CAUSAL INFERENCE METHODS FOR BIAS CORRECTION IN DATA ANALYSES

by
Razieh Nabi

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
March 2021

Abstract

Many problems in the empirical sciences and rational decision making require *causal,* rather than *associative,* reasoning. The field of causal inference is concerned with establishing and quantifying cause-effect relationships to inform interventions, even in the absence of direct experimentation or randomization. With the proliferation of massive datasets, it is crucial that we develop principled approaches to drawing actionable conclusions from imperfect information. Inferring valid causal conclusions is impeded by the fact that data are unstructured and filled with different sources of bias. The types of bias that we consider in this thesis include: confounding bias induced by common causes of observed exposures and outcomes, bias in estimation induced by high dimensional data and curse of dimensionality, discriminatory bias encoded in data that reflect historical patterns of discrimination and inequality, and missing data bias where instantiations of variables are systematically missing.

The focus of this thesis is on the development of novel causal and statistical methodologies to better understand and resolve these pressing challenges. We draw on methodological insights from both machine learning/artificial intelligence and statistical theory. Specifically, we use ideas from graphical modeling to encode our assumptions about the underlying data generating mechanisms in a clear and succinct manner. Further, we use ideas from nonparametric and semiparametric theories to enable the use of flexible machine learning modes in the estimation of causal effects that are identified as functions of observed data.

There are four main contributions to this thesis. First, we bridge the gap between

identification and semiparametric estimation of causal effects that are identified in causal graphical models with unmeasured confounders. Second, we use semiparametric inference theory for marginal structural models to give the first general approach to causal sufficient dimension reduction of a high dimensional treatment. Third, we address conceptual, methodological, and practical gaps in assessing and overcoming disparities in automated decision making using causal inference and constrained optimization. Fourth, we use graphical representations of missing data mechanisms and provide a complete characterization of identification of the underlying joint distribution where some variables are systematically missing and others are unmeasured.

Committee Members

Dr. Ilya Shpitser (Primary Advisor)
John C. Malone Assistant Professor
Department of Computer Science
Whiting School of Engineering
Johns Hopkins University

Dr. Daniel Scharfstein
Professor of Biostatistics
Department of Population Health Sciences
School of Medicine
University of Utah

Dr. Eric Tchetgen Tchetgen
Luddy Family President's Distinguished Professor
Statistics Department
Wharton School of Business
University of Pennsylvania

Dr. Elizabeth Ogburn
Associate Professor
Department of Biostatistics
Bloomberg School of Public Health
Johns Hopkins University

Preface

Prior to working with Ilya, I had almost no exposure to the field of causal inference. My first encounter with "causal reasoning" was a cosmological argument in my pre-college theology courses, which I was not impressed by. Years later when I was doing my masters in statistics, the mantra of "correlation is not causation" got stuck in my head. The summer before applying to PhD programs, I visited my sister, Marzieh, in California and found the Causality book by Judea Pearl in her bookshelf. I started reading parts of it, and came across this quote: "I would rather discover one causal law than be King of Persia" – Democritus. Semi-seriously I thought to myself: maybe if I pursue a degree in causal inference, one day if I am presented with the throne to be the Queen of Persia, I can decline because at that point I might have learned many causal laws! Marzieh's book now sits in my bookshelf.

When I joined the CS program at Hopkins, I started working with Ilya on two separate projects. The first one was on a causal view of algorithmic fairness. There are many stories where AI algorithms demonstrate discriminatory, and potentially harmful, behaviors towards minorities. Initially, I relied on this work as a positive vehicle for addressing the discrimination I felt due to the restrictive immigration policies and the travel ban in the US. Over time, I found more purpose in my research as it seeks to raise awareness and improve the lives of underrepresented minorities. My research on algorithmic fairness (described in Chapter 3) led to the development of a causal framework to interrogate and modify AI algorithms to not rely on sensitive attributes, like race or gender, in inappropriate ways.

The basis for my research on use of semiparametric theory in estimation of causal quantities originally stemmed from my passion for developing a method that establishes the cause-effect relations between the high dimensional treatment of radiation therapy and salivary dysfunctions. This was the second project I was working on in parallel with algorithmic fairness (described in Chapter 2). This launched me into reading the book on Semiparametric Theory and Missing Data by Anastasios Tsiatis. In a few months, we (Ilya's group) joined forces with folks at the Biostats department, and our discussions turned into a regular story time narrated by Dan Scharfstein. For over a year, every Friday we would gather in the library on the 3rd floor of the School of Public Health, and enjoy story time accompanied with coffee and donuts from Dunkin'. Receiving validation from senior researchers in the semiparametrics field like Dan boosted my confidence and I grew to enjoy it even more.

On the other hand, a colleague of mine, Rohit, was not quite as impressed as I was about the theory. His main issue was lack of an automated procedure to derive influence functions and perform projections. Focusing on average causal effects, Rohit and I started thinking about an automated procedure to find influence functions for effects that are identified in causal graphical models with unmeasured confounders (described in part in Chapter 2). Prior to this, Rohit and I worked on two papers on missing data identification (Chapter 4) which stemmed from working on structure learning with missing data and getting stuck at the "wasteland" of non-identifiable laws. We paused the structure learning project and started thinking more carefully about the identifiability aspects of missing data models.

If I am ever given a chance to go back in time and re-do my PhD, I would try harder to stick to the principles beautifully presented in this quote: "The important thing in life is not to conquer but to fight well and not to win but to take part." – Pierre de Coubertin.

To my lovely mom for being a role model of a brave and independent woman

&

To my late hardworking dad for teaching me to be ambitious and responsible

Acknowledgments

Acknowledgments are by far the hardest part of the dissertation to write. Feelings, emotions, and sensations cannot easily be summarized in a few pages. There are many people to thank. There are even people whose names or faces I do not know that have played a crucial role to place me where I am today; like the immigration officers who decided for me that I should not pursue a PhD degree in Aerospace Engineering at UW! Some may think that was unfortunate. I cannot agree *or* disagree with this sentiment as exploring all the counterfactual worlds that I could have ended up in is unattainable. What I can say with certainty however, is that I am happy and thankful to all the causes that brought me to Hopkins.

At Hopkins, I started working with Ilya on multiple causal inference projects. I immediately fell in love with the field and here I am today concluding my PhD work on causal inference. Ilya's passion for research and his support throughout these years has kept me swirling in this world of counterfactuals and I cannot thank him enough for this. Ilya has gathered a wonderful group of scholars in "House-of-Ayli." It has been a pleasure growing up with them as a person and a researcher. I have learned a lot from each and every one of the "fellow-kids." I thank Amir, Dan, Eli, Jaron, Noam, Numair, Ranjani, Rohit, and Zach for all the nice memories, collaborations, and friendships.

I would like to thank Ilya, Dan S, Betsy, Eric, and Emre for writing strong letters of recommendations for me when I was on the job market, and it is a true honor to have Dan S, Betsy, and Eric in my thesis committee. I especially would like to

thank Dan S for his gracious mentorship when I needed it the most during my job search and interviews, for planting the seed of passion for semiparametrics in me, and for narrating the semiparametric story time for over a year to the "merry band." This brought me closer to Bonnie, Lamar, Youjin, and Ryan (who made coffee and donuts a story time tradition.) I learned a lot from them during our discussions and brainstorming sessions over the "clearly" stated claims in the "yellow book." I would also like to thank Dr. Su for his extensive support when I was writing my first paper in grad school. I had the pleasure of having wonderful mentors like the late Dr. Joan Staniswalis, Dr. Ahmet Bulut, and Dr. Tarik Arici.

This whole journey would have been a total wreck if it was not for friends to share with them the ups and downs, the laughter and tears along the way, to celebrate the successes and to get inspired in failures. My heart is filled with memories of my friends in undergrad in Tehran, the Taksim Square and Acibadem in Istanbul, the third and second floor of Malone (especially the time spent procrastinating and making coffee in the kitchen of Malone). Carito and Manirah have made my Baltimore experience extra special. I can spend hours and hours talking with Carito without realizing the passage of time (and yes Carito I'm talking about Holy Frijoles). I had many joyful memories of first-time experiences with Manirah (first lighting of a Christmas tree, first pumpkin carving, first official birthday cake, first roasted turkey during thanksgiving). I am also grateful to have known Shreya as a friend. Her strength, ambition, kindness, and work ethic are all very admirable.

My family is my strength; like branches on a tree we grow in different directions, yet we share the same strong roots. Despite being thousands of miles apart, our love for one another knows no bounds. Thank you Marzieh, Sareh, Aboozar, Soheila and Negin for being examples of truly strong individuals, for your kind hearts, and for your endless love and support. Thank you mom for inspiring me to be the best version of myself, for sitting down in class with me for a while when I started going to school

before you realized it was not a sustainable solution, and teaching me how to be independent in your warm hugs that smell like heaven. Thank you dad for everything you did for me, for believing that your children can reach the stars if they want to, for teaching me math (your favorite topic) which put me ahead of my classmates in school and made me love the subject even more. Your memories keep me moving in hardship.

And Rohit: my colleague, friend, and partner. No matter how busy his schedule looks like, Rohit has always had time for me. He has inspired me to do the best I can to help others, as to him other people always come first. Thanks Rohit for your endless love and support, for getting excited to watch movies with me that you've already watched, for half marathoning, for my first time ice skating, for always volunteering to cut the onions when we cook, for being crazy enough to jump in the car and drive to Miami with me, for your wonderful Spotify playlists, for showing your support on the sticky notes in my lunch boxes when going to Malone for the job interviews, for making me enjoy Tupac and Notorious BIG the way their music deserves, and the time you spare to write papers with me! Rohit also made me fall in love with Baltimore – The Greatest City in America as the broken bench once said.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Many problems in the empirical sciences and rational decision making require *causal,* rather than *associative,* reasoning. For instance, an important task in most studies in the health sciences and public policies is deriving better data-driven treatment decisions and designing optimal interventions. This requires reasoning counterfactually and thinking about the consequences of interventions, e.g., "would patient X have suffered the adverse outcome Y if they had, contrary to fact, been treated with drug A instead of B?" or "what is the expected mortality rate in a regime where every patient is assigned to treatment T?."

Answering causal and counterfactual questions based on data requires a formalism for expressing and evaluating what might be (or might have been) observed in various situations not necessarily represented in the data. This requires certain extensions in the standard mathematical language of statistics. Several (largely equivalent) frameworks have been developed for the theory of causality based on structural equation models, the potential outcomes framework of Neyman, and causal graphical models developed for probabilistic reasoning and causal inference.

With the proliferation of massive datasets, it is crucial that we develop principled approaches to drawing actionable conclusions from imperfect information. Unfortunately, data are commonly unstructured and filled with different sources of bias. This

makes drawing valid causal conclusions challenging. Examples of different types of bias that exist in data include: (i) confounding bias induced by common causes of observed exposures and outcomes, (ii) bias in estimation induced by high dimensional data and curse of dimensionality, (iii) discriminatory bias encoded in data that reflect historical patterns of discrimination and inequality, and (iv) missing data bias where instantiations of variables are systematically missing.

The theme of this thesis is understanding and resolving these complications in data. This entails exploiting tools from statistics, optimization theory, machine learning, and artificial intelligence. Specifically, we use ideas from semiparametric theory to derive estimators for causal effects with desirable statistical properties such as fast rates of convergence and quantification of uncertainty. Further, we use ideas from graphical modeling to encode our assumptions about the underlying data generating mechanisms in a clear and succinct manner. The contributions of this thesis on tackling the four challenges, mentioned above, can be summarized as follows.

It is commonly assumed that all common causes (a.k.a. confounders) between the treatment and outcome are measured. However, in observational data, it is difficult to justify this assumption. In the first section of Chapter 2, we bridge the gap between identification and estimation theories for causal effects in causal graphical models with unmeasured confounders. In particular, we derive *doubly robust* semiparametric estimators for a significant subset of hidden variable causal graphical models. These estimators allow for only partial specification of the data-generating process, and enable the use of flexible ML methods while retaining desirable statistical properties such as $\sqrt{n}$- consistency, asymptotic normality, and in some cases robustness to model misspecification..

In the second section of Chapter 2, we consider scenarios where even if all confouding variables are measured, drawing causal conclusions can still be challenging. In classical causal inference, the treatment variable is often assumed to take on binary values

as in treatment vs. placebo or continuous values as in drug dosages. However, in certain applications, we might encounter a treatment with values that lie in a higher dimensional space. For instance, oncologists are interested in the effect of radiation therapy on salivary dysfunction in head and neck cancer patients. Unlike standard treatments, radiation is represented via 3D voxel maps of exposure dosages on the organs involved in the radiation therapy. In the second section of Chapter 2, we propose a strategy for performing a feasible causal analysis by finding a lower dimensional representation of the treatment in a way that preserves its causal effect on the outcome.

Another challenge in data-driven decision making is counteracting discriminatory biases reflected in data. With the massive expansion of available data and advancements in ML algorithms, increasingly important decisions are being automated. Unfortunately, this increases the potential for discriminatory biases to become "baked in" to automated systems that influence people's lives. Without careful adjustments for these biases during learning and deployment of automated systems, these systems could indeed put certain individuals at risk of discrimination. In Chapter 3, we aim to address conceptual, methodological, and practical gaps in assessing and overcoming disparities in automated decision making by a combination of tools from causal mediation analysis and constrained optimization.

A ubiquitous source of bias in applied data analyses is missing data which results in target distributions that are systematically censored by a missingness process. A common modeling approach assumes data entries are censored in a way that does not depend on the underlying missing data, known as the missing completely at random (MCAR) model, or only depends on observed values in the data, known as the missing at random (MAR) model. These simple models are insufficient however, in problems where missingness status may depend on underlying values that are themselves censored. This type of missingness is known as missing not at random (MNAR). While the underlying target distribution is often not identified from observed

data under MNAR, there exist identified MNAR models. In Chapter 4, we show that the most general currently known methods for identification in graphical models of missing data retain a significant gap, in the sense that they fail to identify the underlying joint distribution in many models where it is indeed identified. Further, we provide a complete characterization of identification of the underlying joint distribution where some variables are systematically missing and others are unmeasured.

In the remainder of this chapter, we provide an overview of causal inference and introduce some preliminaries that are required for the development of methods in the following chapters.

Disclaimer: The results presented in Section 2.1 and Chapter 4 are based on co-first-author papers that the author of this dissertation shares with Rohit Bhattacharya. Some text appearing in these chapters (and related introductory material in Chapter 1 may be similar across our dissertations.

1.1 Causal Inference Workflow

In causal inference, we are interested in consequences of interventions or counterfactual questions. For example, "what would happen to Y if an "upstream" variable T is intervened on and *set* to t?" or "would the value of Y have been different if, contrary to the fact, T had been different?"

Causal targets of interest are often captured via contrasts of random variables of the form $Y(t)$, which denotes the *potential outcome* (a.k.a. *counterfactual*) Y had treatment T been assigned to t, possibly contrary to the fact [1]. The counterfactual variable $Y(t)$ is equivalent to $Y \mid \text{do}(t)$ in Judea Pearl's do-calculus notation [2]. A common causal target of interest is the *average causal effect* (ACE) of treatment T on outcome Y, defined via the following contrast

$$\text{ACE} \coloneqq \mathbb{E}[Y(t) - Y(t')]. \tag{1.1}$$

Alternative quantities of interest are conditional causal effects (a.k.a heterogeneous effect or subgroup effects), direct, indirect, and path-specific effects (capturing decomposition of an effect along different causal pathways), or dynamic treatment regimes (used in precision medicine and longitudinal decision-making processes).

In order to compute causal quantities from data, we must first posit a *causal model*, which encodes a set of conditional independence restrictions on the distribution of potential outcomes and other relevant variables. Given a causal model, the causal workflow starts with determining whether the target of interest is *identified*, i.e., uniquely computable as a function of the observed data distribution, finding efficient ways of *estimating* the target, and performing *sensitivity analysis* on the assumptions made along the way. To illustrate the causal workflow, consider the following example.

Example 1.1. Assume there exist a joint distribution over a set of observed variables Z denoted by $p(Z)$. Assume we have a set of n independent and identically distributed (iid) samples drawn from $p(Z)$. We are interested in the average causal effect of a treatment T on an outcome Y, i.e., ACE defined in (1.1). In order to compute ACE from the observed data, we need to follow four main steps described below.

Step 1: Causal model – A causal model is a set of distributions defined over the counterfactual and factual variables. A popular causal model is known as *conditionally ignorable model*, which encodes three main assumptions:

(1) *Consistency*: The observed outcome Y is equal to the potential outcome $Y(t)$ when the treatment received is t. This is expressed as $Y(T) = Y$, where $Y(T)$ reads as "the random variable Y had treatment T been assigned to whatever value it would have naturally attained,"

(2) *Conditional ignorability*: There exists a set of measured pre-treatment covariates C that renders the treatment T conditionally independent of the potential outcomes given C, i.e., $Y(t) \perp\!\!\!\perp T \mid C,\ \forall t \in \mathfrak{X}_T$, where $(\cdot \perp\!\!\!\perp \cdot \mid \cdot)$ represents

conditional independence and $\mathfrak{X}_V$ denotes the state space of variable V, and

(3) *Positivity*: For each level of the covariates C, the probability of receiving treatment is greater than zero, i.e., $p(T = t \mid C = c) > 0, \forall t \in \mathfrak{X}_T, c \in \mathfrak{X}_C$.

Step 2: Identification – In causal inference, we use assumptions encoded in the causal model to link observed data with counterfactual contrasts of interest. When such a functional exists, we say the causal parameter is *identified* from the observed data under the causal model; otherwise, the parameter is *unidentified*.

Under the assumptions encoded in a conditionally ignorable model, the counterfactual distribution $p(Y(t))$, for any value t of T, is identified via $\sum_C p(Y \mid T = t, C) \times p(C)$. Therefore, the ACE is identified as the following function of the observed data, known as the *adjustment formula*,

$$\text{ACE} = \mathbb{E}\Big[\mathbb{E}\Big[Y \;\Big|\; T = t, C\Big] - \mathbb{E}\Big[Y \;\Big|\; T = t', C\Big]\Big], \tag{1.2}$$

where the outer expectation is taken with respect to $p(C)$ [2, 3].

Step 3: Estimation – After the causal parameter is identified as an observed data functional, the inference problem can be viewed as a pure functional estimation problem. In general, we are interested in deriving estimators with desirable statistical properties, such as estimators that are unbiased and have low variance with fast rates of convergence to normal limiting distributions. There are different approaches to constructing such estimators, e.g., parametric likelihood methods, score matching, inverse weighting methods, and nonparametric/semiparametric estimators.

The identified functional in (1.2) yields a simple "plug-in" estimator. Assume $\mathbb{E}[Y \mid T = t, C]$ is parameterized by a parameter vector η_y and the parametric form of the regression is captured via the function $\mu_t(C; \eta_y)$, i.e., $\mu_t(C; \eta_y) \equiv \mathbb{E}[Y \mid T = t, C; \eta_y]$. The plug-in estimator reduces to

$$\widehat{\text{ACE}}_{\text{plug-in}} = \mathbb{P}_n\Big[\mu_t\Big(C; \widehat{\eta_y}\Big) - \mu_{t'}\Big(C; \widehat{\eta_y}\Big)\Big], \tag{1.3}$$

where $\mathbb{P}_n[.] \coloneqq \frac{1}{n}\sum_{i=1}^{n}(.)$ and $\widehat{\eta_y}$ are the maximum likelihood values of η_y.

Since assuming a correctly specified parametric observed data likelihood, or even a correctly specified outcome regression $\mu_t(C;\eta_y)$ is unrealistic in practice, a variety of other estimators have been developed that place *semiparametric* restrictions on the observed data distribution. One such estimator is known as the *inverse probability weighting* (IPW) which seeks to compensate for a biased treatment assignment by reweighing observed outcomes of units assigned $T = t$ by the inverse of the normalized treatment assignment probability $p(T = t \mid C)$. If this probability has a known parametric form $\pi_t(C;\eta_{tr}) \equiv p(T = t \mid C)$, the IPW estimator takes the form

$$\widehat{\text{ACE}}_{\text{ipw}} = \mathbb{P}_n\Big[\Big\{\frac{\mathbb{I}(T=t)}{\pi_t(C;\widehat{\eta_{tr}})} - \frac{\mathbb{I}(T=t')}{\pi_{t'}(C;\widehat{\eta_{tr}})}\Big\} \times Y\Big], \tag{1.4}$$

where $\mathbb{I}(.)$ is the indicator function and $\widehat{\eta_{tr}}$ are the maximum likelihood estimates of η_{tr}.

The plug-in and IPW estimators of ACE are both $\sqrt{n}$*-consistent and asymptotically normal* if the models they rely on, $\mu_t(C;\eta_y)$ and $\pi_t(C;\eta_{tr})$ respectively, are parametric and correctly specified. Otherwise, these estimators are no longer consistent. If flexible models are used for $\mu_t(C)$ and $\pi_t(C)$ instead, the resulting estimators may remain consistent, but with unacceptably slow rates; see [4] for examples.

A principled alternative is to consider semiparametric influence function-based estimators that converge to normal limiting distributions at desirable rates and come equipped with reliable estimates of uncertainty. The counterfactual mean $\mathbb{E}[Y(t)]$ in the ACE, which is identified via the adjustment formula, can be viewed as a target parameter in a semiparametric model, yielding the following *influence function*, $\frac{\mathbb{I}(T=t)}{p(T=t|C)} \times \Big\{Y - \mathbb{E}[Y \mid T, C]\Big\} + \mathbb{E}[Y \mid T = t, C] - \mathbb{E}[Y(t)]$. This immediately yields the following *augmented IPW* (AIPW) estimator for the ACE,

$$\begin{aligned}\widehat{\text{ACE}}_{\text{aipw}} = \mathbb{P}_n\Big[&\frac{\mathbb{I}(T=t)}{\pi_t(C;\widehat{\eta_{tr}})} \times \Big\{Y - \mu_t(C;\widehat{\eta_y})\Big\} + \mu_t(C;\widehat{\eta_y}) \\ &- \frac{\mathbb{I}(T=t')}{\pi_{t'}(C;\widehat{\eta_{tr}})} \times \Big\{Y - \mu_{t'}(C;\widehat{\eta_y})\Big\} - \mu_{t'}(C;\widehat{\eta_y})\Big].\end{aligned} \tag{1.5}$$

This estimator exhibits the property of *double robustness* which means that AIPW is a consistent estimator for ACE when either of the two models, i.e., $\pi_t(C;\eta_{tr})$ and $\mu_t(C;\eta_y)$, is correctly specified, even if the other model is arbitrarily misspecified. In a semiparametric model, given by restrictions implied by a graphical model, the influence function that yields the AIPW estimator, can be projected onto the tangent space of the model to improve efficiency, see [5] for details and Section 1.4 for an overview on semiparametric theory.

Step 4: Sensitivity analysis – Establishing cause-effect relationships from observational data often relies on untestable assumptions such as the conditional ignorability assumption. It is crucial to know whether, and to what extent, the conclusions drawn from non-experimental studies are robust to potential unmeasured confounding.

There exists a rich literature on sensitivity analysis that looks at ACE as the causal parameter of interest. The literature can be divided into two main approaches: one seeks *set identification* and the other seeks *point identification* of ACE (at each sensitivity parameter value.) In set identification, the ACE is restricted to an interval informed by the observed distribution [6, 7, 8, 9, 10]. In point identification, a number of authors have proposed positing sensitivity analysis parameters to govern the relationship among unmeasured confounder(s), outcome, and treatment [11, 12, 13, 14, 15, 16]. Recent reviews on this topic include [17, 18].

If a causal parameter is not identified in a given causal model, it means that the causal model does not encode the sufficient assumptions to relate the counterfactuals to the factuals. In such scenarios, existing work either derive informative bounds, or iterate between the first and the second steps in the causal workflow to restrict the causal model enough, e.g., by adding extra assumptions, so that the causal parameter is identified as a function of observed data. For instance, there is a huge body of work on identification with instrumental variables which require assumptions such as exclusion restriction (absence of direct effect) assumptions, as well as additional

parametric or semiparametric assumptions [19, 20, 21].

The set of presumed independencies among variables can grow quite rapidly in high dimensional settings. A convenient visual representation to communicate the underlying statistical assumptions in the causal model, e.g., independencies and details on (un)measured confounding mechanisms, is provided though causal graphical models. Much of our focus in this dissertation lies in developing causal methods using the language of graphical models, as they have proven useful in deriving novel results in complex multivariate causal systems. In the next two sections, we provide an overview of causal graphical models.

1.2 Causal Directed Acyclic Graphs

Causal models are often represented graphically through a form of graphical models known as directed acyclic graphs (DAGs). We use capital letters V to denote sets of random variables as well as corresponding vertices in graphs and lowercase letters v to denote values or assignments to those random variables. As before, the state space of variable V_i is denoted by $\mathfrak{X}_{V_i}$. A DAG $\mathcal{G}(V)$ consists of a set of vertices V connected by directed edges $V_i \to V_j$ (for some $\{V_i, V_j\} \subseteq V$) such that there are no directed cycles. The set $\text{pa}_{\mathcal{G}}(V_i) \equiv \{V_j \in V \mid V_j \to V_i\}$ denotes the parents of V_i in DAG $\mathcal{G}(V)$. When the vertex set is clear from the given context, we often abbreviate $\mathcal{G}(V)$ as simply $\mathcal{G}$.

The *statistical models* of a DAG $\mathcal{G}(V)$ are sets of distributions that factorize as,

$$p(V) = \prod_{V_i \in V} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)). \tag{1.6}$$

Each missing edge between pairs of variables in a DAG $\mathcal{G}$ imply conditional independences in $p(V)$. These can be read off directly from $\mathcal{G}$ via the well-known d-separation criterion [2]. That is, for disjoint sets $X, Y,$ and Z, the following *global Markov property* holds: $(X \perp\!\!\!\perp_{\text{d-sep}} Y \mid Z)_{\mathcal{G}} \implies (X \perp\!\!\!\perp Y \mid Z)_{p(V)}$. When the context is clear, we

simply use $X \perp\!\!\!\perp Y \mid Z$ to denote conditional independence between X and Y given Z.

The *causal models* of a DAG $\mathcal{G}(V)$ are defined over counterfactual random variables $V_i(\text{pa}_i)$ for each $V_i \in V$, where pa_i is a set of values for $\text{pa}_{\mathcal{G}}(V_i)$. These counterfactuals can alternatively be viewed as being determined by a system of *structural equations* $f_i(\text{pa}_i, \epsilon_i)$ that map values pa_i, as well as values of an exogenous noise term ϵ_i, to values of V_i [2]. Other counterfactuals may be defined via recursive substitution. Specifically for any set $A \subseteq V$, and a variable V_i, we have:

$$V_i(a) \equiv V_i\Big(a \cap \text{pa}_{\mathcal{G}}(V_i), \{V_j(a) : V_j \in \text{pa}_{\mathcal{G}}(V_i) \setminus A\}\Big), \tag{1.7}$$

where $\{V_j(a) : V_j \in \text{pa}_{\mathcal{G}}(V_i) \setminus A\}$ is taken to mean the (recursively defined) set of counterfactuals associated with variables in $\text{pa}_{\mathcal{G}}(V_i) \setminus A$, had A been set to a.

Consider the joint distribution over the potential outcome variables in $V \setminus A$, where each potential outcome is recursively defined via (1.7). Denote this joint distribution by $p(\{V \setminus A\}(a))$, or $p(V(a))$ for short. In the functional model of a DAG $\mathcal{G}$ (as well as some weaker causal models), $p(V(a))$ is identified via the *g-formula* functional [3] as follows,

$$p(V(a)) = \prod_{V_i \in V \setminus A} p\Big(V_i \mid a \cap \text{pa}_{\mathcal{G}}(V_i), \text{pa}_{\mathcal{G}}(V_i) \setminus A\Big). \tag{1.8}$$

When A is the empty set, we obtain the familiar DAG factorization for $\mathcal{G}$ given in (1.6). This implies that the causal model of a DAG $\mathcal{G}$ implies the statistical model of the DAG.

Example 1.2. Consider the DAG in Fig. 1-1(a). By the recursive substitution in (1.7), $Y(t)$ is defined to be $Y(t, M(t, C), C)$. By the g-formula in (1.8), the marginal distribution of $p(Y(t))$ is identified as

$$\begin{aligned} p(Y(t)) &= \sum_{C,M} p(Y \mid T = t, M, C) \times p(M \mid T = t, C) \times p(C) \\ &= \sum_{C} p(Y \mid T = t, C) \times p(C) \end{aligned}$$

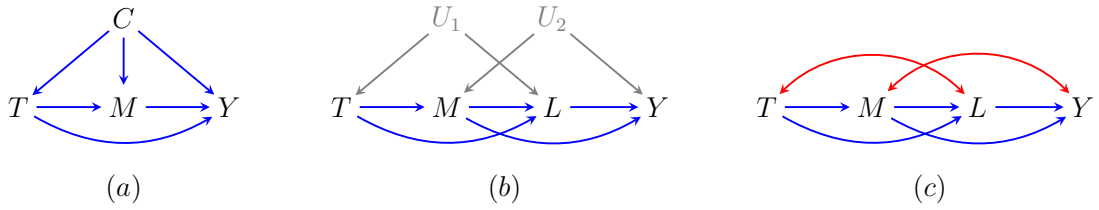


Figure 1-1: (a) A simple causal DAG with treatment T, outcome Y, baseline variables C, and a mediator M. (b) A causal graph with two mediators M and L and unmeasured confounders captured in U. (c) Latent projection of the DAG in (b).

Example 1.3. Using the g-formula in (1.8), it can be easily shown that in all causal models of a DAG $\mathcal{G}$, the ACE is identified via the following simple functional

$$\text{ACE} = \mathbb{E}\Big[\mathbb{E}\Big[Y \mid T = t, \text{pa}_{\mathcal{G}}(T)\Big] - \mathbb{E}\Big[Y \;\Big|\; T = t', \text{pa}_{\mathcal{G}}(T)\Big]\Big]. \tag{1.9}$$

This is also known as the *back-door adjustment* formula [2]. As mentioned earlier, once the target parameter is identified, causal inference reduces to an estimation problem of the identifying functional. There is a number of estimators proposed for this functional, such as the plug-in (1.3), IPW (1.4), and AIPW (1.5) [22, 23, 24, 25].

