas, P.; Theocharous, G.; and Ghavamzadeh, M. 2015. High confidence policy improvement. In *International Conference on Machine Learning*, 2380–2388. PMLR.

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

Tian, J.; and Pearl, J. 2002. A general identification condition for causal effects. In *Aaai/iaai*, 567–573.

Zhang, J.; and Bareinboim, E. 2021. Bounding causal effects on continuous outcome. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, J.; Tian, J.; and Bareinboim, E. 2022. Partial counterfactual identification from observational and experimental data. In *International Conference on Machine Learning*, 26548–26558. PMLR.