1.3 Causal DAGs with Hidden Variables

Causal models most relevant to practical applications are sure to contain variables that are unmeasured or hidden to the data analyst. In such cases, the observed data distribution $p(V)$ may be considered to be a margin of a distribution $p(V \cup H)$ associated with a DAG $\mathcal{G}(V \cup H)$ where vertices in V correspond to observed variables and vertices in H correspond to unmeasured or hidden variables. Two complications arise from the presence of hidden variables. First, the target parameter $\psi(t)$ may not always be identified as a function of the observed data and second, parameterizations of latent variable models are generally not fully identifiable and may contain singularities [26].

A natural alternative to the latent variable model is one that places no restrictions

on $p(V)$ aside from those implied by the Markov restrictions given by the factorization of $p(V \cup H)$ with respect to $\mathcal{G}(V \cup H)$. It was shown in [27] that all *equality constraints* implied by such a factorization are captured by a nested factorization of $p(V)$ with respect to an *acyclic directed mixed graph (ADMG)* $\mathcal{G}(V)$ derived from $\mathcal{G}(V \cup H)$ via the *latent projection* operation described by [28]. Such an ADMG is a smooth supermodel of infinitely many hidden variable DAGs that share the same identification theory for $\psi(t)$ and imply the same equality constraints on the margin $p(V)$ [29, 30]. Thus, our use of ADMGs for identification and estimation of the target $\psi(t)$ is without loss of generality.

The latent projection of a hidden variable DAG $\mathcal{G}(V \cup H)$ onto observed variables V is an ADMG $\mathcal{G}(V)$ with directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges constructed as follows. The edge $V_i \rightarrow V_j$ exists in $\mathcal{G}(V)$ if there exists a directed path from V_i to V_j in $\mathcal{G}(V \cup H)$ with all intermediate vertices in H. An edge $V_i \leftrightarrow V_j$ exists in $\mathcal{G}(V)$ if there exists a collider-free path (i.e., there are no consecutive edges of the form $\rightarrow \circ \leftarrow$) from V_i to V_j in $\mathcal{G}(V \cup H)$ with all intermediate vertices in H, such that the first edge on the path is an incoming edge into V_i and the final edge is an incoming edge into V_j. Conditional independences in $p(V)$ can then be read off from the ADMG $\mathcal{G}(V)$ by a simple analogue of the d-separation criterion, known as *m-separation*, that generalizes the notion of a collider to include mixed edges of the form $\rightarrow \circ \leftrightarrow$, $\leftrightarrow \circ \leftarrow$, and $\leftrightarrow \circ \leftrightarrow$, [31]. An example of the latent projection is provided in Fig. 1-1(b-c).

Factorization of ADMGs

We define the factorization of $p(V)$ relative to an ADMG $\mathcal{G}(V)$ with the use of conditional distributions known as *kernels*. A kernel $q_V(V \mid W)$ is a mapping from values of W to normalized densities over V. That is, $\sum_V q_V(V \mid W = w) = 1, \forall w \in W$ [32]. For any set of variables $X \subseteq V$, marginalization and conditioning in a kernel are defined as $q_{V \setminus X}(V \setminus X \mid W) \equiv \sum_X q_V(V \mid W)$ and $q_V(V \setminus X \mid X, W) \equiv \frac{q_V(V|W)}{q_V(X|W)}$

Further, the bidirected connected components of an ADMG $\mathcal{G}(V)$ are essential in the factorization of $p(V)$ relative to the ADMG $\mathcal{G}(V)$.

The bidirected connected components partition its vertices into distinct subsets known as *districts*. A set $S \subseteq V$ is a district in $\mathcal{G}(V)$ if it forms a maximal connected component via only bidirected edges. We use $\text{dis}_{\mathcal{G}}(V_i)$ to denote the district of V_i in $\mathcal{G}$, which includes V_i itself, and $\mathcal{D}(\mathcal{G})$ to denote the set of all districts in $\mathcal{G}$. A distribution $p(V)$ is said to *district factorize* with respect to an ADMG $\mathcal{G}(V)$ if

$$p(V) = \prod_{D \in \mathcal{D}(\mathcal{G})} q_D(D \mid \text{pa}_{\mathcal{G}}(D)), \tag{1.10}$$

where the parents of a set of vertices D is defined as the set of parents of D not already in D, i.e., $\text{pa}_{\mathcal{G}}(D) \equiv \bigcup_{D_i \in D} \text{pa}_{\mathcal{G}}(D_i) \setminus D$. We follow the same convention for children of a set S, denoted $\text{ch}_{\mathcal{G}}(S)$. For other standard genealogical relations defined for a single vertex V_i, such as ancestors $\text{an}_{\mathcal{G}}(V_i) \equiv \{V_j \in V \mid \exists\, V_j \to \cdots \to V_i \text{ in } \mathcal{G}\}$ and descendants $\text{de}_{\mathcal{G}}(V_i) \equiv \{V_j \in V \mid \exists\, V_i \to \cdots \to V_j \text{ in } \mathcal{G}\}$, both of which include V_i itself by convention, the extension to a set S uses the disjunctive definition which also includes the set itself. For example, $\text{an}_{\mathcal{G}}(S) = \bigcup_{S_i \in S} \text{an}_{\mathcal{G}}(S_i)$.

The use of q in lieu of p in Eq. 1.10 emphasizes the fact that these factors are not necessarily ordinary conditional distributions. Each factor $q_D(D \mid \text{pa}_{\mathcal{G}}(D))$ may in fact be treated as a post-intervention distribution where all variables outside of D are intervened on and held fixed to some constant value [33]. Thus, we use $q_S(\cdot \mid \cdot)$ to denote probability distributions where only variables in S are random and all others are fixed. Such densities are often referred to as *kernels* and are similar to conditional densities in the sense that they provide a mapping from values of elements past the conditioning bar to normalized densities over variables prior to the conditioning bar [32]. Conditioning and marginalization in kernels are defined in the usual way.

In [33], it has been shown that each kernel $q_D(D \mid \text{pa}_{\mathcal{G}}(D))$ in Eq. 1.10 is a function of $p(V)$ as follows. Define the *Markov blanket* of a vertex V_i as the district of V_i and the

parents of its district, excluding V_i itself, i.e., $\mathrm{mb}_{\mathcal{G}}(V_i) = \mathrm{dis}_{\mathcal{G}}(V_i) \cup \mathrm{pa}_{\mathcal{G}}(\mathrm{dis}_{\mathcal{G}}(V_i)) \setminus V_i$. Consider a valid topological order τ on all k vertices in V, that is a sequence $(V_1, \ldots, V_k)$ such that no vertex appearing later in the sequence is an ancestor of vertices earlier in the sequence. Let $\{\preceq_\tau V_i\}$ denote the set of vertices that precede V_i in this sequence, including V_i itself. Then for each $D \in \mathcal{D}(\mathcal{G})$,

$$q_D(D \mid \mathrm{pa}_{\mathcal{G}}(D)) = \prod_{D_i \in D} p(D_i \mid \mathrm{mp}_{\mathcal{G}}(D_i)), \tag{1.11}$$

where $\mathrm{mp}_{\mathcal{G}}(V_i)$, the *Markov pillow* of V_i, is defined as its Markov blanket in a subgraph restricted to V_i and its predecessors according to the topological ordering. More formally, $\mathrm{mp}_{\mathcal{G}}(V_i) \equiv \mathrm{mb}_{\mathcal{G}_S}(V_i)$ where $S = \{\preceq_\tau V_i\}$, and $\mathcal{G}_S$ is the subgraph of $\mathcal{G}$ that is restricted to vertices in S and the edges between these vertices. This leads to a factorization of the observed law as a product of simple conditional factors according to the topological order,

$$p(V) = \prod_{V_i \in V} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)). \qquad \textit{(Topological ADMG factorization)} \tag{1.12}$$

The above factorization (and the district factorization in (1.10)) does not always capture every equality restriction in $p(V)$ implied by the Markov property of the underlying hidden variable DAG $\mathcal{G}(V \cup H)$. However, it is particularly simple to work with, and under some conditions is capable of capturing all such restrictions [34]. It is shown that the *nested Markov factorization* of an ADMG captures all equality constraints on the observed margin $p(V)$ [29]. A description of this factorization is provided in Appendix I.

The ease of conveying statistical assumptions visually, via a DAG [2, 35], prompted further study of the identifiability of counterfactual quantities in causal models that factorize according to a DAG, when some variables may be hidden or unobserved [33]. This led to the development of a *sound* and *complete* characterization of the identifiability of the ACE for a given treatment on a given outcome in all hidden variable causal models associated with a DAG, or simply an ADMG [36, 37, 29]. A

complete identification algorithm provides *necessary* and *sufficient* graphical condition under which the causal parameter is identified as a function of the observed data distribution. For any given field of study, such a characterization is one of the most powerful results that identification theory can offer, as it comes with the guarantee that if these conditions do not hold, the parameter is *provably* not identified in the model.

Despite the sophistication of causal identification theory, estimators based on simple covariate adjustment remain the most common strategy for evaluating the ACE from data. Estimates obtained in this way are often biased due to the presence of unmeasured confounding and/or model misspecification. A popular approach for addressing the latter issue has been to use semiparametric estimators developed using the theory of *influence functions* [38, 39, 40, 41]. To the best of our knowledge, the *front-door* model [42] is the only graphical model with unmeasured confounders (where no valid adjustment set exists but the effect is still identifiable in the corresponding causal model) for which an influence function based estimator has been derived [43]. Other related work includes numerical procedures for approximating the influence function proposed by [44, 45]. However, such methods are either restricted to settings where simple covariate adjustment is valid, or involve numerical approximations of the function itself which may be computationally prohibitive.

1.4 Semiparametric Inference

Assume a statistical model $\mathcal{M} = \{p(Z;\eta) : \eta \in \Gamma\}$ where Γ is the parameter space and η is the parameter indexing a specific distribution. We are often interested in a function $\psi : \eta \in \Gamma \mapsto \psi(\eta) \in \mathbb{R}$; i.e., a parameter that maps the distribution P_η to a scalar number in $\mathbb{R}$, such as an identified average causal effect. (For brevity, we sometimes use ψ instead of $\psi(\eta)$, which should be obvious from context.) The true observed data distributions and true parameters are denoted by P_0 and ψ_0,

respectively.

An estimator $\widehat{\psi}_n$ of a scalar parameter ψ based on n i.i.d copies $Z_1, \ldots, Z_n$ drawn from $p(Z;\eta)$, is *asymptotically linear* if there exists a measurable random function $U_\psi(Z)$ with mean zero and finite variance such that

$$\sqrt{n} \times (\widehat{\psi}_n - \psi) = \frac{1}{\sqrt{n}} \times \sum_{i=1}^{n} U_\psi(Z_i) + o_p(1),$$

where $o_p(1)$ is a term that converges in probability to zero as n goes to infinity. The random variable $U_\psi(Z)$ is called the *influence function* (IF) of the estimator $\widehat{\psi}_n$. The analysis is oftentimes restricted to *regular* and asymptotically linear (RAL) estimators to avoid certain complications, such as super efficiency in Hodges estimator [41]. The RAL estimator $\widehat{\psi}_n$ is *consistent and asymptotically normal* (CAN), with asymptotic variance equal to the variance of its influence function U_ψ,

$$\sqrt{n} \times (\widehat{\psi}_n - \psi) \xrightarrow{d} N\Big(0,\ \mathrm{var}(U_\psi)\Big).$$

Influence functions in semiparametric models are derived as normalized elements of the orthogonal complement of the tangent space of the model. First, define the *Hilbert space*, denoted by H, as the space of all mean-zero scalar functions, equipped with the inner product $\mathbb{E}[h_1 \times h_2], \forall h_1, h_2 \in \mathrm{H}$ [38, 39, 41]. The *tangent space* of the statistical model $\mathcal{M}$ is defined as the mean-square closure of all the linear combinations of the score functions in the corresponding parametric submodels for $\mathcal{M}$. We denote the tangent space by Λ. The orthogonal complement of the tangent space, denoted by $\Lambda^\perp$, is then defined as $\Lambda^\perp = \{h \in \mathrm{H} \mid \mathbb{E}[h \times h'] = 0, \forall h' \in \Lambda\}$. Note that $\mathrm{H} = \Lambda \oplus \Lambda^\perp$, where $\oplus$ denotes the direct sum, and $\Lambda \cap \Lambda^\perp = \{0\}$.

The vector space $\Lambda^\perp$ is of particular importance because we can construct the class of all influence functions, denoted by $\mathcal{U}$, as $\mathcal{U} = \{U_\psi + \Lambda^\perp\}$. In other words, upon knowing a single influence function U_ψ and $\Lambda^\perp$, we can obtain the class of all possible RAL estimators that admit the CAN property. Out of all IFs in $\mathcal{U}$, there

exists a unique one which lies in the tangent space Λ and yields the most efficient RAL estimator by recovering the *semiparametric efficiency bound*. This efficient influence function can be obtained by projecting any influence function, call it U_ψ^*, onto the tangent space Λ. This operation is denoted by $U_\psi^{\text{eff}} = \pi[U_\psi^* \mid \Lambda]$, where U_ψ^{eff} denotes the efficient influence function. In a nonparametric saturated model (one with an unrestricted tangent space), the IF is unique; hence the corresponding estimator is the one that achieves the semiparametric efficiency bound. For a more detailed description of the concepts outlined here, see Appendix II and [39, 41].

In a semiparametric model of a DAG $\mathcal{G}(V)$, which is defined by conditional independence restrictions on the tangent space implied by the DAG factorization, the tangent space Λ can be partitioned into a direct sum of orthogonal subspaces as $\Lambda \equiv \oplus_{V_i \in V}\ \Lambda_i$, where $\Lambda_i \equiv \Big\{\alpha_i(V_i, \text{pa}_{\mathcal{G}}(V_i)) \in \text{H} \mid \mathbb{E}[\alpha_i \mid \text{pa}_{\mathcal{G}}(V_i)] = 0\Big\}$.

If $\mathcal{G}$ is a complete DAG, i.e., every vertex is connected to every other vertex, then there exist no independence relations between any sets of variables. In such scenarios, the tangent space equals the entire Hilbert space. In general, any statistical model with tangent space Λ, where $\Lambda = \mathbb{H}$, is said to be *nonparametric saturated* (NPS).

Chapter 2

Identification and Estimation in Causal Inference

In classical causal inference, the treatment variable is often assumed to take on binary values, where $T = 1$ corresponds to receiving the treatment itself and $T = 0$ corresponds to receiving a placebo. In some applications, treatments may take on continuous values in $\mathbb{R}$. For example, we might be interested in evaluating the effect of a particular treatment dose on viral load. In such cases, in addition to contrasts of responses to two specific doses, we may be interested in the entire dose-response relationship, and choose to model it via a simple functional, for example a logarithmic or sigmoidal function.

In other applications, we might be interested in assessing causal relationships between outcomes and treatments with values that lie in a high dimensional space $\mathbb{R}^p$. These types of causal relationships arise in many applications. In natural language processing interest lies in causal analyses that involve high dimensional text data [46]. Moreover, neuroimaging data are increasingly used to relate neuronal network activity to cognitive processing and behavior. Functional magnetic resonance imaging (fMRI) scans are widely used in psychological science, cognitive science, and neuroscience to inform cognitive theories [47, 48]. Other applications include analysis of heterogeneous data in social networks or healthcare that includes images, time series, and other high

dimensional data sources.

We divide this chapter into two sections. In Section 2.1, we focus on how the presence of unmeasured confounders may complicate the identification and estimation of the causal effect of a single treatment T on a single outcome Y. We consider a class of DAGs with hidden variables (or ADMGs) where there does not exist a valid adjustment set to block all the back-door paths between the treatment and the outcome, as in (1.2). However the effect can still be identified as a function of observed data. In Section 2.2, we focus on a specific situation where presence of confounders (even if they are all measured) can complicate our assessment of causal effects. We restrict our attention to a conditionally ignorable model where all common confounders between the treatment and the outcome of interest are measured and the effect is identified via the adjustment functional. In this section, the treatment of interest is a high dimensional treatment ($T \in \mathbb{R}^p, p > 1$).

Throughout this chapter, we set our target of inference to be the mean of the counterfactual random variable $Y(t)$. That is,

$$\psi(t) \equiv \mathbb{E}[Y(t)]. \qquad \textit{(target parameter)} \tag{2.1}$$

2.1 Single Treatment with Hidden Variables

If a causal model contains unmeasured confounders, causal inference becomes considerably more complicated. The last decade witnessed the development of algorithms that completely solve the identifiability problem for causal effects in hidden variable causal models associated with DAGs [36, 37]. However, much of this machinery remains underutilized in practice owing to the complexity of estimating identifying functionals yielded by these algorithms.

In this section, we provide a simple graphical criterion and semiparametric estimators that bridge the gap between identification and estimation of causal effect in a

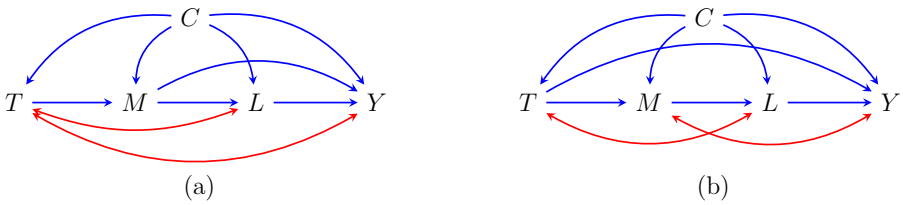


Figure 2-1: Examples of acyclic directed mixed graphs where T is primal fixable.

large class of DAGs with hidden variables where the causal effect is identified, however no valid covariate adjustment is available due to presence of unmeasured confounders.

Consider the ADMGs shown in Fig. 2-1. It is easy to check that in either case there exists no valid adjustment set to identify the causal effect of T on Y. However, such an effect is indeed identified in both graphs. The defining characteristic of these ADMGs that permits identification of the target $\psi(t)$, is that the district of T does not intersect with its children, i.e., variables that have T as their parents on the graph $\mathcal{G}$ and denoted by $\mathrm{ch}_{\mathcal{G}}(T)$.

In this section, we consider ADMGs where $\mathrm{dis}_{\mathcal{G}}(T) \cap \mathrm{ch}_{\mathcal{G}}(T) = \emptyset$. This criterion encompasses many popular models in the literature, including those that satisfy the back-door and front-door criteria [2], as special cases. We name this criterion primal fixability or *p-fixability* for short (due to its generalization of the fixing criterion introduced in the definition of the nested Markov model.)

Primal fixability is known to be a necessary and sufficient condition for the identifiability of the causal effect of T on all other variables $V \setminus T$ [33]. In observed data distributions $p(V)$ that district factorize according to an ADMG $\mathcal{G}(V)$ where T is primal fixable, the resulting identifying functional for the target is as follows.

$$\psi(t) = \sum_{V \setminus T} Y \times \prod_{V_i \in V \setminus D_T} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \times \sum_{T} \prod_{V_j \in D_T} p(V_j \mid \mathrm{mp}_{\mathcal{G}}(V_j)) \Big|_{T=t}, \qquad (2.2)$$

where D_T denotes the district of T [33]. We provide this special notation for the district of T as D_T due to its frequent occurrence in subsequent results.

Assume $p(V)$ factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable and for simplicity of exposition, assume that Y has no descendants in $\mathcal{G}$. The latter assumption is not necessary and our results extend trivially to the setting where this is not true; we use it only to avoid notational complexity [34]. We use a fixed topological order τ where T is preceded by all its non-descendants and Y is succeeded by all its non-descendants non-ancestors. The set of nodes V can then be partitioned into three disjoint sets: $V = \{\mathbb{C}, \mathbb{L}, \mathbb{M}\}$, where

$$\begin{aligned}
\mathbb{C} &= \{C_i \in V \mid C_i \prec T\}, \\
\mathbb{L} &= \{L_i \in V \mid L_i \in D_T, L_i \succeq T\}, \\
\mathbb{M} &= \{M_i \in V \mid M_i \notin \mathbb{C} \cup \mathbb{L}\}.
\end{aligned} \tag{2.3}$$

Rearranging some of the terms in Eq. 2.2, $\psi(t)$ is identified as the following function of the observed data in terms of the sets defined above.

$$\psi(t) = \sum_{V \setminus T} Y \times \prod_{M_i \in \mathbb{M}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\Big|_{T=t} \times \sum_{T} \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times p(\mathbb{C}). \tag{2.4}$$

We derive the corresponding influence function in the following theorem using the *pathwise derivative*; see Appendix II for details. For readability, we use the form $\prod_{L_i \prec M_i}$ as shorthand for $\prod_{L_i \in \mathbb{L} | L_i \prec M_i}$.

Theorem 1 (Nonparametric influence function of augmented primal IPW)**.**
Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, the nonparametric influence function for the target parameter $\psi(t)$, denoted by U_{ψ_t}, is as follows.

$$
\begin{aligned}
U_{\psi_t} = \sum_{M_i \in \mathbb{M}} \Bigg\{ & \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_i} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \Bigg(\sum_{T \cup \{\succ M_i\}} Y \times \prod_{\substack{V_i \in \mathbb{L} \cup \\ \{\succ M_i\}}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))|_{T=t \text{ if } V_i \in \mathbb{M}} \\
& - \sum_{T \cup \{\succeq M_i\}} Y \times \prod_{\substack{V_i \in \mathbb{L} \cup \\ \{\succeq M_i\}}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \; |_{T=t \text{ if } V_i \in \mathbb{M}} \Bigg) \Bigg\} \\
+ \sum_{L_i \in \mathbb{L} \setminus T} \Bigg\{ & \frac{\prod_{M_i \prec L_i} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_i} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times \Bigg(\sum_{\{\succ L_i\}} Y \times \prod_{V_i \succ L_i} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))|_{T=t \text{ if } V_i \in \mathbb{M}} \\
& - \sum_{\{\succeq L_i\}} Y \times \prod_{V_i \succeq L_i} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \; |_{T=t \text{ if } V_i \in \mathbb{M}} \Bigg) \Bigg\} \\
+ \sum_{V \setminus \{T, \mathbb{C}\}} & Y \times \prod_{M_i \in \mathbb{M}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \prod_{L_i \in \mathbb{L} \setminus T} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) - \psi(t), \qquad (2.5)
\end{aligned}
$$

where $\mathbb{C}, \mathbb{L}, \mathbb{M}$ *are defined in display (2.3).*

In the following lemma, we show that the influence function U_{ψ_t} in Theorem 1 uses information in the models for $M_i \in \mathbb{M}$ and $L_i \in \mathbb{L}$ in order to yield an estimator that is doubly robust in these sets.

Lemma 1 (Double robustness of augmented primal IPW)**.**
The estimator obtained by solving the estimating equation $\mathbb{E}[U_{\psi_t}] = 0$*, where* U_{ψ_t} *is given in Theorem 1, is unbiased if all models in either* $\{p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)),\ \forall M_i \in \mathbb{M}\}$ *or* $\{p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)),\ \forall L_i \in \mathbb{L}\}$ *are correctly specified.*

According to Lemma 1, the estimator derived from the nonparametric IF is a *doubly robust* estimator. This allows us to perform consistent inferences for the target parameter $\psi(t)$ even in settings where a large part of the model likelihood is arbitrarily misspecified, provided that conditional models for variables in either $\mathbb{M}$ or $\mathbb{L}$ are specified correctly. In addition, the bias of the estimator has a product form which allows parametric $(\sqrt{n})$ convergence rates for $\psi(t)$ to be obtained even if flexible machine learning models with slower than parametric convergence rates are used to fit nuisance models. See [4] for details.

We can obtain two different estimators for the identified functional in (2.2) with terms that appear in the influence function provided in Theorem 1. This helps us in viewing the influence function in (2.5) as augmenting an IPW-type estimator, in the same way the AIPW estimator for the adjustment functional in (1.5) can be viewed as augmenting the IPW estimator in (1.4) with the outcome regression model that appears in the plug-in estimator in (1.3). We call these two estimators primal IPW and dual IPW.

Lemma 2 (Primal and Dual IPWs)**.**
Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, $\psi(t) = \psi(t)_{primal} = \psi(t)_{dual}$ where

$$\begin{aligned}
\psi(t)_{primal} &\equiv \mathbb{E}\Big[\mathbb{I}(T=t) \times \frac{\sum_T \prod_{V_i \in \mathbb{L}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))}{\prod_{V_i \in \mathbb{L}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))} \times Y\Big], \\
\psi(t)_{dual} &\equiv \mathbb{E}\Big[\frac{\prod_{V_i \in \mathbb{M}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \mid_{T=t}}{\prod_{V_i \in \mathbb{M}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))} \times Y\Big].
\end{aligned} \tag{2.6}$$

The representation of $\psi(t)$ as $\psi(t)_{\text{primal}}$ and $\psi(t)_{\text{dual}}$ in Lemma 2 immediately yields the corresponding primal and dual IPW estimators, via evaluating the expectations empirically and using the plug-in principles. In Theorem 1, $\psi(t)_{\text{primal}}$ and $\psi(t)_{\text{dual}}$ appear in the pieces that correspond to the variables in $\mathbb{M}$ and $\mathbb{L}$, respectively, that are preceded by all the other variables in the topological order.

2.1.1 Restrictions Implied by an ADMG

An ADMG $\mathcal{G}(V)$ may encode two types of equality constraints: ordinary conditional independence statements such as $V_i \perp\!\!\!\perp V_j \mid V_k$, and more general equality constraints, known as *Verma* constraints, that resemble conditional independences albeit in post-intervention distributions [28]. Grouped together, these are known as *equality constraints.* In our earlier work [34], we provide a sound and complete

algorithm that characterizes when the statistical model of an ADMG $\mathcal{G}(V)$, i.e., $\mathcal{M}(\mathcal{G})$, is *nonparametric saturated;* meaning $\mathcal{M}(\mathcal{G})$ imposes no equality restrictions on $p(V)$. As mentioned earlier when $\mathcal{M}(\mathcal{G}) = \mathcal{M}_{\text{nps}}$, then the tangent space of the corresponding ADMG model consists of the entire Hilbert space.

In a nonparametric saturated model, there exists a single unique influence function. Hence, the estimator that we obtain by solving $\mathbb{E}[U_\psi] = 0$, where U_ψ is given by Theorem 1 when T is p-fixable, is not only doubly robust but also the most efficient estimator. On the other hand, constraints in a semiparametric model shrink the tangent space Λ, and thus expand its orthogonal complement $\Lambda^\perp$. As $\Lambda^\perp$ expands, we will have more than one influence function (note that the class of all influence functions is $\{U_\psi + \Lambda^\perp\}$.)

In trying to achieve semiparametric efficiency bounds for our target parameter $\psi(t)$ under the restrictions implied by an ADMG, both ordinary and Verma constraints must be given consideration when deriving the tangent space of the model. Among these two kinds of equality constraints, Verma constraints are more difficult to handle as the restrictions hold in kernels obtained after recursive fixing operations. Instead, we identify a class of ADMGs, termed *mb-shielded ADMGs*, where given a topological order τ, all equality constraints implied by the ADMG $\mathcal{G}(V)$ can be written as ordinary conditional independence statements. For the class of causal models that can be expressed as an mb-shielded ADMG, we derive the form of the efficient influence function under p-fixability, that takes advantage of the Markov restrictions implied on the observed data. See [34] for details.

2.2 High Dimensional Treatments

An illustrative example of a causal relationship between a high dimensional treatment and an outcome that we will use in this section is the relationship between radiation exposure and side effects in cancer patients undergoing radiation therapy. This relationship is of clinical interest in radiation oncology and used to inform radiation treatment planning. In neck and head cancers, for example, minor variations in dose and direction of radiation may result in similar tumor reduction but vastly improve secondary outcomes, such as weight loss, or dysfunction induced by radiation therapy, such as dysphasia or xerostomia [49].

Unlike standard treatments, representable by binary random variables, radiation therapy is complicated and is represented by three dimensional voxel maps of radiation doses in different parts of the body. Since this representation is very high dimensional, the exact dose localization information in the voxel map is sometimes summarized by cumulative dose-volume histograms. Even such summaries are high dimensional, and complicate establishing a clinically relevant causal relationship between treatment and outcomes in this setting.

Seemingly natural approaches to dimension reduction, such as principal component analysis (PCA), are not appropriate in the setting we consider, for two reasons. First, since we are interested in dimension reduction for the sake of explicating a particular relationship between treatments and outcomes, approaches that do not take outcomes into account in the right way run the risk of distorting the estimate of this relationship, or even falsely concluding the relationship is absent. Second, causal relationships between treatments and outcomes (regardless of whether treatments are high dimensional) are difficult to discern due to spurious associations introduced by confounding, which is ubiquitous in observational data sources.

In this section, we provide a framework for structural (causal) models that reduce

dimension of a high dimensional treatment while preserving the causal relationship of this treatment and the outcome [50]. Our framework is based on a novel combination of methods from semiparametric inference, and sufficient dimension reduction (SDR) [51]. Our methods are appropriate in settings where the dimension of the treatment is smaller than effective sample size, leaving open to future work the important case of problems where treatment dimension exceeds the sample size, and ideas from the sparsity literature will likely be required [52].

We start by a quick overview of SDR in Section 2.2.1, before moving to our approach to semiparametric causal SDR in Section 2.2.2. In Section 2.2.3 we describe the estimation and implementation strategy of our estimators in more detail. We report simulation study results in Section 2.2.4, along with a real data application in Section 2.2.4. Our conclusions are in Section 2.3. We defer proofs of all claims to Appendix IV.

2.2.1 Sufficient Dimension Reduction

Given an outcome variable Y and a p-dimensional covariate vector C, the goal of SDR is to find a known function $g_C(.;\beta)$ parameterized by β with a much smaller range than domain such that Y depends on C only through $g_C(C;\beta)$. Often this function is assumed to be linear, in which case the goal is to find $\beta \in \mathbb{R}^{p\times d}$, where $d < p$, such that Y depends on C only through $C^T\beta$. We may be interested in a stronger type of dependence, where the conditional cumulative distribution of Y depends only on $C^T\beta$, i.e., $Pr(Y \leq y \mid C) = Pr(Y \leq y \mid C^T\beta)$, or a weaker type of dependence, where the regression function for Y only depends on C through $C^T\beta$, i.e. $\mathbb{E}[Y \mid C] = \mathbb{E}[Y \mid C^T\beta]$. The space of matrices β for which the former type of dependence holds is called the *central subspace*, while the space of matrices β for which the latter type of dependence holds is called the *central mean subspace*.

There exists a rich literature on how to derive the central (mean) subspace. Ex-

amples include, but are not limited to, sliced inverse regression [51], sliced average variance estimation [53], directional regression [54], kernel inverse regression [55], average derivative estimation [56], nonlinear least squares [57], and principal Hessian directions [58]. However, all these proposed solutions to SDR rely on strong parametric assumptions that are unlikely to hold in practical applications, such as the linearity condition where $\mathbb{E}[C \mid C^T\beta]$ is assumed to be a linear function of C, or the assumption that $\text{cov}(C \mid C^T\beta)$ is constant rather than a function of C. Ma and Zhu [59] introduced a new approach to SDR by recasting the problem in terms of estimation in a semiparametric model. Crucially, this approach relies on far weaker assumptions than is typical in SDR, and is thus much more generally applicable.

A Semiparametric Approach to SDR for the Conditional Mean

If we are interested in SDR on the mean scale, we must find a class of matrices β such that $\mathbb{E}[Y \mid C] = \mathbb{E}[Y \mid C^T\beta]$ is satisfied. The semiparametric approach in [59] recast this problem as a parameter estimation problem in a semiparametric model. To obtain the relevant semiparametric model, we rewrite the above condition as $Y = \ell(C^T\beta) + \epsilon$, where $\ell(C^T\beta) = \mathbb{E}[Y \mid C^T\beta]$ is an unspecified smooth function, and $\mathbb{E}[\epsilon \mid C] = 0$, while the distribution $p(\epsilon \mid C)$ remains otherwise unrestricted. In this model, we are interested in estimating the set of *target parameters* β given the infinite dimensional set of parameters in the *nuisance models* $p(\epsilon \mid C)$ and $\ell(C^T\beta)$. Ma and Zhu [59] derived the orthogonal complement of the nuisance tangent space for this model as,

$$\Lambda^{\perp} = \Big\{ \Big(Y - \mathbb{E}[Y \mid C^T\beta] \Big) \times \Big(\alpha(C) - \mathbb{E}[\alpha(C) \mid C^T\beta] \Big) \Big\}, \tag{2.7}$$

where $\alpha(C)$ is any function of C.

A well-known property of semiparametric models is that all elements of $\Lambda^{\perp}$ are mean zero under the true distribution. Hence, a general class of estimating equations

can be obtained using the sample version of

$$\mathbb{E}[U(\beta)] = \mathbb{E}\Big[\Big(Y - \mathbb{E}[Y \mid C^T\beta]\Big) \times \Big(\alpha(C) - \mathbb{E}[\alpha(C) \mid C^T\beta]\Big)\Big] = 0, \qquad (2.8)$$

where $U(\beta)$ is an arbitrary element in $\Lambda^{\perp}$. The estimator obtained by solving the above estimating equation is doubly robust under any choice of models for $\mathbb{E}[Y \mid C^T\beta]$ and $\mathbb{E}[\alpha(C) \mid C^T\beta]$, meaning that the estimator remains consistent if either of these two models is correctly specified [59].

We are interested in applying SDR ideas to reducing the dimension of a treatment in a way that preserves a *causal* rather than *associational* relationship with the outcome. In addition, we are interested in doing so under the weakest possible assumptions, which entails generalizing the semiparametric approach in [59]. In the remaining of this chapter, we use semiparametric inference theory developed for marginal structural models [60] to give what we believe is the first approach to causal SDR of a high dimensional treatment.

2.2.2 Causal Sufficient Dimension Reduction

We are interested in reducing the dimension of the treatment T such that the causal relationship of T and Y is preserved. Let $g(.;\beta)$ be a function parameterized by β that takes values in $\mathbb{R}^p$ and map them to values in $\mathbb{R}^d$, $d < p$, i.e., $g : T \in \mathbb{R}^p \mapsto g(T;\beta) \in \mathbb{R}^d$. We want to reduce the dimension of T in such a way that the counterfactual response $\mathbb{E}[Y(t)]$ only depends on T via $g(t)$. Specifically, we assume that if $\mathbb{E}[Y(t)]$ is identified, that is if $\mathbb{E}[Y(t)]$ is a mapping f from values t of T to functionals $h_t(p(V))$ of the observed data distribution, where $p(V)$ denotes the joint distribution over the set of observed variables V, then $f(t) = f(g(t;\beta))$. The methodology proposed in this section does not depend on the choice of $g(.;\beta)$, although we fix a particular $g(.;\beta)$ in our data analyses. We assume a conditionally ignorable model which includes the three identification assumptions that were discussed in Chapter 1; namely consistency, conditional ignorability, and positivity. Therefore,

we fix $h_a\Big(p(C,T,Y)\Big) = \mathbb{E}\Big[\mathbb{E}[Y \mid T=t, C]\Big]$, as shown in (1.2). Extensions for other identifying functionals for $\mathbb{E}[Y(t)]$ are possible but left as future work.

A Semiparametric View of Causal SDR

In Chapter 1, we described several estimation strategies for the ACE, that relied on modeling either the outcome regression $\mathbb{E}[Y \mid T, C]$ or the propensity score $p(T \mid C)$ or both. An alternative class of estimators models the relationship between the treatment and the outcome via a *marginal structural model (MSM)*, or a causal regression. A simple version of such a model takes the form $\mathbb{E}[Y(t)] = f(t;\beta)$, for finite set of parameters β. Given such a model, inferences about $\mathbb{E}[Y(t)]$ reduce to inferences about β. For binary treatments, $f(t;\beta)$ can be written as $\beta_0 + \beta_t \times t$ without loss of generality, with ACE $= \beta_t$. A marginal structural model is different from an ordinary regression model, since $\mathbb{E}[Y(t)]$ is equal to (1.2) and not $\mathbb{E}[Y \mid T = t]$ given our causal assumptions. Therefore, one approach to estimating β is to solve a modified set of estimating equations for regression problems appropriately reweighted by the propensity scores

$$\mathbb{E}\left[\frac{p^*(t)}{\pi_t(C;\widehat{\eta}_{tr})} \times \{Y - f(t;\beta)\}\right] = 0, \tag{2.9}$$

where $p^*(t)$ is an arbitrary function of t with the same dimension as β and $\widehat{\eta}_{tr}$ is the maximum likelihood estimate of η_{tr}.

The estimation procedure for MSMs shown in (2.9) can be viewed as a standard set of estimating equations for a regression model relating treatments and outcome, but applied to observed data readjusted via inverse weighting in such a way that treatment variables appear randomly assigned. In other words, MSMs are regressions applied to a version of observed data in such a way that regression parameters can be interpreted causally.

A key observation is that unlike other estimating equations that solve for β by maximizing the feature outcome relationship, the equation in (2.8) fits β to maintain

the identity $\mathbb{E}[Y \mid C] = \mathbb{E}[Y \mid C^T\beta]$. As a consequence, semiparametric causal SDR can be viewed as an MSM version of this regression problem, which seeks to find β which maintains

$$\mathbb{E}\Big[Y(t)\Big] = \mathbb{E}\Big[Y(g(t;\beta))\Big]$$

In other words, in semiparametric causal SDR, our aim is to estimate β by maintaining the following identity

$$\mathbb{E}\Big[\mathbb{E}\Big[Y \mid T = t, C\Big]\Big] = \mathbb{E}\Big[\mathbb{E}\Big[Y \mid g(t;\beta), C\Big]\Big], \tag{2.10}$$

where the outer expectation is with respect to the density $p(C)$. To reiterate, we view the treatment T as a single, albeit high dimensional, variable. By contrast C may include many relevant covariates that need to be controlled for to eliminate the confounding bias.

We note here the different roles that variables play in regression SDR and causal SDR. The goal of regression SDR is to preserve the associative relationship between high dimensional features C and the outcome Y. The goal of causal SDR, as we view it here, is to preserve the causal relationship between a high dimensional treatment T and the outcome Y, which is made complicated by the presence of spurious associations induced by covariates C. Thus, the goal of our causal SDR procedure is *not* to maintain the regression relationship between all features and the outcome by assuming $\mathbb{E}[Y \mid \{T, C\}] = \mathbb{E}[Y \mid g(\{T, C\}; \beta)]$, but to preserve the relationship as in (2.10) where C is marginalized (adjusted for). Note that the set of confounders C could still be high dimensional, but they are not of primary interest in our problem. Incorporating baseline covariates into the dimension reduction strategy along with treatments, as is done in some MSMs, is left as an interesting avenue for future work.

As stated earlier, our objective is to preserve the causal effect of treatment T and outcome Y. However, it suffices to say that if the counterfactual response curve, i.e., $\mathbb{E}[Y(t)]$, is preserved under our dimensionality reduction scheme, then the causal effect

is preserved. Hence, we stated our constraint in (2.10) in terms of the counterfactual mean rather than the counterfactual contrast that would define the effect. Moreover, we fix T to denote the high dimensional treatment. Even though treatment is high dimensional, we emphasize that each unit still receives one treatment. An example of such treatment is receiving a single session of radiation therapy (with no followups). The record of radiation treatment is usually stored as monodimensional cumulative dose-volume histograms, and is summarized as amount of radiation on $k\%$ of the organ's volume, where k ranges from 1 to 100. In this example, we can think of treatment as a vector in $\mathbb{R}^{100}$.

In a conditionally ignorable causal model, intervention on treatment T corresponds to dropping the term $p(T \mid C)$ from the observed density $p(Y, T, C)$. Define $q(Y, T, C)$ as the following modified joint distributions: $p(Y \mid T, C) \times \widetilde{p}(T) \times p(C)$, where $\widetilde{p}(T)$ is any density with the same support as $p(T)$. Then (2.10) can be rewritten as

$$\mathbb{E}_q[Y \mid T = t] = \mathbb{E}_q[Y \mid g(t; \beta)], \tag{2.11}$$

where $\mathbb{E}_q$ is the expectation taken with respect to the density $q(Y, T, C)$ defined above, and $q(Y \mid T) = \sum_C q(Y, C \mid T) = \sum_C p(Y \mid T, C) \times p(C)$ by definition. The notation in (2.11) makes drawing similarities between the constraints in the causal SDR and regular SDR settings more clear.

Equations (2.10) and (2.11) are equivalent forms of our constraint in the causal SDR problem, where the MSM model for $\mathbb{E}[Y(t)] = \mathbb{E}_q[Y \mid T = t]$ is assumed to be a function of the high dimensional treatment intervention t only through its lower dimension representation $g(t; \beta)$. We now describe two approaches to estimating β that maintains the required property based on combining estimation theory of MSMs [60] and the semiparametric SDR method in [59].

Inverse Probability Weighted SDR for the Counterfactual Mean

Let $\ell\Big(g(T;\beta)\Big) \equiv \mathbb{E}_q[Y \mid g(T;\beta)]$ and $\nu\Big(g(T;\beta)\Big) \equiv \mathbb{E}_q[\alpha(T) \mid g(T;\beta)]$ be two unspecified smooth functions of $g(T;\beta)$. A simple estimation strategy for β based on generalizing (2.9), entails solving

$$\mathbb{E}\left[\frac{p^*(t)}{p(T=t \mid C)} \times \widetilde{U}(\beta)\right] = 0, \tag{2.12}$$

where $\widetilde{U}(\beta) = \Big\{Y - \ell(g(t;\beta))\Big\} \times \Big\{\alpha(T) - \nu(g(t;\beta))\Big\}, p^*(t)$ is an arbitrary function of t, and $p(T \mid C)$ is a correctly specified statistical model which governs how the treatment T is assigned based on baseline characteristics C. The above estimation equation may be solved using observed data by evaluating the expectation empirically.

Lemma 3. *An estimator for β which solves (2.12) under the correct specification of $p(T \mid C)$, and either one of $\ell(g(T;\beta)) \equiv \mathbb{E}_q[Y \mid g(T;\beta)]$ or $\nu(g(T;\beta)) \equiv \mathbb{E}_q[\alpha(T) \mid g(T;\beta)]$, is consistent.*

Semiparametric Causal SDR for the Counterfactual Mean

A general approach for deriving RAL estimators of β is based on deriving $\widetilde{\Lambda}_\eta^\perp$, the orthogonal complement of the nuisance tangent space of a semiparametric model that enforces the constraint (2.10), but places no other restrictions on the observed data distribution. One approach is to derive this space explicitly, as was done in [59]. An alternative is to take advantage of general theory relating orthogonal complements of regression problems, and orthogonal complements of "causal regression problems," or MSMs, developed by [60]. Given the semiparametric model $\mathcal{M}$ induced by the restriction (2.10), we take advantage of this theory in the following result.

Theorem 2. *The orthogonal complement of the nuisance tangent space $\widetilde{\Lambda}_\eta^\perp$ for $\mathcal{M}$ contains elements of the form*

$$\widetilde{\Lambda}_\eta^\perp = \left\{\frac{\widetilde{U}(\beta)}{W_t(C)} - \phi(T,C) + \mathbb{E}[\phi(T,C) \mid C]\right\},$$

where $\phi(T, C)$ *is an arbitrary function of* T *and* C, $W_t(C)$ *is the IPW weight* $p(T = t \mid C)/p^*(t)$ *for a fixed* $p^*(t)$, *and* $\widetilde{U}(\beta)$ *is of the form*

$$\widetilde{U}(\beta) = \Big\{Y - \ell(g(t;\beta))\Big\} \times \Big\{\alpha(T) - \nu(g(t;\beta))\Big\},$$

where $\ell(g(t;\beta)) \equiv \mathbb{E}_q[Y \mid g(t;\beta)]$ *and* $\nu(g(t;\beta)) \equiv \mathbb{E}_q[\alpha(T) \mid g(T;\beta)]$. *Moreover, the most efficient estimator in this class, for any fixed* $\alpha(T)$, *is recovered by setting* $\phi^{opt}(T, C) = \mathbb{E}\Big[\frac{\widetilde{U}(\beta)}{W_t(C)} \mid T, C\Big]$.

This result also exists for multiple high dimensional treatments, using the theory for general MSMs with multiple treatments and time-varying confounders, as described in [60].

Lemma 4. *For a fixed choice of* $\alpha(T)$ *and* $p^*(T)$, *the element* $\widetilde{U}(\beta^*) \in \widetilde{\Lambda}_\eta^\perp$ *corresponding to the optimal choice of* $\phi(T, C)$ *has the form.*

$$\frac{p^*(T)}{p(T \mid C)} \times \widetilde{U}(\beta) - \frac{p^*(T)}{p(T \mid C)} \times \mathbb{E}\left[\widetilde{U}(\beta)\Big|T, C\right] + \mathbb{E}_q\left[\mathbb{E}\left[\widetilde{U}(\beta)\Big|T, C\right]\Big|C\right], \tag{2.13}$$

where $\mathbb{E}_q[.]$ *is the expectation taken with respect to the density* $q(Y, T, C) \equiv p(Y \mid T, C) \times p^*(T) \times p(C)$.

Robustness Properties

Just as $\Lambda_\eta^\perp$ in (2.7) entailed double robustness of $U(\beta)$ for semiparametric regression SDR, we now show that the structure of $\widetilde{\Lambda}_\eta^\perp$ yields additional robustness properties.

Lemma 5. *If one of* $\Big\{p(T \mid C)$, $\mathbb{E}[\widetilde{U}(\beta) \mid T, C]\Big\}$ and *one of* $\Big\{\ell(g(T;\beta)) \equiv \mathbb{E}_q[Y \mid g(T;\beta)]$, $\nu(g(T;\beta)) \equiv \mathbb{E}_q[\alpha(T) \mid g(T;\beta)]\Big\}$ *is correctly specified, then the estimator for* β *based on (2.13) is consistent and asymptotically normal with mean zero and variance equal to* $\tau^{-1} \times Var\Big(\widetilde{U}(\beta^*)\Big) \times \tau^{-1'}$, *where* $\widetilde{U}(\beta^*)$ *is given in (2.13) and* τ *is defined as* $\mathbb{E}[\frac{\partial \widetilde{U}(\beta^*)}{\partial \beta}]$.

This result implies that the estimating equation in (2.13) yields a "2×2" robustness property; i.e., (2.13) relies on four nuisance models, arranged in two sets of two. Our robustness property yields an unbiased and consistent estimator if at least one model in each set is correctly specified. In practice, since we will be dealing with high dimensional problems, correct specification of models is difficult to ensure. However, robustness properties of semiparametric estimators also implies that in regions where sufficient subset of models are approximately correct, the overall bias remains small.

Note that if $p(T \mid C)$ and one of the models in $\widetilde{U}(\beta)$ is correctly specified, the AIPW estimator using (2.13) remains consistent for any choice of $\mathbb{E}\Big[\widetilde{U}(\beta) \mid T, C\Big]$. One promising direction of future work is to consider cases where $p(T \mid C)$ and $\widetilde{U}(\beta)$ is known, and search for $\mathbb{E}[\widetilde{U}(\beta) \mid T, C]$ which yields good properties of the overall estimator. This use of the augmented IPW (AIPW) estimator is similar to that in randomized trial data, where $p(T \mid C) = p(T)$ is known by design.

2.2.3 Estimation and Implementation

In order to estimate the parameters β in 2.11, we need to solve the estimating equation $\mathbb{E}[\widetilde{U}(\beta^*)] = 0$, where $\widetilde{U}(\beta^*)$ is given in (2.13). For any $\widetilde{U}(\beta)$ of the form given in Section 2.2.2, Theorem 2, provides the class of all RAL estimators for β^*, which parameterizes the causal central mean subspace in an MSM model, along with the most efficient estimator in this class. Under the general form of $\widetilde{U}(\beta) = \Big\{Y - \ell(g(T;\beta))\Big\} \times \Big\{\alpha(T) - \nu(g(T;\beta))\Big\}$, the term $\mathbb{E}\Big[\widetilde{U}(\beta) \mid T, C\Big]$ in $\widetilde{U}(\beta^*)$ equals $\Big\{\mathbb{E}[Y \mid T, C] - \ell(g(T;\beta))\Big\} \times \Big\{\alpha(T) - \nu(g(T;\beta))\Big\}$. Hence, in the expression in (2.13), four different models are involved in estimating $\widetilde{U}(\beta^*)$, namely (i) $p(T \mid C)$, (ii) $\ell(g(T;\beta)) \equiv \mathbb{E}_q[Y \mid g(T;\beta)]$, (iii) $\nu(g(T;\beta)) \equiv \mathbb{E}_q[\alpha(T) \mid g(T;\beta)]$, and (iv) $\mathbb{E}[Y \mid T, C] = \mathbb{E}_q[Y \mid T, C]$. The last term in (2.13) is equal to $\mathbb{E}_t\Big[\mathbb{E}[U(\beta) \mid T, C]\Big]$, where $\mathbb{E}_t[.]$ is the expectation with respect to the marginal distribution of T which is evaluated empirically without additional modeling.

For a pre-specified functional form of $\ell(g(T;\beta))$, we need to fit three different nuisance models. Given models $\nu(g(T;\beta);\eta_\nu)$, $p(T \mid C;\eta_t)$, and $\mathbb{E}[Y \mid T, C;\eta_y]$ for $\nu(g(T;\beta))$, $p(T \mid C)$, and $\mathbb{E}[Y \mid T, C]$, respectively, it can be shown that if $n^{\frac{1}{4}+\epsilon}(\widehat{\eta}-\eta_0)$ is bounded in probability for some $\epsilon > 0$, then the estimating equation $\mathbb{E}[\widetilde{U}(\beta^*);\widehat{\eta}]$ yields an estimate of β with the same asymptotic properties as if the nuisance models were known. Here $\eta = \{\eta_\nu, \eta_t, \eta_y\}$, and $\widehat{\eta}$, η_0 denote the estimated and the true parameters of the nuisance models, respectively.

Theorem 3. *Let ϕ_0 denote the influence function of the estimator β obtained from the estimating equation $\mathbb{E}[\widetilde{U}(\beta^*,\eta_0)] = 0$. If $n^{\frac{1}{4}+\epsilon}(\widehat{\eta}-\eta_0)$ is bounded in probability for some $\epsilon > 0$, then the influence function corresponding to the estimator $\widehat{\beta}$ obtained from the estimating equation $\mathbb{E}[\widetilde{U}(\beta^*,\widehat{\eta})] = 0$ is the same as ϕ_0. In other words, $\widehat{\beta}$ follows the same asymptotic properties as if we knew the true nuisance models.*

The condition for the rate of convergence of nuisance models in Theorem 3 is a sufficient condition and is potentially too conservative. In practice, we might be able to use models with the slower convergence rates, see [61] for more details. [62] provides a detailed analysis of the convergence rates of nonparametric models.

Implementation

In this section, we describe in detail our procedure for estimating β by solving the empirical version of the estimating equation $\mathbb{E}[\widetilde{U}(\beta^*)] = 0$, where $\widetilde{U}(\beta^*)$ is given in (2.13). In what follows, we assume the structural dimension d, i.e., the cardinality of the range of $g(;\beta)$, is known; later in this section we discuss methods for choosing the structural dimension when it is not known a priori. We denote by $K(.)$ the Epanechnikov kernel and let $K_h(.) \coloneqq \frac{1}{h}K(./h)$ for the choice of bandwidth h. The d-dimensional kernel function is a product of d univariate kernel functions, i.e., $K_h(u) = K(u/h)/h^d = \prod_{j=1}^d K_h(u_j) = \prod_{j=1}^d K(u_j/h)/h^d$ for $u = (u_1, \ldots, u_d)$. In a slight abuse of notation, we use the same K regardless of the dimension of its argument.

Let $T \in \mathbb{R}^p$, $\beta \in \mathbb{R}^{p \times d}$, C be the baseline vector, and Y be the outcome of interest. For a given choice of $p^*(T)$ and $\alpha(T)$,

1. First estimate $\widehat{\eta}_{tr}$ and $\widehat{\eta}_y$ in $p(T \mid C; \eta_{tr})$ and $\mathbb{E}[Y \mid T, C; \eta_y]$ by maximum likelihood or nonparametric methods. These two models do not depend on β and are not updated within the iterations below.

2. Pick starting values $\beta^{(1)}$.

3. At the j^{th} iteration, given a fixed $\beta^{(j)}$, estimate $\widehat{\ell}(g(T; \beta^{(j)}))$ and $\widehat{\nu}(g(T; \beta^{(j)}))$,

$$\widehat{\ell}(g(T; \beta^{(j)})) = \frac{\sum_{i=1}^n Y_i \times K_h(g(T; \beta^{(j)}) - g(T_i; \beta^{(j)}))}{\sum_{i=1}^n K_h(g(T; \beta^{(j)}) - g(T_i; \beta^{(j)}))},$$
$$\widehat{\nu}(g(T; \beta^{(j)})) = \frac{\sum_{i=1}^n \alpha(T_i) \times K_h(g(T; \beta^{(j)}) - g(T_i; \beta^{(j)}))}{\sum_{i=1}^n K_h(g(T; \beta^{(j)}) - g(T_i; \beta^{(j)}))},$$

   and compute the following:

$$U^q(\beta^{(j)}) \equiv \Big\{Y - \widehat{\ell}(g(T; \beta^{(j)}))\Big\} \times \Big\{\alpha(T) - \widehat{\nu}(g(T; \beta^{(j)}))\Big\},$$
$$\mathbb{E}[U^q(\beta^{(j)}) \mid T, C] \equiv \Big\{\mathbb{E}[Y \mid T, C; \widehat{\eta}_y] - \widehat{\ell}(g(T; \beta^{(j)}))\Big\} \times \Big\{\alpha(T) - \widehat{\nu}(g(T; \beta^{(j)}))\Big\}.$$

4. Form the sample version of $\mathbb{E}[\widetilde{U}(\beta^*)]$ as follows, where $\mathbb{P}_n[.] := \frac{1}{n}\sum_{i=1}^n [.]_i$.

$$\zeta(\beta^{(j)}) = \mathbb{P}_n \Bigg[\frac{p^*(T)}{p(T \mid C; \widehat{\eta}_{tr})} \times \Big\{ U^q(\beta^{(j)}) - \mathbb{E}\Big[U^q(\beta^{(j)}) \;\Big|\; T, C\Big] \Big\} + \mathbb{E}_q \Big[\mathbb{E}\Big[U^q(\beta^{(j)}) \;\Big|\; T, C\Big] \;\Big|\; C \Big] \Bigg].$$

5. Calculate the first and second derivatives of $\partial\{||\zeta(\beta)||^2\}/\partial\{vec(\beta)\}$ numerically and evaluate them at $\beta^{(j)}$, then update $\beta^{(j)}$ using the Newton-Raphson rule.

6. Repeat steps (b) through (e) until convergence.

The implementation of the estimating equation in (2.12) follows a similar set of steps, except all steps pertaining to second and third terms of (2.13) are skipped. Moreover, in step (3) of the above implementation, we need to specify individual

models for $\ell(g(T;\beta)) \equiv \mathbb{E}_q[Y \mid g(T;\beta)]$ and $\mathbb{E}[Y \mid T, C] \equiv \mathbb{E}_q[Y \mid T, C]$. However, due to variation dependence of these models, it may be difficult to fit these two models in a congenial way in general. We provide an alternative approach in the following.

Estimation of an "Inverted" Structural Nested Mean Model

In order to deal with the issue of congeniality, we may opt to specify $\mathbb{E}_q[Y \mid g(T;\beta)]$ and $\widetilde{f}(T, C, \beta) = \mathbb{E}_q[Y \mid T, C] - \mathbb{E}_q[Y \mid g(T;\beta)]$, which yield a variationally independent specification of $\mathbb{E}_q[Y \mid g(T;\beta)]$ and $\mathbb{E}_q[Y \mid T, C] = \mathbb{E}_q[Y \mid g(T;\beta)] + \widetilde{f}(T, C, \beta)$. Consequently, the four variationally independent models we need to specify are as follows: $\ell(g(T;\beta))$, $\nu(g(T;\beta))$, $p(T \mid C)$, and $\widetilde{f}(T, C, \beta)$. The last term in (2.13) can be evaluated empirically without additional modeling. In addition, we need to specify one more nuisance model to estimate $\widetilde{f}$, which we describe below.

We fit $\widetilde{f}$ by borrowing ideas from the theory of structural nested mean models (SNMMs) in [63, 60]. Unlike MSMs, which are regression models for causal relationships, SNMMs directly model the so called "blip effects," namely counterfactual differences between the response to a particular treatment, and a response to a reference treatment, given a particular observed trajectory. For a single treatment, this difference simplifies to $\gamma(T, C; \psi) = \mathbb{E}[Y(T) \mid T, C] - \mathbb{E}[Y(0) \mid T, C]$. Let $U_{sn}(\psi) \equiv Y - \gamma(T, C; \psi)$. Consequently, $\mathbb{E}[U_{sn}(\psi) \mid T, C] = \mathbb{E}[Y(0) \mid T, C] = \mathbb{E}[Y(0) \mid C] = \mathbb{E}[U_{sn}(\psi) \mid C]$ (by conditional ignorability). The following estimating equation leads to a consistent estimation of parameters ψ,

$$\mathbb{E}\Big[\{d(T, C) - \mathbb{E}[d(T, C) \mid C]\} \times \{U_{sn}(\psi) - \mathbb{E}[U_{sn}(\psi) \mid C]\}\Big] = 0,$$

where $d(T, C)$ is a function of T and C with the same cardinality as ψ [63]. Assuming $\widetilde{f}$ is parameterized by ψ, we now show that estimating ψ can be viewed as an estimation problem for a kind of "inverted SNMM."

Lemma 6. *Let $U_{dim}(\psi) = Y - \widetilde{f}(T, C, \beta; \psi)$ and fix any $d(T, C)$. If either $\mathbb{E}[d(T, C) \mid$*

$g(T;\beta)]$ *or* $\mathbb{E}[U_{dim}(\psi) \mid g(T;\beta)]$ *are correctly specified, the following estimating equations yield a consistent estimator of* ψ,

$$\mathbb{E}\Big[\{d(T,C) - \mathbb{E}[d(T,C) \mid g(T;\beta)]\} \times \{U_{dim}(\psi) - \mathbb{E}[U_{dim}(\psi) \mid g(T;\beta)]\}\Big] = 0.$$

For the purposes of robustness, specifying both $\mathbb{E}[\widetilde{f} \mid g(T;\beta)]$ and $\mathbb{E}[U_{dim}(\psi) \mid g(T;\beta)]$ correctly is part of the correct specification of $\mathbb{E}[U(\beta) \mid T, C]$, given the type of estimation strategy we use.

The implementation provided earlier can be modified to take advantage of modeling congenial models. Right before step (c), we need to estimate $\widehat{\widetilde{f}^{(j)}}\Big(T, C, \beta^{(j)}; \widehat{\psi}\Big)$ using Lemma 6, and modify step (c) by letting $\mathbb{E}[U^q(\beta^{(j)}) \mid T, C] = \widehat{\widetilde{f}^{(j)}} \times \Big\{\alpha(T) - \widehat{\nu}(g(T;\beta^{(j)}))\Big\}$. A downside of estimating congenial models is that the overall procedure becomes quite computationally intensive.

Choosing the Structural Dimension in Causal SDR

Up until here, we assumed the structural dimension was known a priori. Finding the correct dimension is not an straightforward task and incorrect choices may greatly affect performance. We adapt the technique in [59] that was used to select the structural dimension in regression SDR to causal SDR. Specifically, we utilize a resampling procedure to select the structural dimension. This procedure was originally described by [64] and adapts the idea of [65]. We consider a family of functions $g^1(.;\beta^1), \ldots, g^m(.;\beta^m)$ with different structural dimensions, and use cross-validation procedure we describe below to pick the best dimension.

Let $\widehat{\beta}_\rho$ be the estimate of β from the original sample for the ρth working dimension, where $\rho = 1, \ldots, p-1$, and let $\widehat{\beta}_{\rho,b}$ be the estimate of β from the bth bootstrap sample, for $b = 1, \ldots, B$. The structural dimension can be estimated by finding the dimension ρ to be the cardinality of the range of the function

$$g^* \;=\; \arg\max_{g^i} \;\frac{1}{B}\;\sum_{b=1}^{B}\; r^2\Big(g^i(T;\widehat{\beta}_\rho),\; g^i(T;\widehat{\beta}_{\rho,b})\Big),$$

where $r^2(u,v) = k^{-1}\sum_{i=1}^{k}\lambda_i$ and λ_is are the non-zero eigenvalues of

$$\{\text{var}(u,v)\}^{-1/2}\ \text{cov}(u,v)\ \{\text{var}(v)\}^{-1}\ \text{cov}(v,u)\ \{\text{var}(u)\}^{-1/2}.$$

This procedure uses resampling to choose β to maximize variability of the reduced set of features given by $g^i(.;\beta^i)$ where $g^i(.;\beta^i)$ is chosen in a way that aims to preserve the causal regression relationship between T and the mean of Y. Exploring other alternatives is an interesting area for future work.

2.2.4 Simulations and Data Analysis

We illustrate the utility of our causal SDR proposal through simulations and a real data application in radiation oncology.

Simulation Study

Causal SDR is not well-solved via standard methods for dimension reduction such as PCA, as they do not take the feature/outcome relationship into account, nor by standard SDR methods, as they do not take the confounding issues into account. In this section, we illustrate the utility of our proposal to causal SDR, via simulation studies, and compare them with regression SDR and PCA methods. We also illustrate the consistency of our estimators and illustrate the procedure for selecting the structural dimension. To provide continuity with previous work, our simulation study follows that described in [59].

We perform 50 replications with fixed sample sizes, where the true response $\mathbb{E}[Y(g(t))]$ is an object of dimension $d = 2$, and the observed data distribution $p(Y,T,C)$ is set as follows. The dimension of the baseline factors C is fixed as 4 and the observed treatment dimension p is set to be 6 and 12. The baseline factors C are generated from a standard multivariate normal distribution. We consider two cases for the treatment vector: one where the linearity and the constant covariance conditions in regular SDR are violated, and one where these assumptions are satisfied.

Case 1. We generated $(T_1, T_2)^T$ (when $p = 6$) and $(T_1, T_2, T_{7:12})^T$ (when $p = 12$) from a multivariate normal distribution where the mean of each component is given as: $\mu_1 = \sum_i C_i$, $\mu_2 = \sum_i (-1)^i C_i$, $\mu_7 = C_1$, $\mu_8 = C_2$, $\mu_9 = C_3$, $\mu_{10} = -C_1 + C_2$, $\mu_{11} = -C_2 + C_3$, $\mu_{12} = -C_3 + C_4$, and the covariance matrix is $(\sigma_{ij})_{(p-4)\times(p-4)}$ where $\sigma_{ij} = 0.5^{|i-j|}$. We generated T_3 from a normal distribution with mean $|T_1 + T_2|$ and variance $|T_1|$. T_4 has a normal distribution with mean $|T_1 + T_2|^{1/2}$ and variance $|T_2|$. T_5 and T_6 were generated from Bernoulli distributions with success probabilities $\exp(T_2)/\{1 + \exp(T_2)\}$, and $\Phi(T_2)$, respectively, where $\Phi(.)$ denotes the standard normal cumulative distribution.

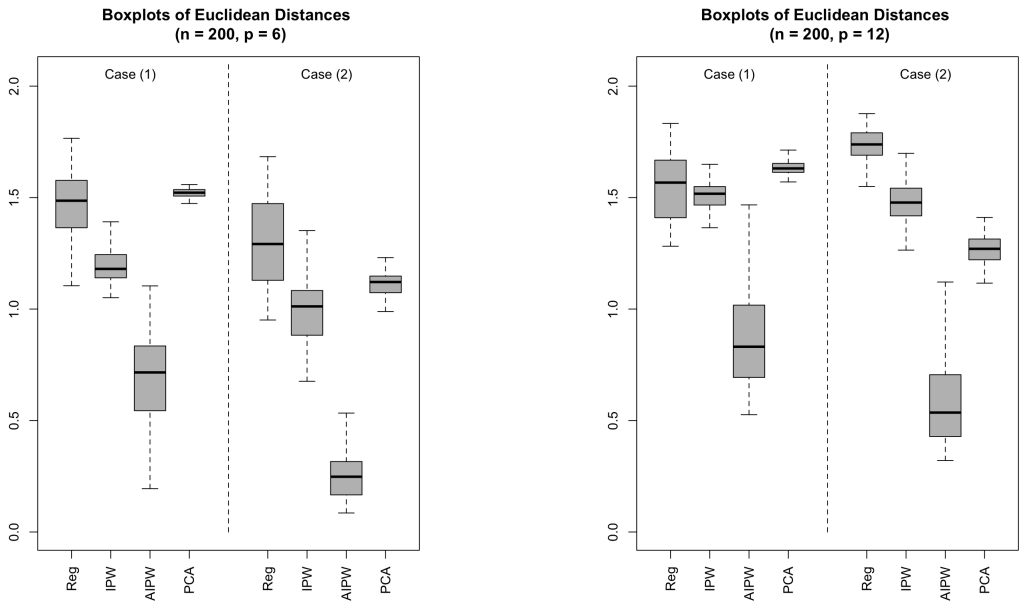
Case 2. The treatment vector is generated from a multivariate normal distribution where the mean of each component is given as follows. $\mu_1 = \sum_i C_i, \mu_2 = \sum_i (-1)^i C_i, \mu_3 = C_1 - C_2 - C_3 + C_4, \mu_4 = -C_1 + C_2 + C_3 - C_4, \mu_5 = \sum_i C_i - 2C_3, \mu_6 = \sum_i C_i - 2C_1$, and $\mu_{6+i} = C_i, \mu_{9+i} = -C_i$ for $i = 1, 2, 3$, and the covariance matrix is $(\sigma_{ij})_{p\times p}$ where $\sigma_{ij} = 0.5^{|i-j|}$.

The response variable is generated using

$$Y = T^T\beta_1 + (T^T)^2\beta_2 + \sum_{i=1}^{4} C_i + \Big\{\sum_{j=1}^{p} T_j\Big\} \times \Big\{\sum_{i=1}^{4} T_i\Big\} + \epsilon,$$

where T^T reads as the transpose of the vector T and the error term ϵ is generated from standard normal. For $p = 6$, we set $\beta_1 = (1, 1, 1, 1, 1, 1)^T/\sqrt{6}$, and $\beta_2 = (1, -1, 1, -1, 1, -1)^T/\sqrt{6}$. For $p = 12$, the last 6 components of β_1 and β_2 are identically zero.

As mentioned in Section 2.2.2, Theorem 2 provides the whole class of estimating equations for a given $\widetilde{U}(\beta)$. For simplicity, we assume $\mathbb{E}[\alpha(T) \mid g(T;\beta)] = 0$, and therefore $\widetilde{U}(\beta) = \{Y - \ell(g(T;\beta))\} \times \alpha(T)$ in the following simulations. The accuracy of the estimates was computed using the distance between the true β, and $\widehat{\beta}$ defined as the Frobenius norm of the matrix $\widehat{\beta}(\widehat{\beta}^T\widehat{\beta})^{-1}\widehat{\beta}^T - \beta(\beta^T\beta)^{-1}\beta^T$.



(a) (b)

Figure 2-2: Boxplots of Frobenius norms between true and estimated parameters in simulations.

Simulation 1. The boxplots of estimation accuracies, with $n = 200$, are reported in Fig. 2-2. The results for both Case 1 and Case 2 when $p = 6$ are presented in Fig. 2-2(a) and the results for both Case 1 and Case 2 when $p = 12$ are presented in Fig. 2-2(b). In each case, there are 4 different boxplots. The first one, from the left hand side, labeled as *Reg*, corresponds to semiparametric SDR estimating equation (2.8). Since regular SDR ignores the influence of confounding variables C, the estimates are not capturing the true causal relationship between T and Y. In the second boxplot, labeled as *IPW*, we use the IPW estimator in (2.12) with the correct model for $p(T \mid C)$, by properly adjusting for all the confounders. This recovers a more reasonable β^* estimate than the first one. However, while *IPW* generally performs better than PCA or regression SDR, the improvement is relatively modest. This might be due to the inefficiency of naive IPW estimators at the reported sample size. The third plot, labelled *AIPW*, uses the augmented IPW (AIPW) estimator corresponding to (2.13), which greatly outperforms the other estimators. The last plot corresponds to the

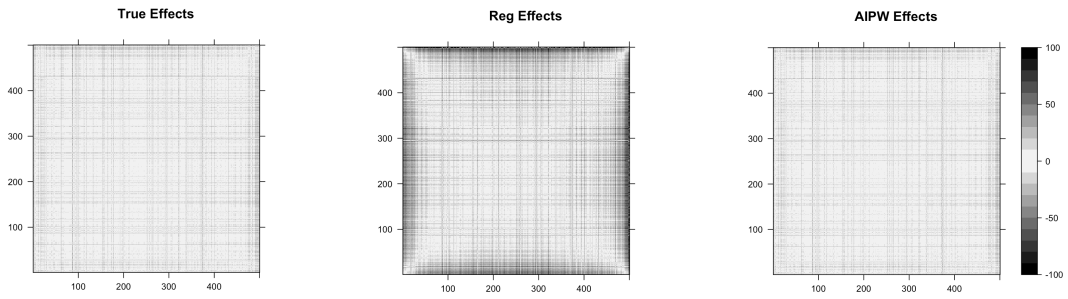


Figure 2-3: Heatmaps of true causal effects and effects computed by estimating β via the regular SDR and the AIPW estimators. Heatmaps are antidiagonally symmetric.

classical PCA dimension reduction technique where the treatment-outcome relation is ignored. In this case, the first two principal directions are reported as estimating the basis of the lower dimensional space. As illustrated in the plots, this naive approach does not seek to preserve a causal, nor indeed *any*, relationship to the outcome.

Note that our original objective was to reduce the dimension of the treatment such that the cause-effect relation between the treatment and the outcome is preserved. In order to show that our estimating procedures actually preserve this relation, we compute the contrast between $E[Y(g(t_i;\beta))]$ and $\mathbb{E}[Y(g(t_j;\beta))]$ for $i,j = 1,\ldots,n$, given the true parameters and the estimated ones. The $n \times n$ heatmap of effects are provided in Fig. 2-3 for the true effects and the ones estimated by regular SDR and AIPW. We used 500 sample points generated from Case 2 with $p = 6$ to plot these heatmaps. The plots in 2-3(a) and (c) demonstrates the significant similarity between the true surface and the one estimated by AIPW. The surface estimated by regression SDR appears to be a very different surface. The root-mean-squared errors between the true causal surface and the ones estimated from AIPW and regular regression SDR are 0.48 and 14.29, respectively.

Simulation 2. We now illustrate the performance of the bootstrap procedure for estimating the structural dimension d. We use the same data generating process as in Simulation 1, with $p = 6$, and $n = 200$. We set the bootstrap size to $B = 50$. The

relative frequency of the selected dimension are reported in Table 2-I. The bootstrap procedure reliably recovers the true structural dimension, namely 2.

Table 2-I: Choosing the structural dimension in Causal SDR

Model ($p = 6$)	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} = 3$	$\hat{d} = 4$	$\hat{d} = 5$
Case 1	0%	98%	2%	0%	0%
Case 2	0%	90%	10%	0%	0%

Simulation 3. In the third set of simulations, we demonstrate the effect of sample size on *IPW* and *AIPW* estimators of β in the causal SDR model. Results are shown in Fig. 2-4. While both estimators are consistent under our model specification, AIPW exhibits favorable convergence rates compared to IPW, as expected.

Real Data Application

We now illustrate our methods using a cohort of patients treated with radiation therapy for head and neck cancer. The cohort consists of 613 patients who received radiation therapy at the Johns Hopkins hospital prior to 2016. Radiation therapy is one of the most effective modalities for the treatment of head and neck cancers. However, because of the complex shape of target volumes in close proximity to sensitive organs, it may be associated with acute and late radiation morbidities such as xerostomia, mucositis, and dysphagia affecting the patient's quality of life. Such morbidities can lead to severe reduction in food intake and undesirable and possibly dangerous weight loss in patients. There are prospective studies that evaluated risk factors for weight loss in patients who undergo radiation therapy [66, 67]. However, a proper analysis of whether radiation causes weight loss has not yet been reported likely due to the methodological challenges involved in using high dimensional variables such as radiation therapy as a treatment in causal analysis.

In our data analysis, we focus on the parotid glands which are incidentally irradiated by radiation and examine the summary measures of radiation therapy given by the

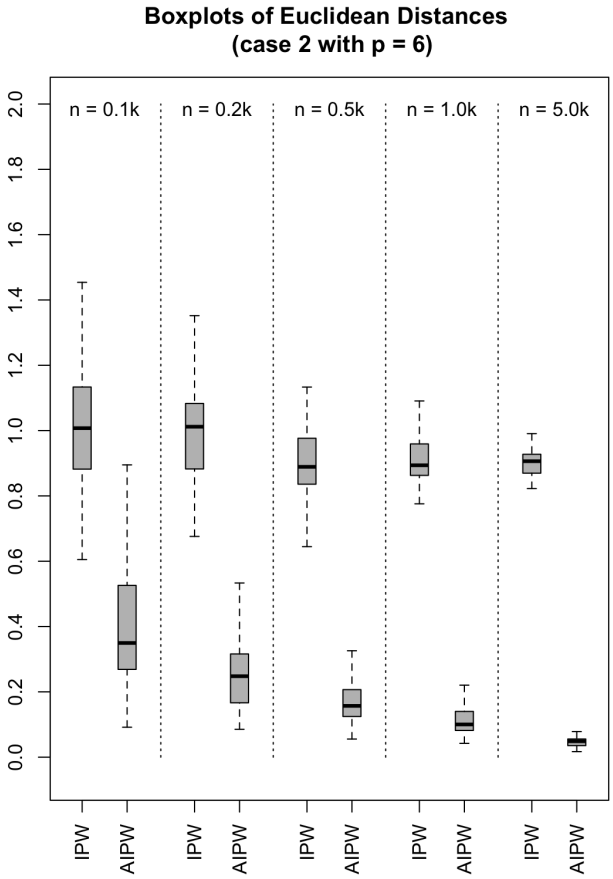


Figure 2-4: Illustration of the effect of sample size on the Frobenius norms between true and estimated parameters using data generated from Case 2 with $p = 6$.

cumulative dose-volume histograms extracted from the raw voxel maps of radiation doses. In particular, we looked at 5 equally spaced percentages of volume to construct a vector of treatment doses. We used weight loss as the outcome of interest, which was defined as the difference between weight measured within 100 to 160 days after the completion of treatment and the weight measured during consultation before the start of treatment. The data has records on demographics such as age, gender, race, and baseline clinical factors such as whether the patient had used feeding tubes and/or received chemotherapy before the RT initiation. We assumed these variables are sufficient to control for confounding and therefore would ensure the conditional ignorability assumption was met.

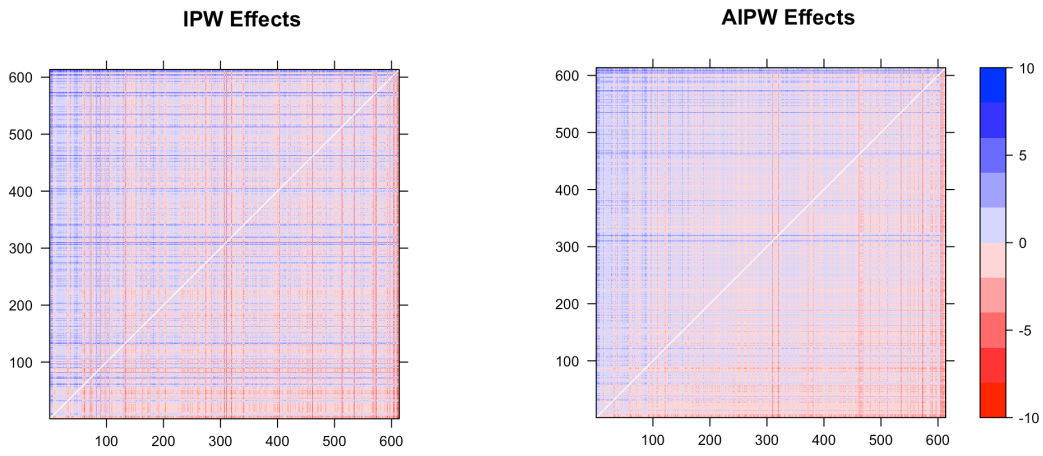


Figure 2-5: Heatmaps to illustrate the causal effect of radiation on weight loss, where effects are computed by estimating β via (a) IPW estimator, and (b) AIPW estimator. Heatmaps are antidiagonally symmetric with opposite color tones.

There exists a rich literature relating parotid dose-volume characteristics to radiotherapy-induced salivary toxicity. It has been shown that the mean dose to the parotid glands correlates strongly with xerostomia and salivary dysfunction which are risk factors of weight loss [68]. In light of such studies, we assume there exists a single dimension in the radiation exposure that captures the relationships between exposure and side effects including weight loss. Therefore, we set the structural dimension d to be one. We set the mapping function $g(.;\beta)$ to be linear in its parameters β, and use Bayesian additive regression trees to fit all nuisance models.

We generated $n \times n$ heatmaps in Fig. 2-5 to illustrate the cause-effect relationship between radiation treatment and weight loss. On the left panel, we use IPW estimator in (2.12) to estimate the parameters β. On the right panel, we use AIPW estimator obtained from Theorem 2. The absolute values on the plots are antidiagonally symmetric. Radiation doses were sorted in increasing values along both axes. We interpret the heatmaps as follows. Consider the $(i,i)^{\text{th}}$ point on the plot and draw a line along the y-coordinate. Since radiation doses were sorted in increasing order, then the radiation value at any point on the line to the right of (i,i) is higher

than the radiation value at the $(i,i)^{\text{th}}$ point. For any point to the left of (i,i), the radiation value is lower. The value at the $(k,i)^{\text{th}}$ coordinate corresponds to the contrast $\mathbb{E}\Big[Y(g(a_k;\beta)) - Y(g(a_i;\beta))\Big]$. Consequently, if $k > i$, then a red dot at (k,i) coordinate implies that an increase in radiation doses leads to an increase in weight loss. On the other hand, a blue dot would imply that an increase in radiation doses would not lead to an increase in weight loss. Similarly, a blue dot at (k,i), for $k < i$, would imply that a decrease in radiation leads to a decrease in weight loss. Reverse is implied when the dot is red.

According to Fig. 2-5, the computed effects using IPW and AIPW agree in most regions of the heatmaps. Focusing on the bottom right triangle, the one below the anti-diagonal, we note that most of the area is filled with red color. It implies that as we increase the amount of radiation, the severity of weight loss increases. In other words, radiation therapy is potentially a cause of weight loss in patients who undergo radiation therapy. In general, AIPW estimator is preferred over IPW estimator due to its doubly robust characterization and efficiency gains.

We investigated the relationship between the treatment and outcome as the treatment size increases by selecting larger numbers of equally spaced percentages of volume in the dose-volume histograms. The plots are provided in Appendix IV. Throughout the analysis, we used a crude summary of the treatment that itself had dimension greater than one. A more fine-tuned approach is to look at the raw voxel maps. A voxel-based approach would identify the relations between radiation-induced morbidity and local dose release, thus providing a potentially better insight into spatial signature of radiation sensitivity in composite regions like the head and neck district [69]. Given the small cohort of patients that we have access to, a voxel-based approach would fall into $p \gg n$ paradigm, and would require strong sparsity assumptions to deal with. This is an interesting and challenging direction for future work.

2.3 Conclusions

In the first section, we bridged the gap between identification and estimation theory for the causal effect of a single treatment on a single outcome in hidden variable causal models associated with directed acyclic graphs (DAGs). We provided a simple graphical criterion, primal fixability, which when satisfied allows for the derivation of two novel IPW estimators – primal and dual IPW. We further derived the nonparametric influence function under p-fixability of the treatment that yields the augmented primal IPW estimator and showed that it is doubly robust in the models used in primal and dual IPW estimators. We considered restrictions on the tangent space implied by the latent projection acyclic directed mixed graph (ADMG) of the hidden variable causal model. In [34], we provide an algorithm that is sound and complete for the purposes of checking the nonparametric saturation status of a hidden variable causal model as long these hidden variables are unrestricted. Further, through the use of mb-shielded ADMGs, we provide a graphical criterion that defines a class of hidden variable causal models whose score restrictions resemble those of a DAG with no hidden variables. For the class of causal models that can be expressed as an mb-shielded ADMG, we then derive the form of the efficient influence function under p-fixability, that takes advantage of the Markov restrictions implied on the observed data. These results are completely generic and may be used to derive the efficient version of any nonparametric influence function in the model with these restrictions.

In the second section, we have described a generalization of the semiparametric sufficient dimension reduction (SDR) approach for regression problems described in [59] to causal SDR. Specifically, we developed a method that reduces the dimension of a high dimensional treatment, while preserving the causal relationship between the treatment and the outcome quantified as a counterfactual mean. Using ideas from structural models [60], we provided semiparametric estimators for parameters

of the function that maps the high dimensional treatment to a lower dimensional subspace. We have shown our estimator exhibits "2x2 robustness," where the estimator remains consistent if one of two models, for two pairs of models, is chosen correctly. In order to scale our methods to high dimensional applied settings, such as fMRI scans, text data, or radiation oncology voxel data, we need to incorporate ideas from parametric modeling, and sparsity within a semiparametric framework. In prior work, we proposed an approach to trading off interpretability and performance in prediction models using our ideas on sufficient dimensionality reduction [70]. Another natural extension for future work is to apply these methods to classical causal inference in longitudinal studies, where multiple time points render a collection of binary treatments a high dimensional object. Our causal SDR approach would provide an alternative to parametric marginal structural models typically employed in such settings.

Chapter 3

A Causal View of Algorithmic Fairness

With the proliferation of comprehensive databases and advancements in artificial intelligence (AI) and machine learning (ML) algorithms, highly impactful decisions are increasingly being automated. Amongst the exciting achievements in ML, there are arising concerns regarding stereotyping and unfair determinations that are present in every corner of AI. ML predictive algorithms have been used in sentencing and parole decisions [71, 72], in child welfare services [73], in evaluating personal loan applications and insurance [74, 75], and as a job applicant screening tool by firms [76].

The fuel of automated decision-making is data, and in order to build an effective intelligent machine, we need as much relevant information as possible. Data may well include sensitive features, such as race and gender, that must be treated with care due to the risk of enabling discrimination. Even in the absence of such variables in individual data, other features may be present that are highly correlated with sensitive features, and so even decisions based on data which has no variables corresponding to (e.g.) race or gender may exhibit significant disparities along these dimensions. For example, an individual's zip code is a very effective *proxy* for race in racially segregated communities and decisions informed by algorithms that use zip code information may thereby introduce (or reproduce) significant racial disparities [77].

Why are unjust disparities potential byproducts of learning algorithms? ML algorithms use training data to learn a function that maps the input to the output by finding patterns in the training data. However, data can sometimes reflect historical patterns of discrimination, bias, and/or inequality due to the way data are collected and stored, the way important variables are defined, or the way hypotheses are framed. As an example, in criminal justice settings recidivism is often defined as a subsequent arrest rather than subsequent conviction. This can have substantial consequences for judicial decisions given background policing practices. Similarly, features such as prior compensation and employment history in resume screenings may be heterogeneous across genders and other traits. There is no information in the data to indicate whether heterogeneity of this type may be due to unfair differences in treatment of these groups.

Consequently, learning algorithms that rely on data from our unfair world can lead to biased or unfair conclusions, and using the output of such algorithms may serve to perpetuate systemic injustice. To break this cycle, three methodological questions must be answered. *First*, how should fairness principles be expressed mathematically, such that these requirements may be productively combined with the statistical models and algorithms used to inform crucial decisions? *Second*, how can learning algorithms be modified such that they produce fair outputs, even when their input training data comes from an unfair world? *Third*, how should we use the fair model on new instances?

The core part of our proposal is to leverage the formal language of causal modeling to mathematically specify fairness constraints and prevent algorithms from perpetuating unfairness by means of causal inference methodology and constrained optimization. We view fairness as an inherently causal notion and characterize the presence of unfairness based on a sensitive feature, like race or gender, with respect to an outcome of interest, as the presence of an effect of the sensitive feature on the outcome along

unfair causal pathways. In other words, we propose to model unfairness based on a sensitive feature with respect to an outcome as the presence of an effect of the feature on the outcome along certain impermissible causal pathways. We divide this chapter into two sections. In the first section, we consider the problem of making fair predictions [78, 79]. In the second section, we consider how to extend learning fair predictions to learning fair policies [80].

3.1 Training Fair Predictive Models

Predictive models trained on imperfect data are increasingly being used in socially-impactful settings. Predictions (such as risk scores) have been used to inform high-stakes decisions in criminal justice [81], healthcare [82], and finance [83]. While automation may bring many potential benefits – such as speed and accuracy – it is also fraught with risks. Predictive models introduce two dangers in particular: the illusion of objectivity and violation of fairness norms. Predictive models may appear to be "neutral," since humans are less involved and because they are products of a seemingly impartial optimization process. However, predictive models are trained on data that reflects the structural inequities, historical disparities, and other imperfections of our society. A particular worry in the context of data-driven decision-making is "perpetuating injustice," which occurs when unfair dependence between sensitive features (e.g., race, gender, age, disability status) and outcomes is maintained, introduced, or reinforced by automated tools.

In this section, we study how to construct fair predictive models by correcting for the unfair causal dependence of predicted outcomes on sensitive features [78]. We propose to model unfairness based on a sensitive feature, such as race or gender, with respect to an outcome as the presence of an effect of the feature on the outcome along certain "disallowed" causal pathways. As a simple example, discussed in [2], job applicants' gender should not *directly* influence the hiring decision, but may influence

the hiring decision indirectly, via secondary applicant characteristics important for the job, and correlated with gender. We argue that fair prediction requires imposing hard constraints on the predictive model in the form of restricting certain causal path-specific effects. This view captures a number of intuitive properties of unfairness, and generalizes existing formal [2, 84] and informal proposals [85]. Impermissible pathways are user-specified and context-specific, hence require input from policymakers, legal experts, or the general public. Some alternative but also causally-motivated constrained prediction methods are proposed in [86, 87]. For a survey and discussion of distinct fairness criteria (both causal and associative) see [88].

This section is organized as follows. We first give a brief introduction to mediation and path-specific effects, which will be necessary to formally define our approach to fair inference. We then formalize unfairness with respect to the sensitive feature and the outcome in terms of unfair path-specific effects. Moving forward, we show that fair inference from finite samples under our definition can be viewed as a certain type of constrained optimization problem, and discuss a number of complications to the basic framework of fair inference. We illustrate our framework via experiments on real datasets in the experimental section.

3.1.1 Mediation and Path-Specific Effects

In causal inference, we might be interested in understanding the mechanisms by which some treatment T influences some outcome Y. A common framework for studying mechanisms is known as *mediation analysis* which seeks to decompose the effect of T on Y into the *direct effect* and the *indirect effect* mediated by a third variable, or more generally into components associated with particular causal pathways. As an example, the direct effect of T on Y in Fig. 1-1(a) corresponds to the effect along the edge $T \to Y$ and the indirect effect corresponds to the effect along the path $T \to M \to Y$, mediated by the variable M.

In the potential outcome notation, the direct and indirect effects can be defined using nested counterfactuals, such as $Y(t, M(t'))$ for $t, t' \in \mathfrak{X}_T$, which reads as the potential outcome Y when T is set to t while M is set to whatever value it would have attained had T been set to t'. The *natural direct effect* (NDE) (on the expectation difference scale) is defined as $\mathbb{E}[Y(t, M(t'))] - \mathbb{E}[Y(t')]$ and the *natural indirect effect* (NID) is defined as $\mathbb{E}[Y(t)] - \mathbb{E}[Y(t, M(t'))]$. Under certain identification assumptions discussed in [89], the distribution of $Y(t, M(t'))$ (and thereby direct and indirect effects) can be nonparametrically identified from observed data by the following formula:

$$p(Y(t, M(t')) = \sum_{C,M} p(Y \mid T = t, C, M) \times p(M \mid T = t', C) \times p(C).$$

More generally, when there are multiple *proper* pathways from T to Y (a proper causal path only intersects T at the source node) one may define various *path-specific effects* (PSEs). The effect along a specific path will be obtained by comparing two potential outcomes, one where for the selected paths all nodes behave as if $T = t$, and along all other paths nodes behave as if $T = t'$.

PSEs are defined by means of nested path-specific potential outcomes. Fix a set of treatment variables T, and a subset of *proper causal paths* π from any element in T. Next, pick a pair of value sets t and t' for elements in T. For any $V_i \in V$, define the potential outcome $V_i(\pi, t, t')$ by setting T to t for the purposes of paths in π, and to t' for the purposes of proper causal paths from T to Y not in π. Formally, for any $V_i \in V$, $V_i(\pi, t, t') \equiv a$ if $V_i \in T$, otherwise

$$V_i(\pi, t, t') \equiv V_i\Big(\Big\{V_j(\pi, t, t') \mid V_j \in \mathrm{pa}_{\mathcal{G}}^{\pi}(V_i)\Big\}, \Big\{V_j(t') \mid V_j \in \mathrm{pa}_{\mathcal{G}}^{\overline{\pi}}(V_i)\Big\}\Big), \tag{3.1}$$

where $V_j(t') \equiv t'$ if $V_j \in T$ and given by recursive substitution otherwise, $\mathrm{pa}_{\mathcal{G}}^{\pi}(V_i)$ is the set of parents of V_i along an edge which is a part of a path in π, and $\mathrm{pa}_{\mathcal{G}}^{\overline{\pi}}(V_i)$ is the set of all other parents of V_i.

A counterfactual $V_i(\pi, t, t')$ is said to be *edge inconsistent* if counterfactuals of the form $V_j(t_k, \ldots)$ and $V_j(t'_k, \ldots)$ occur in $V_i(\pi, t, t')$, otherwise it is said to be

edge consistent. It is known that a joint distribution $p(V(\pi, t, t'))$ containing an edge-inconsistent counterfactual $V_i(\pi, t, t')$ is not identified in the structural causal model (nor weaker causal models) with a corresponding graphical criterion on π and $\mathcal{G}(V)$ called the "recanting witness" [90, 91]. Under some assumptions, PSEs are nonparametrically identified by means of the *edge g-formula* described in [91].

Example 3.1. As an example, consider the DAG in Fig. 3-1(a). The PSE of T on Y along the paths $\pi = \{T \to Y, T \to L \to Y\}$ is encoded by a counterfactual contrast of the form $Y(\pi, t, t') = Y(t, M(t'), L(t, M(t')))$. The corresponding counterfactual density is identified by a special case of the edge g-formula as follows:

$$p(Y(t, M(t'), L(t, M(t'))) = \sum_{C,M,L} p(Y \mid T = t, C, M) \times p(L \mid T = t, M, C) \times p(M \mid T = t', C) \times p(C).$$

3.1.2 Unfair Path-Specific Effects

A common class of approaches for fair inference is to quantify fairness via an associative (rather than causal) relationship between the sensitive feature S and the outcome Y. One difficulty with non-causal formalization of fairness, as in [92, 93], is their inability to distinguish appropriate from inappropriate sources of association. As an example, direct use of a sensitive feature such as race is discriminatory, while denial of services based on a strong proxy for a sensitive feature, such as geographic location, is a form of redlining. However, the use of certain non-sensitive features may be justified, even if they are correlated with a sensitive feature. The associative measures of fairness have difficulties distinguishing these cases. Further, these associative criteria are not easily adaptable to use context-specific information, and they oftentimes are tailored to only classification problems. On the other hand, our causal view to algorithmic fairness takes into account the mechanisms through which variables are related which leads to interesting methodological problems.

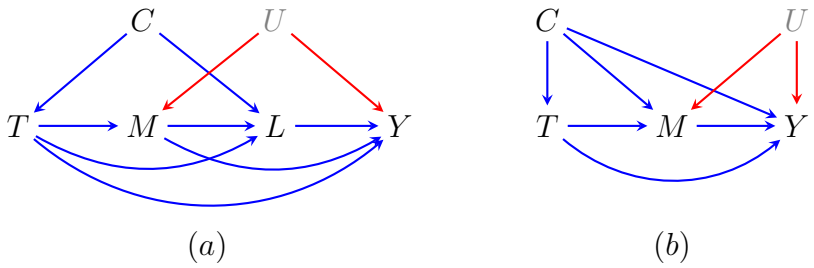


Figure 3-1: (a) A causal graph with two mediators, one confounded with the outcome via an unobserved common cause. (b) A causal graph with a single mediator where the natural direct effect is not identified. Unmeasured confounders are denoted by U.

Consider a hiring example where potential discrimination is with respect to sex (a variable randomized at conception, which means worries about confounding are no longer relevant). As before, consider binary variables S and H for sex and hiring, and an additional vector M, representing applicant characteristics relevant for the job, of the kind that would appear on the resume. One might argue that it is legitimate to consider job characteristics in making hiring decisions *even if* those characteristics are correlated with sex. However, it is not legitimate to consider sex *directly*. This intuition underscores resume "name-swapping" experiments where identical resumes are sent for review with names switched from a male sounding name to a female sounding name [85]. In such experiments, name serves as a proxy for sex as a direct determinant of the hiring decision.

The definition of (un)fairness as related to causal pathways is further supported in the legal literature. The following definition of employment discrimination, which appeared in the legal literature [94] and was cited by [2], makes clear the counterfactual nature of our conception of fairness:

> The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same.

The counterfactual "had the employee been of a different sex" phrase entails considering, for women, the outcome Y had sex been male $S = 1$, while the "everything else had been the same" phrase entails considering job characteristics under the original gender $S = 0$. The resulting counterfactual $Y(S = 1, M(S = 0))$ is precisely the one used in mediation analysis to define natural direct effects.

It is possible to construct examples, discussed further, where some causal pathways from a sensitive variable to the outcome are impermissible and unfair, and others are not. Thus, our view is that unfairness ought to be formalized as the presence of certain path-specific effects. The specific paths which correspond to unfairness are a *domain specific issue*. For example, physical fitness tests may be appropriate to administer for certain physically demanding jobs, e.g., construction, but not for white-collar jobs, such as accounting. As a result, a path from sex to the result of a test to a hiring decision may or may not be (un)fair, depending on the nature of the job.

Non-Identification of the PSE

Suppose our problem entailed the causal model in Fig. 3-1 (a) or (b) where in both cases only the NDE of T on Y is unfair. Existing identification results for PSEs [90] imply that the NDE is not identified in either model. This means estimation of the NDE from observed data is not possible as the NDE is not a function of the observed data distribution in either model.

In such cases, three approaches are possible. In both cases, the unobserved confounders U are responsible for the lack of identification. If it were possible to obtain data on these variables, or obtain reliable proxies for them, the NDE becomes identifiable in both cases. If measuring U is not possible, a second alternative is to consider a PSE that is identified, and that includes the paths in the PSE of interest and *other paths*. For example, in Fig. 3-1 (a), while the NDE of T on Y, which is the PSE including only the path $T \to Y$, is not identified, the PSE which includes paths

$T \to Y$, $T \to M \to Y$, and $T \to M \to L \to Y$.

If we are using the PSE on the mean difference scale, the magnitude of the effect which includes more paths than the PSE we are interested in must be an upper bound on the magnitude of the PSE of interest in order for the bounds we impose to actually limit unfairness. This is only possible if, for instance, all causal influence of T on Y along paths involved in the PSE are of the same sign. In Fig. 3-1 (a), this would mean assuming that if we expect the NDE of T on Y to be negative (due to unfairness), then it is also negative along the paths $T \to M \to L \to Y$, and $T \to M \to Y$.

If measuring U is impossible, and it is not possible to find an identifiable PSE that includes the paths of interest from T to Y and serves as a useful upper bound to the PSE of interest, the other alternative is to use bounds derived for non-identifiable PSEs. While finding such bounds is an open problem in general, they were derived in the context of the NDE with a discrete mediator in [95].

The issue with non-identification of the PSE was also noted in [84]. They proposed to change the causal model, specifically by cutting off some paths from the sensitive variable to the outcome such that the identification criterion in [90] became satisfied, and the PSE became identified. We disagree with this approach, as we believe it amounts to "redefining success." If the original causal model truly represents our beliefs about the structure of the problem, and in particular the pathways corresponding to discrimination, then making any sort of inferences in a model modified away from truth no longer tracks reality. We would certainly not expect any kind of repair within a modified model to result in fair inferences in the real world. The workarounds for non-identification we propose aim to stay within the true model, but try to obtain information on the true non-identified PSE, either by non-parametric bounds, or by including other pathways along with the "unfair" pathways.

3.1.3 Constraining Unfair Path-Specific Effects

Consider an observed data distribution $p(Z)$ induced by a causal model, where $Z = \{Y, C, S, M\}$ includes outcome Y, all baseline factors C, sensitive features S, and mediators M between S and Y. Context and background ethical considerations pick out some path-specific effect of the sensitive feature S on the outcome Y as unfair. We assume this effect is identified as some function of the observed distribution: $g(p_Z)$. Fix upper and lower bounds ϵ_l, ϵ_u for the PSE, representing a tolerable range. The most relevant bounds in practice are $\epsilon_l = \epsilon_u = 0$ or approximately zero. We propose to transform the inference problem on $p(Z)$, the "unfair world," into an inference problem on another distribution $p^*(Z)$, called the "fair world," which is close in the sense of minimal KL-divergence to $p(Z)$ while also having the property that the PSE lies within (ϵ_l, ϵ_u) [78].

Given a dataset $\mathcal{D} = \{Z_i = (Y_i, C_i, S_i, M_i), i = 1, \ldots, n\}$ drawn from $p(Z)$, a likelihood function $\mathcal{L}(\mathcal{D}; \alpha)$ parameterized by α, an estimator $\widehat{g}(p_Z)$ of the unfair PSE, and bounds ϵ_l, ϵ_u, we suggest to approximate $p^*(Z)$ by solving the following constrained maximum likelihood problem [78]:

$$\widehat{\alpha} \;=\; \arg\max_{\alpha} \;\; \mathcal{L}_Z(\mathcal{D}; \alpha) \quad \text{subject to} \quad \epsilon_l \leq \widehat{g}(p_Z) \leq \epsilon_u. \tag{3.2}$$

Having approximated the fair world $p^*(Z; \widehat{\alpha})$ in this way, we point out a key difficulty for using these estimated parameters to predict outcomes for new instances (e.g., new job applicants). A new set of observations Z is not sampled from the "fair world" $p^*(Z)$ but from "unfair world" $p(Z)$. Here, we propose to map new instances from p to p^* and use the result for predicting Y with constrained model parameters $\widehat{\alpha}$. We assume Z can be partitioned into Z_1 and Z_2 such that $p^*(Z) = p^*(Z_1 \mid Z_2) \times p(Z_2)$. In other words, variables in Z_2 are shared between p and p^*, i.e., $p^*(Z_2) = p(Z_2)$, but $p^*(Z_1 \mid Z_2) \neq p(Z_1 \mid Z_2)$. Z_1 typically corresponds to variables that appear in the estimator $\widehat{g}(p_Z)$. There is no obvious principled way of knowing exactly what values of

Z_1 the "fair version" of the new instance would attain. Consequently, all such possible values are averaged out, weighted appropriately by how likely they are according to the estimated p^*. This entails predicting Y as the expected value $\mathbb{E}^*[Y \mid Z_2]$, with respect to the distribution $\sum_{Z_1} p^*(Y, Z_1 \mid Z_2)$.

The optimization problem in (3.2) involves complex non-linear constraints on the parameter space. This makes the proposed constrained optimization a daunting task that relies on complex optimization software (or computationally expensive methods such as rejection sampling), which do not always find high quality local optima. In [79], we provide a novel reparameterization of the observed data likelihood in which unfair path-specific effects appear directly as parameters. This allows us to greatly simplify the constrained optimization problem.

Fair Inference via Reparameterized Likelihoods

We now describe how to reparameterize the observed data likelihood in terms of causal parameters that correspond to path-specific effects. The result presented in the following theorem greatly simplifies the constrained optimization problem (3.2) in settings where the PSE includes the direct influence of S on Y. This is due to the fact that the constrained parameter, corresponding to the PSE of interest, now appears as a single coefficient in the outcome regression model.

Theorem 4. *Assume the observed data distribution $p(Z)$ is induced by a causal model where $Z = \{Y, X\}$ denotes the observed data and $X := \{C, S, M\}$ includes baseline measures C, binary sensitive feature S, and a set of mediators M, between S and Y. Let $p(Y(\pi, s, s'))$ denote the potential outcome distribution that corresponds to the effect of S on Y along unfair causal paths in π, where π includes the direct edge $S \to Y$, and let $p(Y_0(\pi, s, s'))$ denote the identifying functional for $p(Y(\pi, s, s'))$ obtained from the edge g-formula, where the term $p(Y \mid X)$ is evaluated at $\{X \setminus S\} = 0$. Then $\mathbb{E}[Y \mid X]$*

can be written as follows:

$$\mathbb{E}[Y \mid X] = f(X) - \Big(\mathbb{E}[Y(\pi, s, s')] - \mathbb{E}[Y_0(\pi, s, s')]\Big) \; + \; \phi(S),$$

where $f(X) := \mathbb{E}[Y \mid X] - \mathbb{E}[Y \mid S, \{X \setminus S\} = 0]$ *and* $\phi(S) = w_0 + w_s S$. *Furthermore,* w_s *corresponds to* π*-specific effect of* S *on* Y.

Given Theorem 4, the constrained optimization problem in eq. (3.2) significantly simplifies to the following optimization problem:

$$\widehat{\alpha} \;=\; \arg\max_{\alpha} \;\; \mathcal{L}_Z(\mathcal{D}; \alpha) \quad \text{subject to} \quad \epsilon_l \leq w_s \leq \epsilon_u, \tag{3.3}$$

where α contains w_s and the nonlinear constraint has been replaced by a box-constraint on the parameter w_s.

Furthermore, in the optimization problem in (3.2), we propose to constrain only part of the likelihood. Specifically we do not constrain the density $p(C)$ over the baseline features (since this is high-dimensional and thus inplausible to model accurately in their parametric approach). The baseline density is instead estimated by placing $1/n$ mass at every observed data point. This is sub-optimal in the specific setting we consider, where we do not need to average over constrained variables. Constraining a larger part of the joint distribution should lead to a fair world distribution KL-closer to the observed distribution, which leads to better predictive performance as long as the likelihood is correctly specified. In [79], we demonstrate how tools from the empirical likelihood literature [96] can be readily adapted to construct hybrid (semiparametric) observed data likelihoods that satisfy given fairness criteria. With this approach, the entire likelihood is constrained, rather than only part of the likelihood as proposed above. As a result, we are able to use the data more efficiently and achieve better performance.

Fair Inference with Computational Bayesian Methods

Methods for fair inference described so far are fundamentally frequentist in character, in a sense that they assumed a particular true parameter value, and parameter fitting was constrained in a way that an estimate of this parameter was within specified bounds. Here, we do not extend our approach to a fully Bayesian setting, where we would update distributions over causal parameters based on data, and use the resulting posterior distributions for constraining inferences. Instead, we consider how Bayesian methods for estimating conditional densities can be adapted, as a computational tool, to our frequentist approach.

Many Bayesian methods do not compute a posterior distribution explicitly, but instead sample the posterior using Markov chain Monte Carlo approaches [97]. These sampling methods can be used to compute any function of the posterior distribution, including conditional expectations, and can be modified to obey constraints in our problem in a straightforward way. As an example, we consider BART, a popular Bayesian random forest method described in [98]. This method constructs a distribution over a forest of regression trees, with a prior that favors small trees, and samples the posterior using a variant of Gibbs sampling, where a new tree is chosen while all others are held fixed. A well known result [99] states that a Gibbs sampler will generate samples from a constrained posterior directly if it rejects all draws that violate the constraint. We implemented this simple method by modifying the R package (with a C++ backend) *BayesTree*. The experiment using the resulting constrained outcome model is described in the next section.

3.1.4 Data Analyses

We now illustrate our approach to fair inference via two datasets: the COMPAS dataset [71] and the Adult dataset [100].

The COMPAS Dataset

COMPAS is a risk assessment tool that is being used across courts in the US to determine whether to release or detain a defendant before their trial. Each pretrial defendant receives several COMPAS scores based on factors including but not limited to demographics, criminal history, family history, and social status. Among these scores, we are primarily interested in "Risk of Recidivism." Propublica [71] has obtained two years worth of COMPAS scores from the Broward County Sheriff's Office in Florida that contains scores for over 11000 people who were assessed at the pretrial stage and scored in 2013 and 2014. COMPAS score for each defendant ranges from 1 to 10, with 10 being the highest risk. Besides the COMPAS score, the data also includes records on defendant's age, sex, race, prior convictions, and whether or not recidivism occurred in a span of two years. We limited our attention to the cohort consisting of African Americans and Caucasians.

We are interested in predicting whether a defendant would reoffend using the COMPAS data. For illustration, we assume the use of prior convictions, possibly influenced by race, is fair for determining recidivism. Thus, we defined unfairness as effect along the direct path from race to the recidivism prediction outcome. The simplified causal graph model for this task is given in Figure 3-2 (a), where S denotes race, prior convictions is the mediator M, demographic information such as age and sex are collected in C, and Y is recidivism. The disallowed path in this problem is drawn in green in Figure 3-2(a). The effect along this path is the NDE. The objective is to learn a fair model for Y, i.e., a model where NDE is minimized.

We obtained the posterior sample representation of $\mathbb{E}[Y \mid S, M, C]$ via both regular and constrained BART. Under the unconstrained posterior, the NDE (on the odds ratio scale) was equal to $1.3(1.01, 1.45)$. This number is interpreted to mean that the odds of recidivism for Caucasians (on average) would have been 1.3 times higher had

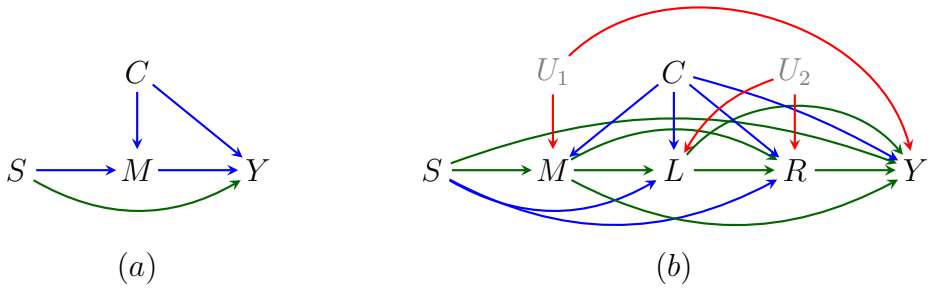


Figure 3-2: Causal graphs for (a) the COMPAS dataset, and (b) the Adult dataset.

they, contrary to the fact, been African American. In our experiment, we restricted NDE to lie between 0.95 and 1.05. Using unconstrained BART, our prediction accuracy on the test set was 67.8%, removing treatment from the outcome model dropped the accuracy to 64.0%, and using constrained BART led to the accuracy of 66.4%. As expected, dropping race led to a greater decrease in accuracy, compared to simply constraining the outcome model to obey the constraint on the NDE.

In addition to our approach to removing unfair NDE, we are also interested in assessing the extent to which the existing COMPAS recidivism classifier is biased. Unfortunately, we do not have access to the exact model which generated COMPAS scores, since it is proprietary, nor all the input features used. Instead, we used our dataset to predict a binarized COMPAS score by fitting the model $\tilde{p}(Y \mid M, C)$ using BART. We dropped race, as we know the COMPAS tool does not use that feature. Unfairness, as we defined it, may still be present even if we drop race. To assess this, we estimate the NDE, our measure of unfairness, in the semiparametric model of $p(Y, M, S, C)$, where the only constraint is that $p(Y \mid M, C)$ is equal to $\tilde{p}$ above. This model corresponds to (our approximation of) the "world" used in the COMPAS tool. Measuring the NDE on the odds ratio scale using this model yielded $2.1(2.06, 2.40)$, which is far from 1 (the null effect value). In other words, assuming the defendant is Caucasian, then the odds of recidivism would be 2.1 times higher had they been, contrary to fact, African American. Thus, our best guess on the model used in the

COMPAS tool is that it is severely unbiased against African Americans.

The Adult Dataset

The adult dataset from the UCI repository has records on 14 attributes such as demographic information, level of education, and job related variables such as occupation and work class on 48842 instances along with their income that is recorded as a binary variable denoting whether individuals have income above or below $50k$ – high vs low income. The objective is to learn a statistical model that predicts the class of income for a given individual. Suppose banks are interested in using this model to identify reliable candidates for loan application. Raw use of data might construct models that are biased against females who are perceived to have lower income in general compared to males. The causal model for this dataset is drawn in Figure 3-2(b). Gender is the sensitive variable in this example denoted by S in figure 3-2(b) and income class is denoted by Y. M denotes the marital status, L is the level of education, and R consists of three variables, occupation, hours per week, and work class. The baseline variables including age and nationality are collected in C. U_1 and U_2 capture the unobserved confounders between M, Y and L, R, respectively.

Here, besides the direct effect ($S \to Y$), we would like to remove the effect of sex on income through marital status ($S \to M \to \ldots \to Y$). The disallowed paths are drawn in green in Figure 3-2(b). The PSE along the green paths is identifiable via the recanting district criterion in [90], and can be computed by calculating odds ratio or contrast comparison of the counterfactual variable $Y(s, M(s), L(s', M(s)), R(s', M(s), L(s', M(s))), C)$, where s' is set to a baseline value, $s = 1$ in one counterfactual, and $s = 0$ in the other. The counterfactual distribution can be estimated from the following functional:

$$\sum_{Z \setminus S} \{p(Y \mid s, m, l, r, c) \times \prod_{i=1}^{3} p(r_i \mid s', m, l, c) \times p(l \mid s', m, c) \times p(m \mid s, c) \times p(c)\}.$$

If we use logistic regression to model Y and linear regression to model other variables given their past, and compute the PSE on the odds ratio scale, it is straightforward to show that the PSE simplifies to $\exp\left(\theta_s^y + \theta_m^y\theta_s^m + \theta_l^y\theta_m^l\theta_s^m + \sum_i \theta_{r_i}^y(\theta_m^{r_i}\theta_s^m + \theta_l^{r_i}\theta_m^l\theta_s^m)\right)$, where θ_i^j denotes the coefficient associated with variable i in modeling the variable j, [101]. Therefore, the constraint in (3.2) is an easy function to compute, and the resulting constrained optimization problem relatively easy to solve. Unfortunately, adapting the constrained BART procedure is computationally expensive.

We trained two models for Y, one by maximizing the constrained likelihood in (3.2) using the R package *nloptr*, and the other by using the full model with no constrain. For performance evaluation on test set, we should use $\mathbb{E}[Y \mid S, C]$ in constrained model and $\mathbb{E}[Y \mid S, M, L, R, C]$ in unconstrained model. The PSE in the unconstrained model is 3.16. This means, the odds of having a high income would have been more than 3 times higher for a female if her sex and marital status would have been the same as if she was a male. We solve the constrained problem by restricting the PSE to lie between 0.95 and 1.05. Accuracy in the unconstrained model is 82%, and drops to 72% in the constrained model while assuring that the constrained model is fair.

3.2 Optimal Fair Policies

Making optimal and adaptive intervention decisions in the face of uncertainty is a central task in precision medicine, computational social science, and artificial intelligence. In healthcare, the problem of learning optimal policies is studied under the heading of *dynamic treatment regimes* [102]. The same problem is called *reinforcement learning* in artificial intelligence [103], and *optimal stochastic control* [104] in engineering and signal processing. In all of these cases, a policy (a function of historical data to some space of possible actions, or a sequence of such functions) is chosen to maximize some pre-specified outcome quantity, which might be abstractly considered a *utility* (or

reward in reinforcement learning).

Increasingly, ideas from optimal policy learning are being applied in new contexts. In some areas, particularly socially-impactful settings like criminal justice, social welfare policy, hiring, and personal finance, it is essential that automated decisions respect principles of fairness since the relevant data sets include potentially sensitive attributes (e.g., race, gender, age, disability status) and/or features highly correlated with such attributes, so ignoring fairness considerations may have socially unacceptable consequences. A particular worry in the context of automated sequential decision making is "perpetuating injustice," i.e., when maximizing utility maintains, reinforces, or even introduces unfair dependence between sensitive features, decisions, and outcomes. Though there has been growing interest in the issues of fairness in machine learning [105, 92, 93, 106, 107, 108], so far methods for optimal policy learning subject to fairness constraints have not been well-explored.

As a motivating example, we consider a simplified model for a children's welfare screening program, recently discussed in [109, 110]. A hotline for child abuse and neglect receives many thousands of calls a year, and call screeners must decide on the basis of calculated risk estimates what action to take in response to any given call, e.g., whether or not to follow up with an in-person visit from a caseworker. The idea is that only cases with substantial potential risk to the child's welfare should be prioritized. The information used to determine the calculated risk level and thereby the agency's action includes potentially sensitive features, such as race and gender, as well as a myriad of other factors such as perhaps whether family members receive public assistance, have an incarceration history, record of drug use, and so on. Though many of these factors may be predictive of subsequent negative outcomes for the children, there is a legitimate worry that both risk calculations and policy choices based on them may depend on sensitive features in inappropriate ways, and thereby lead to unfair racial disparities in the distribution of families investigated, and perhaps

separated, by child protective services.

Learning high-quality policies that satisfy fairness constraints is difficult due to the fact that multiple sources of bias may occur in the problem simultaneously. One kind of bias, which we call *retrospective bias*, has its origin in the historical data used as input to the policy learning procedure. This data may reflect various systematic disparities and discriminatory historical practices in our society, including prior decisions themselves based on poor data. Algorithms trained on such data can maintain these inequities. Furthermore, decision making algorithms may suffer from what we call *prospective* sources of bias. For instance, suppose the functional form of the chosen decision rule explicitly depends on sensitive features in inappropriate ways. In that case, making decisions based on the new decision rule may perpetuate existing disparities or even introduce disparities that were previously absent. Avoiding this sort of bias may involve imposing non-trivial restrictions on the policy learning procedure. Finally, learning high-quality policies from observational data requires dealing with *confounding bias*, where associations between decision and reward cannot be used directly to assess decision quality due to the presence of confounding variables, as well as *statistical bias* due to misspecified statistical models. Policy learning algorithms that respect fairness constraints must address all of these sources of bias.

Our main theoretical result illustrates in what sense enacting fair policies can "break the cycle of injustice": we show how to learn policies such that the joint distribution induced by these policies (in conjunction with reward/utility mechanisms outside the policy-maker's control) will satisfy specified fairness constraints while remaining "close" to the generating distribution. To our knowledge, this work constitutes the first attempt to integrate algorithmic fairness and policy learning with the possible exception of [108], which addressed what we call prospective bias in the context of Markov Decision Processes.

To precisely describe our approach, we must introduce some necessary concepts

and tools from counterfactual policies and optimal policy learning. Then, we adapt our perspective on algorithmic fairness in prediction problems, outlined in the previous section, to learning optimal fair policies. We illustrate our proposal via experiments on synthetic and real data.

3.2.1 Policy Counterfactuals and Policy Learning

Consider a multi-stage decision problem with K pre-specified decision points, indexed by $k = 1, \ldots, K$. Let Y denote the final outcome of interest and A_k denote the action made (treatment administered) at decision point k with the finite state space of $\mathcal{A}_k$. Let X denote the available information prior to the first decision, and Y_k denote the information collected between decisions k and $k+1$, $(Y \equiv Y_K)$. $\overline{A}_k$ represents all treatments administered from time 1 to k; likewise for $\overline{Y}_k$. We combine the treatment and covariate history up to treatment decision A_k into a history vector H_k. The state space of H_k is denoted by $\mathcal{H}_k$. While our proposal applies to arbitrary state spaces, we present examples with continuous outcomes and binary decisions for simplicity.

The goal of policy learning is to find policies that map vectors in $\mathcal{H}_k$ to values in $\mathcal{A}_k$ (for all k) that maximize the expected value of outcome Y. In offline settings, where exploration by direct experimentation is impossible, finding such policies requires reasoning counterfactually, as is common in causal inference. Let $f_A = \{f_{A_1}, \ldots, f_{A_K}\}$ be a sequence of decision rules. At the kth decision point, the kth rule f_{A_k} maps the available information prior to the kth treatment decision H_k to treatment decision a_k, i.e., $f_{A_k} : \mathcal{H}_k \mapsto \mathcal{A}_k$. Given f_A we define the counterfactual response of Y had A been assigned according to f_A, or $Y(f_A)$, by the following recursive definition [111, 112]:

$$Y\Big(\Big\{f_{A_k}\Big(H_k(f_A)\Big) : A_k \in \mathrm{pa}_{\mathcal{G}}(Y) \cap A\Big\}, \Big\{\mathrm{pa}_{\mathcal{G}}(Y) \setminus A\Big\}(f_A)\Big).$$

In words: the potential outcome Y had any parent of Y that is in A been set to f_A in response to counterfactual history H_k up to k, where this history behaves as if A were set to f_A *and* any parent of Y that is not in A, behaves as if A were set to f_A.

Under a causal model associated with the DAG $\mathcal{G}$, the distribution $p(Y(f_A))$, is identified by the following generalization of the g-formula:

$$\sum_{Z\setminus\{Y,A\}} \prod_{V\in Z\setminus A} p\Big(V|\{f_{A_k}(H_k) : A_k \in \text{pa}_{\mathcal{G}}(V)\cap A\}, \text{pa}_{\mathcal{G}}(V)\setminus A\Big). \tag{3.4}$$

Given an identified response to a fixed set of policies f_A, we consider search for the optimal policy set f_A^*, defined to be one that maximizes $\mathbb{E}[Y(f_A)]$. Since $Y(f_A)$ is a counterfactual quantity, validating the found set of policies is difficult given only retrospective data, with statistical bias due to model misspecification being a particular worry. This stands in contrast with online policy learning problems in reinforcement learning, where new data under any policy may be generated and validation is therefore automatic. Partly in response to this issue, a set of orthogonal methods for policy learning have been developed that model different parts of the observed data likelihood function. Q-learning, value search, and g-estimation are common methods used in dynamic treatment regimes literature for learning optimal policies [102]. We defer detailed descriptions to later in the section.

3.2.2 From Fair Prediction to Fair Policies

In Section 3.1 and [78], we argued that fair inference for prediction requires imposing hard constraints on the prediction problem, in the form of restricting certain path-specific effects. We adapt this approach to optimal sequential decision-making [80]. We summarize this proposal with a brief example, inspired by the aforementioned child welfare case.

Consider a simple causal model for this scenario, shown in Fig. 3-3(a). Hotline operators receive thousands of calls per year, and must decide on an action A for each call, e.g., whether or not to send a caseworker. These decisions are made on the basis of a (high-dimensional) vectors of covariates X and M, as well as possibly sensitive features S, such as race. M consists of mediators of the effect of S on A. Y_1

corresponds to an indicator for whether the child is separated from their family by child protective services, and Y_2 corresponds to child hospitalization (presumably attributed to domestic abuse or neglect). The observed joint distribution generated by this causal model would be $p(Y_1, Y_2, A, M, S, X)$. Our proposal from [78] is that fairness corresponds to the impermissibility of certain path-specific effects, and so fair inference requires decisions to be made from a counterfactual distribution $p^*(Y_1, Y_2, A, M, S, X)$ which is "nearby" to p (in the sense of minimal Kullback-Leibler divergence) but where these PSEs are constrained to be zero; p^* is the distribution generated by a "fair world."

Multiple fairness concerns have been raised by experts and advocates in discussions of the child protection decision-making process [109, 110]. For example, it is clearly impermissible that race has any direct effect on the decision made by the hotline screener, i.e., that all else being held fixed, members from one group have a higher probability of being surveilled by the agency. However, it is perhaps permissible that race has an indirect effect via some mediated pathway, e.g., if race is associated with some behaviors or features which themselves ought to be taken into consideration by hotline staffers, because they are predictive of abuse. If that's true, then $S \to A$ would be labeled an impermissible pathway whereas $S \to M \to A$ (for some M) would be permissible. Similarly, it would be unacceptable if race had an effect on whether children are separated from their families; arguably both the direct pathway $S \to Y_1$ and indirect pathway though hotline decisions $S \to A \to Y_1$ should be considered impermissible. Rather than defend any particular choice of path-specific constraints, we note that our fairness framework for prediction problems can flexibly accommodate any set of given constraints, as long as the PSEs are identifiable from the observed distribution.

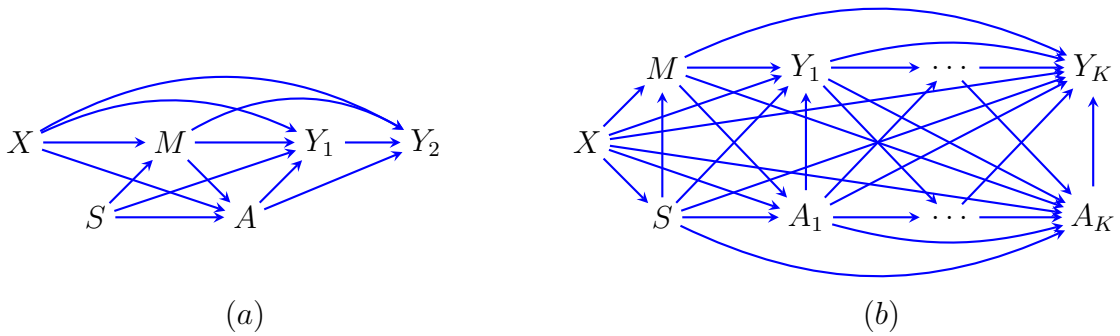


Figure 3-3: (a) A causal DAG corresponding to our (simplified) child welfare example with baseline factors X, sensitive feature S, action A, vector of mediators (including e.g. socioeconomic variables, histories of drug treatment) M, an indicator Y_1 of whether a child is separated from their parents, and an indicator of child hospitalization Y_2. (b) A multistage decision problem, which corresponds to a complete DAG over vertices $X, S, M, A_1, Y_1, \cdots, A_K, Y_K$.

Inference in a nearby "fair world"

We now describe the specifics of the proposal. We assume the data is generated according to some (known) causal model, with observed data distribution $p(\cdot)$, and that we can characterize the fair world by a fair distribution $p^*(\cdot)$ where some set of pre-specified PSEs are constrained to be zero, or within a tolerance range. Without loss of generality we can assume the utility variable Y is some deterministic function of Y_1 and Y_2 (i.e., $Y \equiv u(Y_1, Y_2)$) and thus use Y in place of Y_1 and Y_2 in what follows. Then $Z = (Y, X, S, M, A)$ in our child welfare example. For the purposes of illustration, assume the following two PSEs are impermissible: PSE^{sa}, corresponding to the direct effect of S on A and defined as $\mathbb{E}[A(s, M(s'))] - \mathbb{E}[A(s')]$, and PSE^{sy}, corresponding to the effect of S on Y along the edge $S \to Y$, and the path $S \to A \to Y$ and defined as $\mathbb{E}[Y(s, A(s, M(s')), M(s'))] - \mathbb{E}[Y(s')]$.

If the PSEs are identified under the considered causal model, they can be written as functions of the observed distribution. For example, the unfair PSE of the sensitive feature S on outcome Y in our child welfare example may be written as a functional $\text{PSE}^{sy} = g_1(p_Z) \equiv g_1(p(Y, X, S, M, A))$. Similarly the unfair PSE of S on A is $\text{PSE}^{sa} =$

$g_2(p_Z) \equiv g_2(p(Y, X, S, M, A))$. Generally, given a set of identified PSEs $g_j(p_Z) \forall j \in \{1, ..., J\}$ and corresponding tolerated lower/upper bounds $\epsilon_j^-, \epsilon_j^+$, the fair distribution $p^*(Z)$ is defined as

$$\begin{aligned} p^*(Z) \;&\equiv\; \arg\min_q \;\; D_{KL}(p \;||\; q) \\ \text{subject to} \;\; &\epsilon_j^- \leq g_j(p_Z) \leq \epsilon_j^+, \quad \forall j \in \{1, ..., J\}, \end{aligned} \tag{3.5}$$

where D_{KL} is the KL-divergence and J is the number of constraints.[1] In finite sample settings, we propose solving the following constrained maximum likelihood problem:

$$\begin{aligned} \widehat{\alpha} \;&=\; \arg\max_\alpha \;\; \mathcal{L}(Z; \alpha) \\ \text{subject to} \;\; &\epsilon_j^- \leq \widehat{g}_j(p_Z) \leq \epsilon_j^+, \quad \forall j \in \{1, ..., J\}, \end{aligned} \tag{3.6}$$

where $\widehat{g}_j(p_Z)$ are estimators for the chosen PSEs and $\mathcal{L}(Z; \alpha)$ is the likelihood function. The most relevant bounds in practice are the null values for ϵ_j^- and ϵ_j^+.

Fair decision-making

In the sequential decision setting, there are multiple complications. In particular, we aim to learn high-quality policies while simultaneously making sure that the joint distribution induced by the policy satisfies our fairness criteria, potentially involving constraints on multiple causal pathways. This problem must be solved in settings where distributions of some variables, such as outcomes, are not under the policy-maker's control. Finally, we must show that if the learned policy is adapted to new instances (drawn from the original observed distribution) in the right way, then these new instances combined with the learned policy, constrained variables, and variables outside our control, together form a joint distribution where our fairness criteria remain satisfied.

[1] Note that in our examples J will typically be $K + 1$, i.e., one constraint for the S to Y paths and one constraint for each set of paths from S to A_k. We allow for J constraints in general to accommodate more complex settings (e.g., where there are multiple sensitive features, multiple outcomes, or a different set of pathways are constrained).

Consider a K-stage decision problem given by a DAG where every vertex pair is connected, and with vertices in a topological order $X, S, M, A_1, Y_1, \ldots, A_K, Y_K$; see Fig. 3-3(b). Note that the setting where S can be assumed exogenous is a special case of this model with missing edge between X and S. Though we only assume a single set of permissible mediators M here, at the expense of some added cumbersome notation all of the following can be extended to the case where there are distinct sets of mediators $M_1, \ldots, M_K$ preceding every decision point. (We extend the results below to that setting in Appendix V.) We will consider the following PSEs as inadmissible: PSE^{sy}, representing the effect of S on Y along all paths *other than* the paths of the form $S \to M \to \ldots \to Y$; and PSE^{sa_k}, representing the effect of S on A_k along all paths *other than* the paths of the form $S \to M \to \ldots \to A_k$. That is, we consider *only* pathways connecting S and A_k or Y through the allowed mediators M to be fair. In this model, these PSEs are identified by [90]:

$$\begin{aligned}
\text{PSE}^{sy} &= \mathbb{E}[Y(s, M(s'))] - \mathbb{E}[Y(s')] \\
&= \sum_{X,M} \Big\{ \mathbb{E}[Y \mid S = s, M, X] - \mathbb{E}[Y \mid S = s', M, X] \Big\} \times p(M \mid S = s', X) \times p(X), \\
\text{PSE}^{sa_k} &= \mathbb{E}[A_k(s, M(s'))] - \mathbb{E}[A_k(s')] \\
&= \sum_{X,M} \Big\{ \mathbb{E}[A_k \mid S = s, M, X] - \mathbb{E}[A_k \mid S = s', M, X] \Big\} \times p(M \mid S = s', X) \times p(X).
\end{aligned}$$

Numerous approaches for estimating and constraining these identified PSEs are possible. Here, we restrict our attention to semiparametric estimators, which model only a part of the likelihood function while leaving the rest completely unrestricted. Estimators of this sort share some advantages with parametric methods (e.g., often being uniformly consistent at favorable rates), but do not require specification of the full probability model. Specifically, we use estimators based on the following result.

Theorem 5. *Assume S is binary. Under the causal model above, the followings are consistent estimators of PSE^{sy} and PSE^{sa_k}, assuming all models are correctly specified:*

$$\widehat{g}^{sy}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)} \times \frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} \times Y_n,$$

$$\widehat{g}^{sa_k}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)} \times \frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} \times A_{kn}.$$

These inverse probability weighted (IPW) estimators use models for M and S. Thus, we can approximate p^* by constraining only the M and S models, i.e., obtaining estimates $\hat{\alpha}_m$ and $\hat{\alpha}_s$ of the parameters α_m and α_s in $p^*(M \mid S, X; \alpha_m)$ and $p^*(S \mid X; \alpha_s)$ by solving (3.6). The outcomes Y_k and decisions A_k are left unconstrained. This is subtle and important, since it enables us to choose our optimal decision rules f_A^* without restriction of the policy space and allows the mechanism determining outcomes Y_k (based on decisions A_k and history H_k) to remain outside the control of the policy-maker. Consequently, we can show that implementing this procedure guarantees that the joint distribution over all variables Z induced by 1) the constrained M and S models, 2) the conditional distributions for A_k given H_k implied by the optimal policy choice, and 3) *any* choice of $p(Y_k \mid A_k, H_k)$ will (at the population-level) satisfy the specified fairness constraints. We prove the following result in the Supplement:

Theorem 6. *Consider the K-stage decision problem described by the DAG in Fig. 3-3(c). Let $p^*(M \mid S, , X; \alpha_m)$ and $p^*(S \mid X; \alpha_s)$ be the constrained models chosen to satisfy $PSE^{sy} = 0$ and $PSE^{sa_k} = 0$. Let $\widetilde{p}(Z)$ be the joint distribution induced by $p^*(M \mid S, X; \alpha_m)$ and $p^*(S \mid X; \alpha_s)$, and where all other distributions in the factorization are unrestricted. That is,*

$$\widetilde{p}(Z) \equiv p(X) \times p^*(S \mid X; \alpha_s) \times p^*(M \mid S, X; \alpha_m) \times \prod_{k=1}^{K} p(A_k \mid H_k) \times p(Y_k \mid A_k, H_k).$$

Then the functionals PSE^{sy} and PSE^{sa_i} taken w.r.t. $\widetilde{p}(Z)$ are also zero.

This theorem implies that any approach for learning policies based on $\widetilde{p}(Z)$ addresses both retrospective bias (since the fairness criterion violation present in $p(Z)$ is absent in $\widetilde{p}(Z)$) and prospective bias (since the criterion holds in $\widetilde{p}(Z)$ for any choice of policy on A_k inducing $p(A_k \mid H_k)$). As we discuss in detail in the next following, modified policy learning based on $\widetilde{p}(Z)$ requires special treatment of the constrained variables S and M. New instances (e.g., new calls to the child protection hotline) will be drawn from the unfair distribution p, not $\widetilde{p}$. So, the enacted policy cannot use empirically observed values of S or M. In what follows, our approach is to either average over S and M (following the procedure in Section 3.1 and [78]), or resample observations of S and M from the constrained models.

3.2.3 Estimation of Optimal Policies in the Fair World

In the following, we describe several strategies for learning optimal policies, and our modifications to these strategies based on the above fairness considerations.

Q-learning

In Q-learning, the optimal policy is chosen to optimize a sequence of counterfactual expectations called Q-functions. These are defined recursively in terms of value functions $V_k(\cdot)$ as follows:

$$Q_K(H_K, A_K) = \mathbb{E}[Y_K(A_K) \mid H_K], \quad V_K(H_K) = \max_{a_K} \; Q_K(H_K, a_K), \tag{3.7}$$

and for $k = K-1, \ldots, 1$

$$Q_k(H_k, A_k) = \mathbb{E}[V_{k+1}(H_{k+1}, A_k) \mid H_k], \quad V_k(H_k) = \max_{a_k} \; Q_k(H_k, a_k). \tag{3.8}$$

Assuming $Q_k(H_k, A_k)$ is parameterized by β_k, the optimal policy at each stage may be easily derived from Q-functions as $f^*_{A_k}(H_k) = \arg\max_{a_k} Q_k(H_k, a_k; \widehat{\beta}_k)$. Q-functions are recursively defined regression models where outcomes are value functions, and features are histories up to the current decision point. Thus, parameters β_k $(k = 1, \ldots, K)$ of

all Q-functions may be learned recursively by maximum likelihood methods applied to regression at stage k, given that the value function at stage $k+1$ was already computed for every row; see [102] for more details.

Note that at each stage k, the identity $Q_k(H_k, A_k) = \mathbb{E}[V_{k+1}(H_{k+1}, A_k) \mid H_k] = \mathbb{E}[V_{k+1}(H_{k+1}) \mid A_k, H_k]$ only holds under our causal model if the *entire past* H_k is conditioned on. In particular, $\mathbb{E}[V_{k+1}(H_{k+1}, A_k) \mid H_k \setminus \{M, S\}] \neq \mathbb{E}[V_{k+1}(H_{k+1}) \mid A_k, H_k \setminus \{M, S\}]$. To see a simple example of this, note that $Y_K(a_1)$ is not independent of A_1 conditional on just X in Fig. 3-3(b), due to the presence of the path $Y_K \leftarrow M \rightarrow A_1$; however the indepenence does hold conditional on the entire $H_1 = \{X, S, M\}$ [112].

In a fair policy learning setting, though $\{M, S\}$ may be in H_k, we cannot condition on values of M, S to learn fair policies since these values were drawn from p rather than p^*. There are multiple ways of addressing this issue. One approach is to modify the procedure to obtain optimal policies that condition on all history *other than* $\{M, S\}$. We first learn Q_ks using (3.7) and (3.8). We then provide the following modified definition of Q-functions defined directly on p^*:

$$Q_k^*(H_k \setminus \{M, S\}, A_k; \beta_k) = \frac{1}{W} \times \sum_{m,s} Q_k(H_k, A_k; \beta_k) \times \prod_{i=1}^{k} p(A_i \mid H_i \setminus \{M, S\}, m, s) \times \prod_{i=2}^{k-1} p(M_i \mid A_i, H_i \setminus \{M, S\}, m, s) \times p^*(m, s \mid X),$$

where for $k = K, \ldots, 1$,

$$W = \sum_{m,s} p^*(m, s|X) \prod_{i=1}^{k} p(A_i|H_i \setminus \{M, S\}, m, s) \prod_{i=2}^{k-1} p(M_i|A_i, H_i \setminus \{M, S\}, m, s).$$

The optimal fair policy at each stage is then derived from Q^*-functions as

$$f_{A_k}^*(H_k) \;=\; \arg\max_{a_k} \; {Q^*}_k(H_k \setminus \{M, S\}, a_k; \widehat{\beta}_k).$$

As an alternative approach, we can compute the original Q-functions defined in (3.7) and (3.8) with respect to $p^*(Z)$ by ignoring the observed values M_n and S_n for the nth individual and replacing them with samples drawn from $p^*(M \mid S, X; \alpha_m)$ and

$p^*(S \mid X; \alpha_s)$. Then, in (3.7) and (3.8), the history at the kth stage, H_k, gets replaced with $H_k^* = \{H_k \setminus \{M, S\}, M^*, S^*\}$.

Value search

It may be of interest to estimate the optimal policy within a restricted class $\mathcal{F}$. One approach to learning the optimal policy within $\mathcal{F}$ is to directly search for the optimal $f_A^{*,\mathcal{F}} \equiv \arg\max_{f_A \in \mathcal{F}} \mathbb{E}[Y(f_A)]$, which is known as *value search*. The expected response to an arbitrary policy $\phi = \mathbb{E}[Y(f_A)]$, for $f_A \in \mathcal{F}$ can be estimated in a number of ways. Often $\widehat{\phi}$ takes the form of a solution to some estimating equation $\mathbb{E}[h(\phi)] = 0$ solved empirically given samples from $p(Z)$. A simple estimator for ϕ that uses only the treatment assignment model $\pi(H_k; \psi) \equiv p(A_k = 1 | H_k)$ is the IPW estimator that solves the following estimating equation:

$$\mathbb{E}\left[\prod_{k=1}^{K} \left\{ C_{f_{A_k}} / \pi_{f_{A_k}}(H_k; \widehat{\psi}) \right\} \times Y - \phi \right] = 0, \tag{3.9}$$

where the expectation is evaluated empirically and $\widehat{\psi}$ is fit by maximum likelihood. Further, $\pi_{f_{A_k}}(H_k; \psi) \equiv \pi(H_k; \psi) \times f_{A_k}(H_k) + (1 - \pi(H_k; \psi)) \times (1 - f_{A_k}(H_k))$ and $C_{f_{A_k}} \equiv \mathbb{I}(A_k = f_{A_k}(H_k))$.

Finding fair policies via value search involves solving the same problem with respect to $p^*(Z)$ instead. Given known models $p^*(M \mid S, X; \alpha_m)$ and $p^*(S \mid X; \alpha_s)$, we may consider two approaches. The first one involves solving a modified estimating equation of the form

$$\mathbb{E}^*[h(\phi)] \equiv \mathbb{E}\left[\sum_{m,s} \mathbb{E}[h(\phi) \mid M, S, X] \times p^*(M \mid S, X; \alpha_m) \times p^*(S \mid X; \alpha_s) \right] = 0,$$

with respect to $p^*(Z \setminus \{M, S\})$.

The alternative is to solve the original estimating equation $\mathbb{E}[h(\phi)] = 0$ with respect to $p^*(Z)$ by replacing observed values M_n and S_n for the nth individual with sampled values M_n^* and S_n^* drawn from $p^*(M|S, X; \alpha_m)$ and $p^*(S|X; \alpha_s)$. In both approaches, the optimal fair policy at each stage is then derived by replacing the

history at the kth stage, H_k, with $H_k^* = \{H_k \setminus \{M, S\}, M^*, S^*\}$. Given constrained models $p^*(M|S, X; \alpha_m)$, and $p^*(S|X; \alpha_s)$ representing $p^*(Z)$, we can perform value search by solving the given estimating equation empirically on a dataset where every row x_n, s_n, m_n in the data is replaced with I rows x_n, s_{ni}^*, m_{ni}^* for $i = 1, \ldots I$, with m_{ni}^* and s_{ni}^* drawn from $p^*(M|S, x_n; \alpha_m)$ and $p^*(S|x_n; \alpha_s)$, respectively.

G-estimation

A third method for estimating policies is to directly model the counterfactual contrasts known as *optimal blip-to-zero functions* and then learn these functions by a method called g-estimation [111]. In the interest of space, we defer a full description of blip-to-zero functions and g-estimation to Appendix V, where we also present some results for our implementation of fair g-estimation.

Tradeoffs and treatment of constrained variables

We have proposed to constrain the M and S models to satisfy given fairness constraints. Since empirically observed values of M and S are sampled from p rather than p^* (or $\widetilde{p}$), our approach requires resampling or averaging over these features. The choice of models to constrain involves a tradeoff. The more models are constrained, the closer the KL distance between p and p^*, but the more features have to be resampled or averaged out; that is, some information on new instances is "lost." Alternative approaches may constrain fewer or different models in the likelihood (for example, we could have selected to constrain the Y model instead of S). However, the benefit of our approach here is that we can guarantee, with outcomes Y outside the policy-maker's control, that the induced joint distribution will satisfy the given fairness constraints (by Theorem 6), whereas alternative procedures which aim to avoid averaging or resampling will typically have no such guarantees. Another alternative that avoids averaging over variables altogether is to consider likelihood parameterizations where

the absence of a given PSE directly corresponds to setting some variation-independent likelihood parameter for the Y model to zero. While such a parameterization is possible for linear structural equation models, it is an open problem in general for arbitrary PSEs and nonlinear settings. Developing novel, general-purpose alternatives that transfer observed distributions to their "fair versions," while avoiding resampling and averaging, is an open problem left to future work.

3.2.4 Data Analyses

We now illustrate our approach to learning optimal fair policies with both synthetic data and COMPAS criminal justice data.

Synthetic data

We generated synthetic data for a two-stage decision problem according to the causal model shown in Fig. 3-3(c) ($K = 2$), where all variables are binary except for the continuous response utility $Y \equiv Y_2$. Details on the specific models used are reported in Appendix V. We generated a dataset of size 5000, with 100 bootstrap replications, where the sensitive variable S is randomly assigned and where S is chosen to be an informative covariate in estimating Y.

We use estimators in Theorem 5 to compute PSE^{sy}, PSE^{sa_1}, and PSE^{sa_2} which entail using M and S models. In this setting, the PSE^{sy} is 1.918 (on the mean scale) and is restricted to lie between -0.1 and 0.1. The PSE^{sa_1} is 0.718, and PSE^{sa_2} is 0.921 (on the odds ratio scale) and both are restricted to lie between 0.95 and 1.05. We only constrain M and S models to approximate p^* and fit these two models by maximizing the constrained likelihood using the R package `nloptr`. The parameters in all other models were estimated by maximizing the likelihood.

Optimal fair polices along with optimal (unfair) policies were estimated using the two techniques described in Section 3.2.3 (where we used the "averaging" approach

in both cases). We evaluated the performance of both techniques by comparing the population-level response under fair policies versus unfair policies. One would expect the unfair policies to lead to higher expected outcomes compared to fair policies since satisfying fairness constraints requires sacrificing some policy effectiveness. The expected outcomes under unfair polices obtained from Q-learning and value search were 7.219 ± 0.005 and 7.622 ± 0.265, respectively. The values dropped to 6.104 ± 0.006 and 6.272 ± 0.133 under fair polices, as expected. In addition, both fair and unfair optimal polices had higher expected outcomes than the observed population-level outcome, using both methods. In our simulations, the population outcome under observed policies was 4.82 ± 0.007. Some additional results are reported in the Supplement.

The COMPAS dataset

The COMPAS dataset includes records on risk scores (A), defendant's age ($X_1 \in X$), gender ($X_2 \in X$), race (S), prior convictions (M), and whether or not recidivism occurred in a span of two years (R). We limited our attention to the cohort consisting of African Americans and Caucasians, and to individuals who either had not been arrested for a new offense or who had recidivated within two years. Our sample size is 5278. All variables were binarized including the COMPAS score, which we treat as an indicator of a binary decision to incarcerate versus release (pretrial) "high risk" individuals, i.e., we assume those with score ≥ 7 were incarcerated. In this data, 28.9% of individuals had scores ≥ 7.

Since the data does not include any variable that corresponds to utility, and there is no uncontroversial definition of what function one should optimize, we define a heuristic utility function from the data as follows. We assume there is some (social, economic, and human) cost, i.e., negative utility, associated with incarceration (deciding $A = 1$), and that there is some cost to releasing individuals who go on to reoffend (i.e., for whom $A = 0$ and $R = 1$). Also, there is positive utility associated with releasing

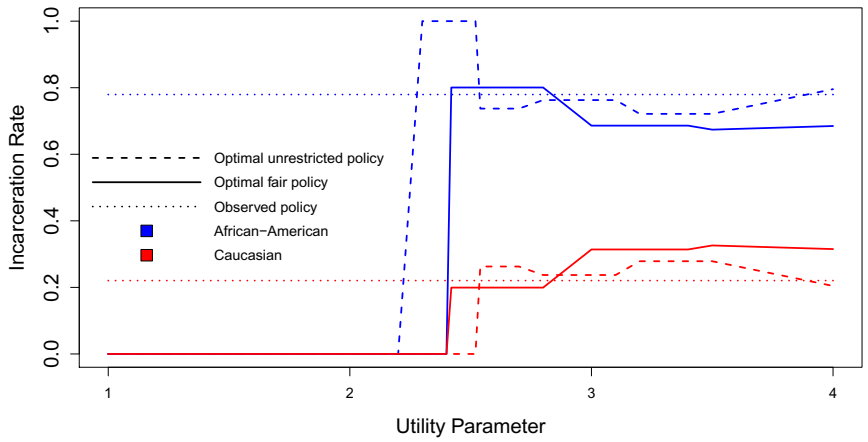


Figure 3-4: Group-level incarceration rates for the COMPAS data as a function of the utility parameter θ.

individuals who do not go on to recidivate (i.e., for whom $A = 0$ and $R = 0$). A crucial feature of any realistic utility function is how to balance these relative costs, e.g., how much (if any) "worse" it is to release an individual who goes on to reoffend than to incarcerate them. To model these considerations we define utility $Y \equiv (1 - A) \times \{-\theta R + (1 - R)\} - A$. The utility function is thus parameterized by θ, which quantifies how much "worse" is the case where individuals are released and reoffend as compared with the other two possibilities which are treated symmetrically. We emphasize that this utility function is a heuristic we use to illustrate our optimal policy learning method, and that a realistic utility function would be much more complicated (possibly depending also on factors not recorded in the available data).

We apply our proposed Q-learning procedure to optimize $\mathbb{E}[Y]$, assuming $K = 1$ and exogenous S. The fair policy constrains $S \to A$ and $S \to Y$ pathways; see Appendix V for details of our implementation as well as additional results. The proportion of individuals incarcerated ($A = 1$) is a function of θ, plotted in Fig. 3-4 and stratified by racial group. See the supplement for results on *overall* incerceration rates, which also vary among the policies. The region of particular interest is between

$\theta = 2$ and 3, where fair and unrestricted optimal policies differ and both recommend lower-than-observed overall incarceration rates; see the supplement. For most θ values, the fair policy recommends a decision rule which narrows the racial gap in incarceration rates as compared with the unrestricted policy, though does not eliminate this gap entirely. (Constraining the causal effects of race through mediator M would go further in eliminating this gap.) In regions where $\theta > 3$, both optimal policies in fact recommend higher-than-observed overall incarceration rates but a narrower racial gap, particularly for the fair policy. Comparing fair and unconstrained policy learning on this data serves to simultaneously illustrate how the proposed methods can be applied to real problems and how the choice of utility function is not innocuous.

3.3 Conclusions

Algorithms are opinions embedded in code. Despite their illusion of objectivity, they make use of subjective judgements of human beings at every step of their development, from data collection and naming of variables, to the way algorithms are trained and their output presented to decision makers. The goal of algorithmic fairness is to build algorithms that minimize the potential harm that they may place on underrepresented minorities. This entails devising algorithms that are sensitive to different sources of bias, tackling and removing these biases in the training step, and realizing the limitations and generalizations of what we create.

In this chapter, we considered the problem of fair statistical inference in two settings: fair predictions and fair policies, where we wish to minimize unfairness with respect to a particular sensitive feature, such as race or gender. We formalized the presence of unfairness as the presence of a certain *path-specific effect (PSE)* [89, 90], and framed the problem as one where we maximize the likelihood subject to constraints that restrict the magnitude of the PSE. We explored the implications of this view for cases where the PSE of interest is not identified, and for computational Bayesian

methods. We illustrated our approach using experiments on real datasets.

One of the advantages of our approach is it can be readily extended to concepts like affirmative action and "the wage gap." For instance, to conceptualize affirmative action, we propose to define a set of "valid paths" from S (e.g., race/sexual orientation) to Y (e.g., admission decision), perhaps paths through academic merit, or extracurriculars, or even the direct path, and solve a constrained optimization problem that *increases* the PSE along these paths. Here we mean placing a lower bound ϵ_l on the PSE away from the value corresponding to "no effect". Then, we learn p^* as the KL-closest distribution to the observed data distribution p that satisfies the constraint on the PSE. Finally, we predict the admission decision of a new instance Z in a similar way as the proposal in this chapter, by using the information in the new instance Z shared between p and p^*, and predicting/averaging over other information using p^*. We thus "count the causal influence of the sensitive feature on admission via prescribed paths" more highly among disadvantaged minorities. Defining these paths is a domain-specific issue. Increasing the PSE potentially lowers predictive performance, just as decreasing the PSE did in our experiments on reducing unfair biases. This makes sense since we are moving away from the PSE implied by the "unfair world" given by the MLE towards something else that we deem more "fair". A similar definition can be made for "the wage gap", which we believe should be meaningfully defined as a comparison of the PSE of gender on salary with respect to "inappropriate paths."

One methodological difficulty with our approach is the need for a computationally challenging constrained optimization problem. We discuss an alternative to reparameterize the observed data likelihood to include the causal parameter corresponding to the unfair PSE, in a way causal parameters have been added to the likelihood in structural nested mean models [60]. Under such a reparameterization, minimizing the PSE always corresponds to imposing box constraints on the likelihood [79].

Furthermore, we have extended a formalization of algorithmic fairness from [78] to

the setting of learning optimal policies under fairness constraints. We demonstrated how to constrain a set of statistical models and learn a policy such that subsequent decision making given new observations from the "unfair world" induces high-quality outcomes while satisfying the specified fairness constraints in the induced joint distribution. In this sense, our approach can be said to "break the cycle of injustice" in decision-making. We investigated the performance of our proposals on synthetic and real data, where in the latter case we have supplemented the data with a heuristic utility function.

Chapter 4

Graphical Models of Missing Data

Missing data has the potential to affect analyses conducted in all fields of scientific study, including healthcare, economics, and the social sciences. Strategies to cope with missingness that depends only on the observed data, known as the missing at random (MAR) mechanism, are well-studied [113, 114, 115, 41]. However, the setting where missingness depends on covariates that may themselves be missing, known as the missing not at random (MNAR) mechanism, is substantially more difficult and under-studied [116, 117]. MNAR mechanisms are expected to occur quite often in practice, for example, in longitudinal studies with complex patterns of dropout and re-enrollment, or in studies where social stigma may prompt non-response to questions pertaining to drug-use, or sexual activity and orientation, in a way that depends on other imperfectly collected or censored covariates [118, 119, 120].

Previous work on MNAR models has proceeded by imposing a set of restrictions on the full data distribution (the target distribution and its missingness mechanism) that are sufficient to yield identification of the parameter of interest. While there exist MNAR models whose restrictions cannot be represented graphically [121], the restrictions posed in several popular MNAR models such as the permutation model [118], the block-sequential MAR model [122], the itemwise conditionally independent nonresponse (ICIN) model [123, 124], and those in [125, 126, 127, 128, 129, 130] are either explicitly graphical or can be interpreted as such.

In our earlier work [131], we considered the identifiability of the target distribution within the class of graphical models of missing data, and showed that the most general identification strategies, [129, 132, 131], retain a significant gap in that they fail to identify a wide class of identifiable distributions. We proposed a new algorithm that significantly narrowed the identifiability gap in existing methods.

In this chapter, we show that even our most general algorithm [131] still retains a significant gap in that there exist target distributions that are identified which the algorithm fails to identify. We then present what is, to our knowledge, the first completeness result for missing data models representable as DAGs – a necessary and sufficient graphical condition under which the full data distribution is identified as a function of the observed data distribution [133]. For any given field of study, such a characterization is one of the most powerful results that identification theory can offer, as it comes with the guarantee that if these conditions do not hold, the model is provably not identified.

We further generalize these graphical conditions to settings where some variables are not just missing, but completely unobserved. Such distributions are typically summarized using acyclic directed mixed graphs (ADMGs), describes in Chapter 1 and [29]. We prove, once again, that our graphical criteria are sound and complete for the identification of full laws that are Markov relative to a hidden variable DAG and the resulting summary ADMG. This new result allows us to address two of the most critical issues in practical data analyses simultaneously, those of missingness and unmeasured confounding [133].

Finally, in the course of proving our results on completeness, we show that the proposed graphical conditions also imply that all missing data models of directed acyclic graphs or acyclic directed mixed graphs that meet these conditions, are in fact sub-models of the MNAR models in [123, 124]. This simple, yet powerful result implies that the joint density of these models may be identified using an odds ratio

parameterization that also ensures congenial specification of various components of the likelihood [134, 135]. Our results serve as an important precondition for the development of score-based model selection methods for graphical models of missing data, as an alternative to the constraint-based approaches proposed in [136, 137, 138].

4.1 Missing Data Models

A missing data model is a set of distributions defined over a set of random variables $\{O, X^{(1)}, R, X\}$, where O denotes the set of variables that are always observed, $X^{(1)}$ denotes the set of variables that are potentially missing, R denotes the set of missingness indicators of the variables in $X^{(1)}$, and X denotes the set of the observed proxies of the variables in $X^{(1)}$. By definition missingness indicators are binary random variables; however, the state space of variables in $X^{(1)}$ and O are unrestricted. Given $X_i^{(1)} \in X^{(1)}$ and its corresponding missingness indicator $R_i \in R$, the observed proxy X_i is defined as $X_i \equiv X_i^{(1)}$ if $R_i = 1$, and $X_i =?$ if $R_i = 0$. Hence, $p(X \mid R, X^{(1)})$ is deterministically defined. We call the non-deterministic part of a missing data distribution, i.e, $p(O, X^{(1)}, R)$, the *full law*, and partition it into two pieces: the *target law* $p(O, X^{(1)})$ and the *missingness mechanism* $p(R \mid X^{(1)}, O)$. The censored version of the full law $p(O, R, X)$, that the analyst actually has access to is known as the *observed data distribution*.

Following the convention in [128], let $\mathcal{G}(V)$ be a missing data DAG, where $V = \{O \cup X^{(1)} \cup R \cup X\}$. In addition to acyclicity, edges of a missing data DAG are subject to other restrictions: outgoing edges from variables in R cannot point to variables in $\{X^{(1)}, O\}$, each $X_i \in X$ has only two parents in $\mathcal{G}$, i.e., R_i and $X_i^{(1)}$ (these edges represent the deterministic function above that defines X_i, and are shown in gray in all the figures below), and there are no outgoing edges from X_i (i.e., the proxy X_i does not cause any variable on the DAG, however the corresponding full data variable $X_i^{(1)}$ may cause other variables.) A missing data model associated with a missing data

DAG $\mathcal{G}$ is the set of distributions $p(O, X^{(1)}, R, X)$ that factorizes as,

$$\prod_{V_i \in O \cup X^{(1)} \cup R} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)) \prod_{X_i \in X} p(X_i \mid X_i^{(1)}, R_i).$$

By standard results on DAG models, conditional independences in $p(X^{(1)}, O, R)$ can still be read off from $\mathcal{G}$ by the d-separation criterion [2]. For convenience, we will drop the deterministic terms of the form $p(X_i \mid X_i^{(1)}, R_i)$ from the identification analyses in the following sections since these terms are always identified by construction.

As an extension, we also consider a hidden variable DAG $\mathcal{G}(V \cup U)$, where $V = \{O, X^{(1)}, R, X\}$ and variables in U are unobserved, to encode missing data models in the presence of unmeasured confounders. In such cases, the full law would obey the nested Markov factorization [29] with respect to a missing data ADMG $\mathcal{G}(V)$, obtained by applying the latent projection operator [28] to the hidden variable DAG $\mathcal{G}(V \cup U)$. As a result of marginalization of latents U, there might exist bi-directed edges (to encode the hidden common causes) between variables in V (bi-directed edges are shown in red in all the figures below). It is straightforward to see that a missing data ADMG obtained via projection of a hidden variable missing data DAG follows the exact same restrictions as stated in the previous paragraph (i.e., no directed cycles, $\text{pa}_{\mathcal{G}}(X_i) = \{X_i^{(1)}, R_i\}$, every $X_i \in X$ is childless, and there are no outgoing edges from R_i to any variables in $\{X^{(1)}, O\}$.)

Identification in Missing Data Models

The goal of non-parametric identification in missing data models is twofold: identification of the target law $p(O, X^{(1)})$ or functions of it $f(p(O, X^{(1)}))$, and identification of the full law $p(O, X^{(1)}, R)$, in terms of the observed data distribution $p(O, R, X)$.

A compelling reason to study the problem of identification of the full law in and of itself, is due to the fact that many popular methods for model selection or causal discovery, rely on the specification of a well-defined and congenial joint

distribution [139, 140, 141]. A complete theory of the characterization of missing data full laws that are identified opens up the possibility of adapting such methods to settings involving non-ignorable missingness, in order to learn not only substantive relationships between variables of interest in the target distribution, but also the processes that drive their missingness. This is in contrast to previous approaches to model selection under missing data that are restricted to submodels of a single fixed identified model [136, 137, 138]. Such an assumption may be impractical in complex healthcare settings, for example, where discovering the factors that lead to missingness or study-dropout may be just as important as discovering substantive relations in the underlying data.

Though the focus of this chapter is on identification of the full law of missing data models that can be represented by a DAG (or a hidden variable DAG), some of our results naturally extend to identification of the target law (and functionals therein) due to the fact that the target law can be derived from the full law as $\sum_R p(O, X^{(1)}, R)$.

Remark 1. *By chain rule of probability, the target law $p(O, X^{(1)})$ is identified if and only if $p(R = 1 \mid O, X^{(1)})$ is identified. The identifying functional is given by*

$$p(O, X^{(1)}) = \frac{p(O, X^{(1)}, R = 1)}{p(R = 1 \mid O, X^{(1)})}.$$

(the numerator is a function of observed data by noting that $X^{(1)} = X$, and is observed when $R = 1$).

Remark 2. *The full law $p(O, X^{(1)}, R)$ is identified if and only if $p(R \mid O, X^{(1)})$ is identified. According to Remark 1, the identifying functional is given by*

$$p(O, X^{(1)}, R) = \frac{p(O, X^{(1)}, R = 1)}{p(R = 1 \mid O, X^{(1)})} \times p(R \mid O, X^{(1)}).$$

The rest of the chapter is organized as follows. In Section 4.2, we explain, through examples, why none of the existing identification algorithms put forward in the literature are *complete* in the sense that there exist missing data DAGs whose full law and target law are identified but these algorithms fail to derive an identifying functional for them. In Section 4.3, we provide a complete algorithm for full law identification. In Section 4.4, we further extend our identification results to models where unmeasured confounders are present. We defer all proofs to Appendix VI.

4.2 Incompleteness of Current Identification Methods

In this section, we show that even the most general methods proposed for identification in missing data DAG models remain *incomplete*. In other words, we show that there exist *identified* MNAR models that are representable by DAGs, however all existing algorithms fail to identify both the full and target law for these models. For brevity, we use our procedure proposed in [131] as an exemplar. However, as it is the most general procedure in the current literature, failure to identify via this procedure would imply failure by all other existing ones. For each example, we also provide alternate arguments for identification that eventually lead to the general theory.

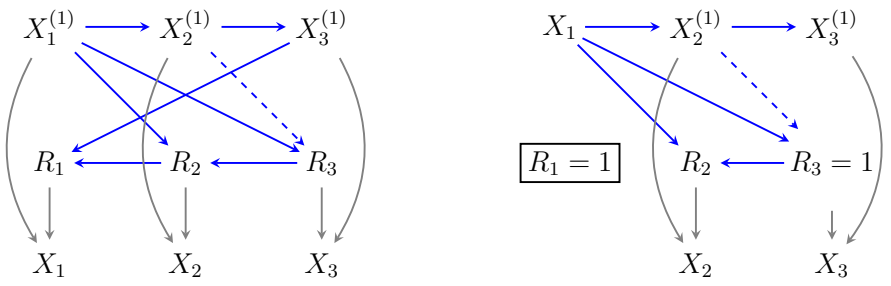
Our algorithm proposed in [131] proceeds as follows. For each missingness indicator R_i, the algorithm tries to identify the distribution $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))|_{R=1}$, sometimes referred to as the *propensity score* of R_i. It does so by checking if R_i is conditionally independent (given its parents) of the corresponding missingness indicators of its parents that are potentially missing. If this is the case, the propensity score is identified by a simple conditional independence argument (d-separation). Otherwise, the algorithm checks if this condition holds in post-fixing distributions obtained through recursive application of the *fixing* operator, which roughly corresponds to inverse weighting the current distribution by the propensity score of the variable

being fixed [29] (a more formal definition is provided in Appendix I.) If the algorithm succeeds in identifying the propensity score for each missingness indicator in this manner, then it succeeds in identifying the target law as Remark 1 suggests, since $p(R = 1 \mid O, X^{(1)}) = \prod_{R_i \in R} p(R_i \mid \text{pa}_{\mathcal{G}}(R))|_{R=1}$. Additionally, if it is the case that in the course of execution, the propensity score $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))$ for each missingness indicator is also identified at all levels of its parents, then the algorithm also succeeds in identifying the full law (due to Remark 2).

In order to ground our theory in reality, we now describe a series of hypotheses that may arise during the course of a data analysis that seeks to study the link between the effects of smoking on bronchitis, through the deposition of tar or other particulate matter in the lungs. For each hypothesis, we ask if the investigator is able to evaluate the goodness of fit of the proposed model, typically expressed as a function of the full data likelihood, as a function of just the observed data. In other words, we ask if the full law is identified as a function of the observed data distribution. If it is, this enables the analyst to compare and contrast different hypotheses and select one that fits the data the best.

Setup. To start, the investigator consults a large observational database containing the smoking habits, measurements of particulate matter in the lungs, and results of diagnostic tests for bronchitis on individuals across a city. She notices however, that several entries in the database are missing. This leads her to propose a model like the one shown in Fig. 4-1(a), where $X_1^{(1)}, X_2^{(1)}$, and $X_3^{(1)}$ correspond to smoking, particulate matter, and bronchitis respectively, and R_1, R_2, and R_3 are the corresponding missingness indicators.

For the target distribution $p(X^{(1)})$, she proposes a simple mechanism that smoking leads to increased deposits of tar in the lungs, which in turn leads to bronchitis ($X_1^{(1)} \to X_2^{(1)} \to X_3^{(1)}$). For the missingness process, she proposes that a suspected diagnosis of bronchitis is likely to lead to an inquiry about the smoking status of



(a) $\mathcal{G}_a$ (b) $\mathcal{G}_b \coloneqq \mathcal{G}_a(V \setminus R_1)$

Figure 4-1: (a) The missing data DAG used in scenario 1 (without the dashed edge $X_2^{(1)} \to R_3$) and scenario 2 (with the dashed edge $X_2^{(1)} \to R_3$) (b) Conditional DAG corresponding to the missing data DAG in (a) after fixing R_1, i.e., inverse weighting by the propensity score of R_1.

the patient ($X_3^{(1)} \to R_1$), smokers are more likely to get tested for tar and bronchitis ($X_1^{(1)} \to R_2, X_1^{(1)} \to R_3$), and ordering a diagnostic test for bronchitis, increases the likelihood of ordering a test for tar, which in turn increases the likelihood of inquiry about smoking status ($R_1 \leftarrow R_2 \leftarrow R_3$).

We now show that for this preliminary hypothesis, if the investigator were to utilize the procedure described in [131] she may conclude that it is not possible to identify the full law. We go on to show that such a conclusion would be incorrect, as the full law is, in fact, identified, and provide an alternative means of identification.

Scenario 1. Consider the missing data DAG model in Fig. 4-1(a) by excluding the edge $X_2^{(1)} \to R_3$, corresponding to the first hypothesis put forth by the investigator. The propensity score for R_1 can be obtained by simple conditioning, noting that $R_1 \perp\!\!\!\perp R_3 \mid X_3^{(1)}, R_2$ by d-separation. Hence, $p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) = p(R_1 \mid X_3^{(1)}, R_2) = p(R_1 \mid X_3, R_2, R_3 = 1)$.

Conditioning is not sufficient in order to identify the propensity score for R_2, as $R_2 \not\perp\!\!\!\perp R_1 \mid X_1^{(1)}, R_3$. However, it can be shown that in the distribution $q(V \setminus R_1 \mid R_1 = 1) \equiv \frac{p(V)}{p(R_1=1 \mid \mathrm{pa}_{\mathcal{G}}(R_1))}$, $R_2 \perp\!\!\!\perp R_1 \mid X_1, R_3 = 1$, since this distribution is Markov relative

to the graph in Fig. 4-1(b) (see the Appendix for details). We use the notation $q(\cdot \mid \cdot)$ to indicate that while q acts in most respects as a conditional distribution, it was not obtained from $p(V)$ by a conditioning operation. This implies that the propensity score for R_2 (evaluated at $R = 1$) is identified as $q(R_2 \mid X_1, R_3 = 1, R_1 = 1)$.

Finally, we show that the algorithm in [131] is unable to identify the propensity score for R_3. We first note that $R_3 \not\perp\!\!\!\perp R_1 \mid X_1^{(1)}$ in the original problem. Furthermore, as shown in Fig. 4-1(b), fixing R_1 leads to a distribution where R_3 is necessarily selected on as the propensity score $p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1))$ is identified by restricting the data to cases where $R_3 = 1$. It is thus impossible to identify the propensity score for R_3 in this post-fixing distribution. The same holds if we try to fix R_2 as identification of the propensity score for R_2 required us to first fix R_1, which we have seen introduces selection bias on R_3.

Hence, the procedure in [131] fails to identify both the target law and the full law for the problem posed in Fig. 4-1(a). However, both these distributions are, in fact, identified as we now demonstrate.

A key observation is that even though the identification of $p(R_3 \mid X_1^{(1)})$ might not be so straightforward, $p(R_3 \mid X_1^{(1)}, R_2)$ is indeed identified, because by d-separation $R_3 \perp\!\!\!\perp R_1 \mid X_1^{(1)}, R_2$, and therefore $p(R_3 \mid X_1^{(1)}, R_2) = p(R_3 \mid X_1, R_2, R_1 = 1)$. Given that $p(R_3 \mid X_1^{(1)}, R_2)$ and $p(R_2 \mid X_1^{(1)}, R_3 = 1)$ are both identified (the latter is obtained through as described earlier), we consider exploiting an odds ratio parameterization of the joint density $p(R_2, R_3 \mid \mathrm{pa}_{\mathcal{G}}(R_2, R_3)) = p(R_2, R_3 \mid X_1^{(1)})$. As we show below, such a parameterization immediately implies the identifiability of this density and consequently, the individual propensity scores for R_2 and R_3.

Given disjoint sets of variables A, B, C and reference values $A = a_0, B = b_0$, the odds ratio parameterization of $p(A, B \mid C)$, given in [134], is as follows:

$$\frac{1}{Z} \times p(A \mid b_0, C) \times p(B \mid a_0, C) \times \mathrm{OR}(A, B \mid C), \tag{4.1}$$

where

$$\text{OR}(A = a, B = b \mid C) = \frac{p(A = a \mid B = b, C)}{p(A = a_0 \mid B = b, C)} \times \frac{p(A = a_0 \mid B = b_0, C)}{p(A = a \mid B = b_0, C)},$$

and Z is the normalizing term and is equal to

$$\sum_{A,B} p(A \mid B = b_0, C) \times p(B \mid A = a_0, C) \times \text{OR}(A, B \mid C).$$

Note that $\text{OR}(A, B \mid C) = \text{OR}(B, A \mid C)$, i.e., the odds ratio is symmetric; see [134].

A convenient choice of reference value for the odds ratio in missing data problems is the value $R_i = 1$. Given this reference level and the parameterization of the joint in Eq. (4.1), we know that $p(R_2, R_3 \mid X_1^{(1)}) = \frac{1}{Z} \times p(R_2 \mid R_3 = 1, X_1^{(1)}) \times p(R_3 \mid R_2 = 1, X_1^{(1)}) \times \text{OR}(R_2, R_3 \mid X_1^{(1)})$, where Z is the normalizing term, and

$$\text{OR}(R_2 = r_2, R_3 = r_3 \mid X_1^{(1)}) = \frac{p(R_3 = r_3 \mid R_2 = r_2, X_1^{(1)})}{p(R_3 = 1 \mid R_2 = r_2, X_1^{(1)})} \times \frac{p(R_3 = 1 \mid R_2 = 1, X_1^{(1)})}{p(R_3 = r_3 \mid R_2 = 1, X_1^{(1)})}.$$

The conditional pieces $p(R_2 \mid R_3 = 1, X_1^{(1)})$ and $p(R_3 \mid R_2 = 1, X_1^{(1)})$ are already shown to be functions of the observed data. To see that the odds ratio is also a function of observables, recall that $R_3 \perp\!\!\!\perp R_1 \mid R_2, X_1^{(1)}$. This means that $R_1 = 1$ can be introduced into each individual piece of the odds ratio functional above, making it so that the entire functional depends only on observed quantities. Since all pieces of the odds ratio parameterization are identified, we can conclude that $p(R_2, R_3 \mid X_1^{(1)})$ is identified as the normalizing term is always identified if all the conditional pieces and the odds ratio are identified. This result, in addition to the fact that $p(R_1 \mid R_2, X_3^{(1)})$ is identified as before, leads us to the identification of both the target law and the full law, as the missingness process $p(R \mid X^{(1)})$ is identified.

Scenario 2. Suppose the investigator is interested in testing an alternate hypothesis to see whether detecting high levels of particulate matter in the lungs, also serves as an indicator to physicians that a diagnostic test for bronchitis should be ordered. This corresponds to the missing data DAG model in Fig. 4-1(a) by including the edge

$X_2^{(1)} \to R_3$. Since this is a strict super model of the previous example, the procedure in [131] still fails to identify the target and full laws in a similar manner as before.

However, it is still the case that both the target and full laws are identified. The justification for why the odds ratio parameterization of $p(R_2, R_3 \mid \text{pa}_{\mathcal{G}}(R_2, R_3)) = p(R_2, R_3 \mid X_1^{(1)}, X_2^{(1)})$ is identified in this scenario, is more subtle. We have,

$$p(R_2, R_3 \mid X_1^{(1)}, X_2^{(1)}) = \frac{1}{Z} \times p(R_2 \mid R_3 = 1, X_1^{(1)}, X_2^{(1)}) \times p(R_3 \mid R_2 = 1, X_1^{(1)}, X_2^{(1)}) \times \text{OR}(R_2, R_3 \mid X_1^{(1)}, X_2^{(1)}).$$

Note that $R_2 \perp\!\!\!\perp X_2^{(1)} \mid R_3, X_1^{(1)}$, and $R_3 \perp\!\!\!\perp R_1 \mid R_2, X_1^{(1)}, X_2^{(1)}$. Therefore, $p(R_2 \mid R_3 = 1, X_1^{(1)}, X_2^{(1)}) = p(R_2 \mid R_3 = 1, X_1^{(1)})$ is identified the same way as described in Scenario 1, and $p(R_3 \mid R_2 = 1, X_1^{(1)}, X_2^{(1)}) = p(R_3 \mid R_1 = 1, R_2 = 1, X_1, X_2)$ is a function of the observed data and hence is identified. Now the identification of $p(R_2, R_3 \mid X_1^{(1)}, X_2^{(1)})$ boils down to identifiability of the odds ratio term. By symmetry, we can express the odds ratio $\text{OR}(R_2, R_3 \mid X_1^{(1)}, X_2^{(1)})$ in two different ways,

$$\begin{aligned}\text{OR}(R_2, R_3 \mid X_1^{(1)}, X_2^{(1)}) &= \frac{p(R_2 \mid R_3, X_1^{(1)})}{p(R_2 = 1 \mid R_3, X_1^{(1)})} \times \frac{p(R_2 = 1 \mid R_3 = 1, X_1^{(1)})}{p(R_2 \mid R_3 = 1, X_1^{(1)})} \\ &= \frac{p(R_3 \mid R_2, X_1^{(1)}, X_2^{(1)})}{p(R_3 = 1 \mid R_2, X_1^{(1)}, X_2^{(1)})} \times \frac{p(R_3 = 1 \mid R_2 = 1, X_1^{(1)}, X_2^{(1)})}{p(R_3 \mid R_2 = 1, X_1^{(1)}, X_2^{(1)})}.\end{aligned}$$

The first equality holds by d-separation ($R_2 \perp\!\!\!\perp X_2^{(1)} \mid R_3, X_1^{(1)}$). This implies that $\text{OR}(R_2, R_3 \mid X_1^{(1)}, X_2^{(1)})$ is not a function of $X_2^{(1)}$. Let us denote this functional by $f_1(R_2, R_3, X_1^{(1)})$. On the other hand, we can plug-in $R_1 = 1$ to pieces in the second equality since $R_3 \perp\!\!\!\perp R_1 \mid R_2, X_1^{(1)}, X_2^{(1)}$ (by d-separation.) This implies that $\text{OR}(R_2, R_3 \mid X_1^{(1)}, X_2^{(1)})$ is a function of $X_1^{(1)}$ only through its observed values (i.e. X_1). Let us denote this functional by $f_2(R_2, R_3, X_1, X_2^{(1)}, R_1 = 1)$. Since odds ratio is symmetric (by definition), then it must be the case that $f_1(R_2, R_3, X_1^{(1)}) = f_2(R_2, R_3, X_1, X_2^{(1)}, R_1 = 1)$; concluding that f_2 cannot be a function of $X_2^{(1)}$, as the

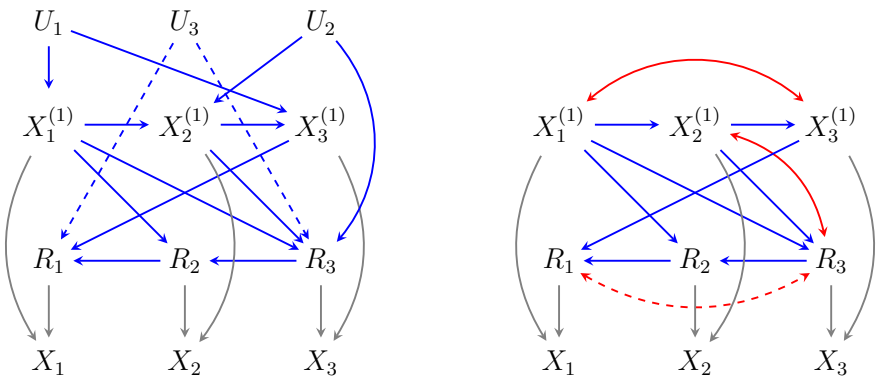
left hand side of the equation does not depend on $X_2^{(1)}$. This renders f_2 to be a function of only observed quantities, i.e. $f_2 = f_2(R_2, R_3, X_1, R_1 = 1)$. This leads to the conclusion that $p(R_2, R_3 \mid X_1^{(1)}, X_2^{(1)})$ is identified and consequently the missingness process $p(R \mid X^{(1)})$ in Fig. 4-1(a) is identified. According to Remarks 1 and 2, both the target and full laws are identified.

Adding any directed edge to Fig. 4-1(a) (including the dashed edge) allowed by missing data DAGs results in either a *self-censoring* edge ($X_i^{(1)} \to R_i$) or a special kind of collider structure called the *colluder* ($X_j^{(1)} \to R_i \leftarrow R_j$), that we first defined in [131]. We discuss in detail, the link between identification of missing data models of a DAG and the absence of these structures in Section 4.3.

Scenario 3. So far, the investigator has conducted preliminary analyses of the problem while ignoring the issue of unmeasured confounding. In order to address this issue, she first posits an unmeasured confounder U_1, corresponding to genotypic traits that may predispose certain individuals to both smoke and develop bronchitis. She posits another unmeasured confounder U_2, corresponding to the occupation of an individual, that may affect both the deposits of tar found in their lungs (for e.g., construction workers may accumulate more tar than an accountant due to occupational hazards) as well as limit an individual's access to proper healthcare, leading to the absence of a diagnostic test for bronchitis.

The missing data DAG with unmeasured confounders, corresponding to the aforementioned hypothesis is shown in Fig. 4-2(a) (excluding the dashed edges). The corresponding missing data ADMG, obtained by latent projection is shown in Fig. 4-2(b) (excluding the dashed bidirected edge). A procedure to identify the full law of such an MNAR model, that is nested Markov with respect to a missing data ADMG, is absent from the current literature. The question that arises, is whether it is possible to adapt the odds ratio parameterization from the previous scenarios, to this setting.

We first note that by application of the chain rule of probability and Markov



(a) $\mathcal{G}(V, U)$ (b) $\mathcal{G}(V)$

Figure 4-2: (a) The missing data DAG with unobserved confounders used in scenario 3 (without the dashed edges) and scenario 4 (with the dashed edges). (b) The corresponding missing data ADMGs obtained by applying the latent projection rules to the hidden variable DAG in (a).

restrictions, the missingness mechanism still factorizes in the same way as in Scenario 2, i.e., $p(R \mid X^{(1)}) = p(R_1 \mid R_2, X_3^{(1)}) \times p(R_2, R_3 \mid X_1^{(1)}, X_2^{(1)})$ [33]. Despite the addition of the bidirected edges $X_1^{(1)} \leftrightarrow X_3^{(1)}$ and $X_2^{(1)} \leftrightarrow R_3$, corresponding to unmeasured confounding, it is easy to see that the propensity score for R_1 is still identified via simple conditioning. That is, $p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1)) = p(R_1 \mid X_3, R_2, R_3 = 1)$ as $R_1 \perp\!\!\!\perp R_3 \mid X_3^{(1)}, R_2$ by m-separation. Furthermore, it can also be shown that the two key conditional independences that were exploited in the odds ratio parameterization of $p(R_2, R_3 \mid X^{(1)})$, still hold in the presence of these additional edges. In particular, $R_2 \perp\!\!\!\perp X_2^{(1)} \mid R_3, X_1^{(1)}$, and $R_3 \perp\!\!\!\perp R_1 \mid R_2, X_1^{(1)}, X_2^{(1)}$, by m-separation. Thus, the same odds ratio parameterization used for identification of the full law in Scenario 2, is also valid for Scenario 3. The full odds ratio parameterization of the MNAR models in Scenarios 2 and 3 is provided in Appendix VI.

Scenario 4. Finally, the investigator notices that a disproportionate number of missing entries for smoking status and diagnosis of bronchitis, correspond to individuals from certain neighborhoods in the city. She posits that such missingness

may be explained by systematic biases in the healthcare system, where certain ethnic minorities may not be treated with the same level of care. This corresponds to adding a third unmeasured confounder U_3, which affects the ordering of a diagnostic test for bronchitis as well as inquiry about smoking habits, as shown in Fig. 4-2(a) (including the dashed edges.) The corresponding missing data ADMG is shown in Fig. 4-2(b) (including the bidirected dashed edge.) Once again, we investigate if the full law is identified, in the presence of an additional unmeasured confounder U_3, and the corresponding bidirected edge $R_1 \leftrightarrow R_3$.

The missingness mechanism $p(R \mid X^{(1)})$ in Fig. 4-2(b) (including the dashed edge) no longer follows the same factorization as the one described in Scenarios 2 and 3, due to the presence of a direct connection between R_1 and R_3. According to [33], this factorization is given as $p(R \mid X^{(1)}) = p(R_1 \mid R_2, R_3, X_1^{(1)}, X_2^{(1)}, X_3^{(1)}) \times p(R_2 \mid R_3, X_1^{(1)}) \times p(R_3 \mid X_1^{(1)}, X_2^{(1)})$. Unlike the previous scenarios, the propensity score of R_1, $p(R_1 \mid R_2, R_3, X_1^{(1)}, X_2^{(1)}, X_3^{(1)})$, includes $X_1^{(1)}, X_2^{(1)}$, and R_3 past the conditioning bar. Thus, the propensity score of R_1 seems to be not identified, since there is no clear way of breaking down the dependency between R_1 and $X_1^{(1)}$. The problematic structure is the path $X_1^{(1)} \rightarrow R_3 \leftrightarrow R_1$ which contains a collider at R_3 that opens up when we condition on R_3 in the propensity score of R_1.

In light of the discussion in previous scenarios, another possibility for identifying $p(R \mid X^{(1)})$ is through analysis of the odds ratio parameterization of the entire missingness mechanism. In Section 4.3, we provide a description of the general odds ratio parameterization on an arbitrary number of missingness indicators. For brevity, we avoid re-writing the formula here. We simply point out that the first step in identifying the missingness mechanism via the odds ratio parameterization is arguing whether conditional densities of the form $p(R_i \mid R \setminus R_i = 1, X^{(1)})$ are identified, which is true if $R_i \perp\!\!\!\perp X_i^{(1)} \mid R \setminus R_i, X^{(1)} \setminus X_i^{(1)}$.

Such independencies do not hold in Fig. 4-2(b) (including the dashed edge) for any

of the Rs, since there exist collider paths between every pair $(X_i^{(1)}, R_i)$ that render the two variables dependent when we condition on everything outside $X_i^{(1)}, R_i$ (by m-separation). Examples of such paths are $X_1^{(1)} \to R_3 \leftrightarrow R_1$ and $X_2^{(1)} \leftrightarrow R_3 \leftrightarrow R_1 \leftarrow R_2$ and $X_3^{(1)} \to R_1 \leftrightarrow R_3$.

In Section 4.4, we show that the structures arising in the missing data ADMG presented in Fig. 4-2(b) (including the dashed edge), give rise to MNAR models that are provably not identified without further assumptions.

4.3 Full Law Identification in DAGs

[131] proved that two graphical structures, namely the self-censoring edge $(X_i^{(1)} \to R_i)$ and the colluder $(X_j^{(1)} \to R_i \leftarrow R_j)$, prevent the identification of full laws in missing data models of a DAG. In this section we exploit an odds ratio parameterization of the missing data process to prove that these two structures are, in fact, the *only* structures that prevent identification, thus yielding a complete characterization of identification for the full law in missing data DAG models.

We formally introduce the odds ratio parameterization of the missing data process introduced in [134], as a more general version of the simpler form mentioned earlier in Eq. (4.1). Assuming we have K missingness indicators, $p(R \mid X^{(1)}, O)$ can be expressed as follows.

$$p(R \mid X^{(1)}, O) = \frac{1}{Z} \times \prod_{k=1}^{K} p(R_k \mid R_{-k} = 1, X^{(1)}, O) \times \prod_{k=2}^{K} \mathrm{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X^{(1)}, O), \tag{4.2}$$

where $R_{-k} = R \setminus R_k, R_{\prec k} = \{R_1, \ldots, R_{k-1}\}, R_{\succ k} = \{R_{k+1}, \ldots, R_K\}$, and

$$\mathrm{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X^{(1)}, O) = \frac{p(R_k \mid R_{\succ k} = 1, R_{\prec k}, X^{(1)}, O)}{p(R_k = 1 \mid R_{\succ k} = 1, R_{\prec k}, X^{(1)}, O)} \times \frac{p(R_k = 1 \mid R_{-k} = 1, X^{(1)}, O)}{p(R_k \mid R_{-k} = 1, X^{(1)}, O)}.$$

Z in Eq. (4.2) is the normalizing term and is equal to $\sum_r \{\prod_{k=1}^K p(r_k \mid R_{-k} = 1, X^{(1)}, O) \times \prod_{k=2}^K \text{OR}(r_k, r_{\prec k} \mid R_{\succ k} = 1, X^{(1)}, O)\}$.

Using the odds ratio reparameterization given in Eq. (4.2), we now show that under a standard *positivity assumption*, stating that $p(R \mid X^{(1)}, O) > \delta > 0$, with probability one for some constant δ, the full law $p(R, X^{(1)}, O)$ of a missing data DAG is identified in the absence of self-censoring edges and colluders. Moreover, if any of these conditions are violated, the full law is no longer identified. We formalize this result below.

Theorem 7. *A full law $p(R, X^{(1)}, O)$ that is Markov relative to a missing data DAG $\mathcal{G}$ is identified if $\mathcal{G}$ does not contain edges of the form $X_i^{(1)} \to R_i$ (no self-censoring) and structures of the form $X_j^{(1)} \to R_i \leftarrow R_j$ (no colluders), and the stated positivity assumption holds. Moreover, the resulting identifying functional for the missingness mechanism $p(R \mid X^{(1)}, O)$ is given by the odds ratio parameterization provided in Eq. 4.2, and the identifying functionals for the target law and full law are given by Remarks 1 and 2.*

In what follows, we show that the identification theory that we have proposed for the full law in missing data models of a DAG is *sound* and *complete*. Soundness implies that when our procedure succeeds, the model is in fact identified, and the identifying functional is correct. Completeness implies that when our procedure fails, the model is *provably* not identified (non-parametrically). These two properties allow us to derive a precise boundary for what is and is not identified in the space of missing data models that can be represented by a DAG.

Theorem 8. *The graphical condition of no self-censoring and no colluders, put forward in Theorem 7, is sound and complete for the identification of full laws $p(R, O, X^{(1)})$ that are Markov relative to a missing data DAG $\mathcal{G}$.*

We now state an important result that draws a connection between missing data

models of a DAG $\mathcal{G}$ that are devoid of self-censoring and colluders, and the itemwise conditionally independent nonresponse (ICIN) model described in [123, 124]. As a substantive model, the ICIN model implies that no partially observed variable directly determines its own missingness, and is defined by the restrictions that for every pair $X_i^{(1)}, R_i$, it is the case that $X_i^{(1)} \perp\!\!\!\perp R_i \mid R_{-i}, X_{-i}^{(1)}, O$. We utilize this result in the course of proving Theorem 8.

Lemma 7. *A missing data model of a DAG $\mathcal{G}$ that contains no self-censoring edges and no colluders, is a submodel of the ICIN model.*

4.4 Full Law Identification in ADMGs

We now generalize identification theory of the full law to scenarios where some variables are not just missing, but completely unobserved, corresponding to the issues faced by the analyst in Scenarios 3 and 4 of Section 4.2. That is, we shift our focus to the identification of full data laws that are (nested) Markov with respect to a missing data ADMG $\mathcal{G}$.

Previously, we noted that the absence of colluders and self-censoring edges in a missing data DAG $\mathcal{G}$ imply a set of conditional independence restrictions of the form $X_i^{(1)} \perp\!\!\!\perp R_i \mid R_{-i}, X_{-i}^{(1)}, O$, for any pair $X_i^{(1)} \in X^{(1)}$ and $R_i \in R$. We now describe necessary and sufficient graphical conditions that must hold in a missing data ADMG $\mathcal{G}$ to imply this same set of conditional independences. Going forward, we ignore (without loss of generality), the deterministic factors $p(X \mid X^{(1)}, R)$, and the corresponding deterministic edges in $\mathcal{G}$, in the process of defining this graphical criterion.

A *colliding path* between two vertices A and B is a path on which every non-endpoint node is a collider. We adopt the convention that $A \to B$ and $A \leftrightarrow B$ are trivially collider paths. We say there exists a *colluding path* between the pair $(X_i^{(1)}, R_i)$ if $X_i^{(1)}$ and R_i are connected through at least one non-deterministic colliding path i.e.,

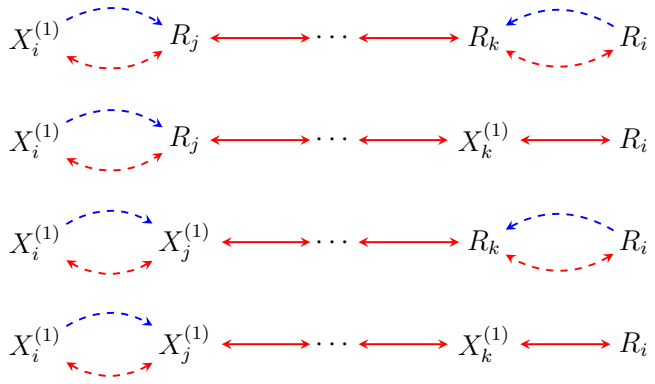


Figure 4-3: All possible colluding paths between $X_i^{(1)}$ and R_i. Each pair of dashed edges imply that the presence of either (or both) result in formation of a colluding path.

one which does not pass through (using deterministic edges) variables in X.

We enumerate all possible colluding paths between a vertex $X_i^{(1)}$ and its corresponding missingness indicator R_i in Fig. 4-3. Note that both the self-censoring structure and the colluding structure introduced in [131] are special cases of a colluding path. Using the m-separation criterion for ADMGs, it is possible to show that a missing data model of an ADMG $\mathcal{G}$ that contains no colluding paths of the form shown in Fig. 4-3, is also a submodel of the ICIN model in [123, 124].

Lemma 8. *A missing data model of an ADMG $\mathcal{G}$ that contains no colluding paths is a submodel of the ICIN model.*

This directly yields a sound criterion for identification of the full law of missing data models of an ADMG $\mathcal{G}$ using the odds ratio parameterization as before.

Theorem 9. *A full law $p(R, X^{(1)}, O)$ that is Markov relative to a missing data ADMG $\mathcal{G}$ is identified if $\mathcal{G}$ does not contain any colluding paths and the stated positivity assumption in Section 4.3 holds. Moreover, the resulting identifying functional for the missingness mechanism $p(R \mid X^{(1)}, O)$ is given by the odds ratio parametrization provided in Eq. 4.2.*

We now address the question as to whether there exist missing data ADMGs which contain colluding paths but whose full laws are nevertheless identified. We show (see Appendix for proofs), that the presence of a single colluding path of any of the forms shown in Fig. 4-3, results in a missing data ADMG $\mathcal{G}$ whose full law $p(X^{(1)}, R, O)$ cannot be identified as a function of the observed data distribution $p(X, R, O)$.

Lemma 9. *A full law $p(R, X^{(1)}, O)$ that is Markov relative to a missing data ADMG $\mathcal{G}$ containing a colluding path between any pair $X_i^{(1)} \in X^{(1)}$ and $R_i \in R$ is not identified.*

Revisiting our example in scenario 4, we note that every $(R_i, X_i^{(1)})$ pair is connected through at least one colluding path. Therefore, according to Lemma 9, the full law in Fig. 4-2(a) including the dashed edge, is not identified. It is worth emphasizing that the existence of at least one colluding path between any pair $(R_i, X_i^{(1)})$ is sufficient to conclude that the full law is not identified.

In what follows, we present a result on the soundness and completeness of our graphical condition that represents a powerful unification of non-parametric identification theory in the presence of non-ignorable missingness and unmeasured confounding. To our knowledge, such a result is the first of its kind. We present the theorem below.

Theorem 10. *The graphical condition of the absence of colluding paths, put forward in Theorem 9, is sound and complete for the identification of full laws $p(X^{(1)}, R, O)$ that are Markov relative to a missing data ADMG $\mathcal{G}$.*

Throughout this chapter, we have focused on identification of the full law which, according to Remark 1, directly yields identification for the target law. However, identification of the full law is a sufficient but not necessary condition for identification of the target law. In other words, the target law may still be identified despite the presence of colluding paths. Fig. 4(a) in [131] is an example of such a case.

4.5 Conclusions

In this chapter, we closed an important open problem in the non-parametric identification theory of missing data models represented via directed acyclic graphs, possibly in the presence of unmeasured confounders. We provided a simple graphical condition to check if the full law, Markov relative to a (hidden variable) missing data DAG, is identified. We further proved that these criteria are *sound* and *complete*. Moreover, we provided an identifying functional for the missingness process, through an odds ratio parameterization that allows for congenial specification of components of the likelihood. Our results serve as an important precondition for the development of score-based model selection methods that consider a broader class of missing data distributions than the ones considered in prior works. An interesting avenue for future work is exploration of the estimation theory of functionals derived from the identified full data law. To conclude, we note that while identification of the full law is sufficient to identify the target law, there exist identified target laws where the corresponding full law is not identified. We leave a complete characterization of target law identification to future work.

Chapter 5

Discussions and Conclusions

Making valid causal and statistical inferences is complicated by many types of biases in data. The aim of this thesis is to provide useful tools to mitigate some of these biases in our data analyses. Examples of biases that we considered and discussed include confounding bias induced by common causes of observed exposures and outcomes, bias in estimation induced by high dimensional data and curse of dimensionality, discriminatory bias encoded in data that reflect historical patterns of discrimination and inequality, and missing data bias where instantiations of variables are systematically missing. We used tools from statistics, optimization theory, and graphical models to understand and address these issues.

There are certain assumptions that enable us to tackle both identification and estimation problems in causal inference. For instance, the proposed complete identification algorithm in [36] assumes that the causal model is representable by a known graphical model. However, the causal model may not be known a priori. Under certain assumptions, the causal graph or a family of equivalent causal graphs can be identified from available data. This has been widely studied under the heading of structure learning (a.k.a. causal discovery in the literature). There is a rich literature on model selection from observational data in the context of causal inference [35]. This includes constraint-based algorithms such as PC [35, 142], score-based algorithms such as GES [143], and continuous optimization based algorithms such as the ones in [144, 145].

Evaluating cause-effect relationships provides us with aggregated population-level information on whether a certain treatment is effective or not. However, in order to account for inherent heterogeneity among individuals and optimize individual-level experiences, we might be interested in *personalized interventions*, where treatment is assigned according to a policy that takes into account the individual prior history; hence it is not fixed across individuals. For instance, personalized medicine aims at systematic use of individual patient history including biological information and biomarkers to improve patient's health care. Personalized actions can be viewed as realizations of decision rules where available information is mapped to the space of possible decisions. Making good personalized decisions often involves acting in multiple stages. For instance, multiple successive medical interventions may be required for long-term care of patients with chronic diseases. The goal of personalized medicine is to tailor a sequence of decision rules on treatment, known as dynamic treatment regimes or policies, based on patient characteristics seen so far, to maximize the likelihood of a desirable outcome. A number of algorithms have been developed for *estimating optimal treatment regimes* [102].

A natural step in causal inference is to understand the mechanisms by which the treatment influences the outcome. Understanding causal mechanisms may lead to designing better policies by optimizing a *part* of the effect of the treatment on the outcome. For example, we may wish to maximize the chemical effect of a drug given data from an observational study where the chemical effect of the drug on the outcome is entangled with the indirect effect mediated by differential adherence. In such cases, we may wish to optimize the direct effect of a drug, while keeping the indirect effect to that of some reference treatment. Policies of this type may be more directly relevant in precision medicine contexts where adherence varies among patients. In prior work, we derived a variety of methods for learning high quality policies of this type by combining tools from causal mediation analysis and reinforcement learning [146].

In classical causal inference, inferring cause-effect relations from data relies on the assumption that units are independent and identically distributed (iid). This assumption is often implausible and is violated in settings where units are related through a network of dependencies, known as interference. The most common example is that of infectious diseases where treatment of one individual may have a protective effect on others in the population. There is a growing literature on causal inference with interference and dependent data [147, 148, 149]

Despite the fascinating methodological advances in the field of causal inference over the past few decades, there still remain many open problems and exciting challenges in this research area. In future, we plan to pursue multiple directions to continue to provide solutions to open problems and bridge the gap between theory and scientific applications in healthcare, social justice, public policy, and social science.

Appendix I

Overview of Nested Markov Models

Here, we introduce the necessary graphical preliminaries to describe the *nested Markov factorization* of an ADMG that captures all equality constraints on the observed margin $p(V)$. Given a DAG $\mathcal{G}(V \cup U)$ where U contains variables that are unobserved, the *latent projection operator* onto the observed margin produces an acyclic directed mixed graph $\mathcal{G}(V)$ that consists of directed and bidirected edges [28]. The bidirected connected components of an ADMG $\mathcal{G}(V)$, partition the vertices V into distinct sets known as districts. The district membership of a vertex V_i in $\mathcal{G}$ is denoted $\text{dis}_{\mathcal{G}}(V_i)$, and the set of all districts in $\mathcal{G}$ is denoted $\mathcal{D}(\mathcal{G})$.

CADMGs and Kernels

The nested Markov factorization of $p(V)$ relative to an ADMG $\mathcal{G}(V)$ is defined with the use of conditional distributions known as *kernels* and their associated *conditional ADMGs* (CADMGs) that are derived from $p(V)$ and $\mathcal{G}(V)$ respectively, via repeated applications of the *fixing operator* [29]. A CADMG $\mathcal{G}(V, W)$, is an ADMG whose nodes can be partitioned into random variables V and *fixed* variables W, with the restriction that only outgoing edges may be adjacent to variables in W. A kernel $q_V(V \mid W)$ is a mapping from values of W to normalized densities over V. That is,

$\sum_V q_V(V \mid W = w) = 1, \forall w \in W$ [32]. For any set of variables $X \subseteq V$, marginalization and conditioning in a kernel are defined as follows.

$$q_{V\setminus X}(V \setminus X \mid W) \equiv \sum_X q_V(V \mid W), \text{ and}$$
$$q_V(V \setminus X \mid X, W) \equiv \frac{q_V(V \mid W)}{q_V(X \mid W)}.$$

The notation $q_V(\cdot \mid X)$ makes clear which variables appearing past the "conditioning" bar in a kernel are fixed as opposed to simply conditioned on. That is, if a variable $X_i \notin V$, then it is fixed, else it is conditioned on.

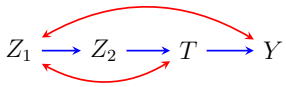
Fixing and Fixability

A variable $A \in V$ is said to be *fixable* if the paths $A \to \cdots \to X$ and $A \leftrightarrow \cdots \leftrightarrow X$ do not both exist for all $X \in V \setminus \{A\}$. Given a CADMG $\mathcal{G}(V, W)$ where A is fixable, the graphical operator of fixing, denoted $\phi_A(\mathcal{G})$, yields a new CADMG $\mathcal{G}(V \setminus A, W \cup A)$ with all incoming edges into A being removed, and A being set to a fixed value a. Given a kernel $q_V(V \mid W)$, the corresponding probabilistic operation of fixing, denoted $\phi_A(q_V; \mathcal{G})$ yields a new kernel

$$q_{V\setminus A}(V \setminus A \mid W \cup A) \equiv \frac{q_V(V \mid W)}{q_V(A \mid \mathrm{mb}_{\mathcal{G}}(A), W)},$$

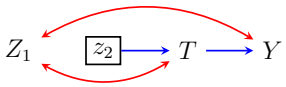
where $\mathrm{mb}_{\mathcal{G}}(A)$ is the *Markov blanket* of A, defined as the bidirected connected component (district) of A (excluding A itself) and the parents of the district of A, i.e., $\mathrm{mb}_{\mathcal{G}}(A) \equiv \mathrm{dis}_{\mathcal{G}}(A) \cup \mathrm{pa}_{\mathcal{G}}(\mathrm{dis}_{\mathcal{G}}(A)) \setminus \{A\}$. It is easy to check that when $\mathcal{G}$ is a DAG, i.e., there are no bidirected edges, the denominator in the probabilistic operation of fixing, reduces to the familiar definition of a simple propensity score.

The notion of fixability can be extended to a set of variables $S \subseteq V$ as follows. A set S is said to be fixable if elements in S can be ordered into a sequence $\sigma_S = \langle S_1, S_2, \dots \rangle$ such that S_1 is fixable in $\mathcal{G}$, S_2 is fixable in $\phi_{S_1}(\mathcal{G})$, and so on. This notion of fixability on sets of variables is essential to the description of the nested Markov model that we present in the following section.



(a) $\mathcal{G}$

$p(V) = p(Z_2 \mid Z_1) \times p(Z_1) \times p(T, Y \mid Z_1, Z_2)$



(b) $\phi_{Z_2}(\mathcal{G})$

$q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2) = p(Z_1) \times p(T, Y \mid Z_1, Z_2)$

(c) $\phi_{\{Z_1,Z_2\}}(\mathcal{G})$

$q_{T,Y}(T, Y \mid Z_1, Z_2) = \sum_{Z_1} p(Z_1) \times p(T, Y \mid Z_1, Z_2)$

(d) $\phi_{\{Z_1,Z_2,T\}}(\mathcal{G})$

$q_Y(Y \mid T, Z_1, Z_2) = \frac{\sum_{Z_1} p(Z_1) \times p(T,Y \mid Z_1,Z_2)}{\sum_{Z_1} p(Z_1) \times p(T \mid Z_1,Z_2)}$

Figure I-1: An example to illustrate fixing and kernel operations.

Occasionally, fixing operations may also simplify to marginalization or conditioning events. We illustrate these concepts with a simple example.

Example I.1. Consider the ADMG shown in Fig. I-1(a) and fix the kernel of interest to be $q_Y(Y \mid T, Z_1, Z_2)$, i.e., a kernel where all other variables except Y are fixed. A valid fixing sequence in order to obtain such a kernel from the joint $p(V)$ is (Z_2, Z_1, T). Fixing Z_2 entails dividing by the simple conditional $p(Z_2 \mid Z_1)$ and yields the CADMG $\phi_{Z_2}(\mathcal{G})$ and corresponding kernel $q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)$ shown in Fig. I-1(b). In order to fix Z_1, we must divide by the kernel $q_{Z_1,T,Y}(Z_1 \mid Z_2, T, Y)$. By rules of conditioning and marginalization in kernels,

$$q_{Z_1,T,Y}(Z_1 \mid Z_2, T, Y) \equiv \frac{q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)}{q_{Z_1,T,Y}(T, Y \mid Z_2)} \equiv \frac{q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)}{\sum_{Z_1} q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)}$$

Fixing Z_1 and evaluating the above expression gives us the CADMG and corresponding kernel shown in Fig. I-1(c). That is, fixing Z_1 in the kernel $q_{Z_1,T,Y}(Z_1 \mid Z_2, T, Y)$, simplifies to marginalization of Z_1. Finally, applying rules of conditioning and marginalization to the kernel $q_{T,Y}(T, Y \mid Z_1, Z_2)$ we can obtain the kernel $q_{T,Y}(T \mid Z_1, Z_2, Y)$. Dividing by this corresponds to fixing T, giving us the CADMG and desired kernel shown in Fig. I-1(d).

Nested Markov Factorization

Given a CADMG $\mathcal{G}$, A set $S \subseteq V$ is said to be *reachable* if there exists a valid sequence of fixing operations on vertices $V \setminus S$. Further, S is said to be *intrinsic* if it is reachable, and forms a single bidirected connected component or district in $\phi_{\sigma_{V\setminus S}}(\mathcal{G})$, i.e., the CADMG obtained upon executing all fixing operations given by a valid fixing sequence $\sigma_{V\setminus S}$.

A distribution $p(V)$ is said to obey the nested Markov factorization relative to an ADMG $\mathcal{G}(V)$ if for every fixable set S, and any valid fixing sequence σ_S,

$$\phi_{\sigma_S}(p(V);\mathcal{G}) = \prod_{D\in\mathcal{D}(\phi_{\sigma_S}(\mathcal{G}))} q_D(D \mid \mathrm{pa}_{\phi_{\sigma_S}(\mathcal{G})}(D)),$$

where all kernels appearing in the product above can be constructed by combining kernels corresponding to intrinsic sets i.e., $\{q_I(I \mid \mathrm{pa}_{\mathcal{G}}(I)) \mid I \text{ is intrinsic in } \mathcal{G}\}$. Such a construction is made possible by the fact that all the sets D quantified in the product are districts in a reachable graph derived from $\mathcal{G}$.

It was noted in [29] that when a distribution $p(V)$ is nested Markov relative to an ADMG $\mathcal{G}$, all valid fixing sequences yield the same CADMG and kernel so that recursive applications of the fixing operator on a set $V \setminus S$ can simply be denoted as $\phi_{V\setminus S}(\mathcal{G})$ and $\phi_{V\setminus S}(q_V;\mathcal{G})$ without explicitly specifying any particular valid order. Thus, the construction of the set of kernels corresponding to intrinsic sets can be characterized as $\{q_I(I \mid \mathrm{pa}_{\mathcal{G}}(I)) \mid I \text{ is intrinsic in } \mathcal{G}\} = \{\phi_{V\setminus I}(p(V;\mathcal{G})) \mid I \text{ is intrinsic in } \mathcal{G}\}$, and the nested Markov factorization can be re-stated more simply as, for every fixable set S we have,

$$\phi_S(p(V;\mathcal{G})) = \prod_{D\in\mathcal{D}\left(\phi_S(\mathcal{G})\right)} \phi_{V\setminus D}(p(V);\mathcal{G}),$$

An important result from [29] states that if $p(V \cup U)$ is Markov relative to a DAG $\mathcal{G}(V \cup U)$, then $p(V)$ is nested Markov relative to the ADMG $\mathcal{G}(V)$ obtained by latent projection.

Binary Parameterization of Nested Markov Models

From the above factorization, it is clear that intrinsic sets given their parents form the atomic units of the nested Markov model. Using this observation, a smooth parameterization of discrete nested Markov models was provided by [150]. We now provide a short description of how to derive the so-called Moebius parameters of a *binary* nested Markov model.

For each district $D \in \mathcal{D}(\mathcal{G})$, consider all possible subsets $S \subseteq D$. If S is intrinsic (that is, reachable and bidirected connected in $\phi_{V \setminus S}(\mathcal{G})$), define the head H of the intrinsic set to be all vertices in S that are childless in $\phi_{V \setminus S}(\mathcal{G})$, and the tail T to be all parents of the head in the CADMG $\phi_{V \setminus S}(\mathcal{G})$, excluding the head itself. More formally, $H \equiv \{V_i \in S \mid \mathrm{ch}_{\phi_{V \setminus S}(\mathcal{G})}(V_i) = \emptyset\}$, and $T \equiv \mathrm{pa}_{\phi_{V \setminus S}(\mathcal{G})}(H) \setminus H$. The corresponding set of Moebius parameters for this intrinsic head and tail pair parameterizes the kernel $q_S(H = 0 \mid T)$, i.e., the kernel where all variables outside the intrinsic set S are fixed, and all elements of the head are set to zero given the tail. Note that these parameters are, in general, *variationally dependent* (in contrast to variationally independent in the case of an ordinary DAG model) as the heads and tails in these parameter sets may overlap. The joint density for any query $p(V = v)$, can be obtained through the Moebius inversion formula; see [32, 150] for details. For brevity, we will denote $q_S(H = 0 \mid T)$ as simply $q(H = 0 \mid T)$, as it will be clear from the given context what variables are still random in the kernel corresponding to a given intrinsic set.

Appendix II

Overview of Semiparametric Theory

Assume a statistical model $\mathcal{M} = \{p_\eta(Z) : \eta \in \Gamma\}$ where Γ is the parameter space and η is the parameter indexing a specific model. We are often interested in a function $\psi : \eta \in \Gamma \mapsto \psi(\eta) \in \mathbb{R}$; i.e., a parameter that maps the distribution P_η to a scalar number in $\mathbb{R}$, such as an identified average causal effect. (For brevity, we sometimes use ψ instead of $\psi(\eta)$, which should be obvious from context.) Truth is denoted by P_{η_0} and ψ_0. An estimator $\widehat{\psi}_n$ of a scalar[1] parameter ψ based on n i.i.d copies $Z_1, \ldots, Z_n$ drawn from $p_\eta(Z)$, is *asymptotically linear* if there exists a measurable random function $U_\psi(Z)$ with mean zero and finite variance such that

$$\sqrt{n} \times (\widehat{\psi}_n - \psi) = \frac{1}{\sqrt{n}} \times \sum_{i=1}^{n} U_\psi(Z_i) + o_p(1), \tag{II.1}$$

where $o_p(1)$ is a term that converges in probability to zero as n goes to infinity. The random variable $U_\psi(Z)$ is called the *influence function* of the estimator $\widehat{\psi}_n$. The term influence function comes from the robustness literature [151].

Before mentioning the asymptotic properties of an asymptotically linear estimator, it is worth noting that in asymptotic theory, we can sometimes construct *super efficient* estimators, e.g. Hodges estimator, that have undesirable local properties

[1]Here, our focus is on estimation of $\psi = \mathbb{E}[Y(t)]$ which is a scalar parameter. For an extension to a vector valued functional in $\mathbb{R}^q, q > 1$, refer to [41, 38].

associated with them. Therefore, the analysis is oftentimes restricted to *regular*[2] and asymptotically linear (RAL) estimators to avoid such complications. Although most reasonable estimators are RAL, regular estimators do exist that are not asymptotically linear. However, as a consequence of [152] representation theorem, the most efficient regular estimator is asymptotically linear; hence, it is reasonable to restrict attention to RAL estimators. According to [153], the influence function of a RAL estimator is the same as the influence function of its estimand. Further, there is a bijective correspondence between RAL estimators and influence functions.

By a simple consequence of the central limit theorem and Slutsky's theorem, it is straightforward to show that the RAL estimator $\widehat{\psi}_n$ is *consistent and asymptotically normal* (CAN), with asymptotic variance equal to the variance of its influence function U_ψ,

$$\sqrt{n} \times (\widehat{\psi}_n - \psi) \xrightarrow{d} N\Big(0,\ \mathrm{var}(U_\psi)\Big). \tag{II.2}$$

The first step in dealing with a semiparametric model, is to consider a simpler finite-dimensional parametric submodel that is contained within the semiparametric model and it contains the truth. Consider a (regular) parametric submodel $\mathcal{M}_{\text{sub}} = \{P_{\eta_\kappa} : \kappa \in [0,1) \text{ where } P_{\eta_{\kappa=0}} = P_{\eta_0}\}$ of the model $\mathcal{M}$. Given P_{η_0}, define the corresponding score to be $S_{\eta_0}(Z) = \dfrac{d}{d\kappa} \log p_{\eta_\kappa}(Z)\Big|_{\kappa=0}$. It is known that

$$\frac{d}{d\kappa}\psi(\eta_\kappa)\Big|_{\kappa=0} = \mathbb{E}\Big[U_\psi(Z) \times S_{\eta_0}(Z)\Big], \tag{II.3}$$

where $\psi(\eta_\kappa)$ is the target parameter in the parametric submodel, $U_\psi(Z)$ is the corresponding influence function evaluated at law P_{η_0}, $S_{\eta_0}(Z)$ is the score of the law P_{η_0}, and the expectation is taken with respect to P_{η_0}. Equation II.3 provides an easy way to derive an influence function for the parameter ψ. In the next subsection, we use

[2]Given a collection of probability laws $\mathcal{M}$, an estimator $\widehat{\psi}$ of $\psi(P)$ is said to be regular in $\mathcal{M}$ at P if its convergence to $\psi(P)$ is locally uniform [39].

this equation to derive an influence function for our target $\psi = \mathbb{E}[Y(t)]$ and discuss its properties.

Influence functions provide a geometric view of the behavior of RAL estimators. Consider a Hilbert space[3] H of all mean-zero scalar functions, equipped with an inner product defined as $\mathbb{E}[h_1 \times h_2], h_1, h_2 \in \mathrm{H}$. The *tangent space* in the model $\mathcal{M}$, denoted by Λ, is defined to be the mean-square closure of parametric submodel tangent spaces, where a parametric submodel tangent space is the set of elements $\Lambda_{\eta_\kappa} = \{\alpha S_{\eta_\kappa}(Z)\}, \alpha$ is a constant and S_{η_κ} is the score for the parameter ψ_{η_κ} for some parametric submodel. In mathematical form, $\Lambda = \overline{[\Lambda_{\eta_\kappa}]}$.

The tangent space Λ is a closed linear subspace of the Hilbert space H ($\Lambda \subseteq \mathrm{H}$). The orthogonal complement of the tangent space, denoted by $\Lambda^\perp$, is defined as $\Lambda^\perp = \{h \in \mathrm{H} \mid \mathbb{E}[h \times h'] = 0, \forall h' \in \Lambda\}$. Note that $\mathrm{H} = \Lambda \oplus \Lambda^\perp$, where $\oplus$ is the direct sum, and $\Lambda \cap \Lambda^\perp = \{0\}$. Given an arbitrary element $h \in \Lambda^\perp$, it holds that for any submodel $\mathcal{M}_{\text{sub}}$, with score S_{η_0} corresponding to P_{η_0}, $\mathbb{E}[h \times S_{\eta_0}] = 0$. Consequently, using Eq. II.3, $h + U_\psi(Z)$ is also an influence function. The vector space $\Lambda^\perp$ is then of particular importance because we can now construct the class of all influence functions, denoted by $\mathcal{U}$, as $\mathcal{U} = U_\psi(Z) + \Lambda^\perp$. Upon knowing a single IF $U_\psi(Z)$ and the tangent space orthogonal complement $\Lambda^\perp$, we can obtain the class of all possible RAL estimators that admit the CAN property.

Out of all the influence functions in $\mathcal{U}$ there exists a unique one which lies in the tangent space Λ, and which yields the most efficient RAL estimator by recovering the *semiparametric efficiency bound*. This efficient influence function can be obtained by projecting any influence function, call it U^*_ψ, onto the tangent space Λ. This operation is denoted by $U^{\text{eff}_\psi} = \pi[U^*_\psi \mid \Lambda]$, where U^{eff}_ψ denotes the efficient IF.

On the other hand, if the tangent space contains the entire Hilbert space, i.e., $\Lambda =$

[3]The Hilbert space of all mean-zero scalar functions is the L^2 space. For a precise definition of Hilbert spaces see [154].

H, then the statistical model $\mathcal{M}$ is called a *nonparametric* model. In a nonparametric model, we only have one influence function since $\Lambda^{\perp} = \{0\}$. This unique influence function can be obtained via Eq. II.3 and corresponds to the efficient influence function U_{ψ}^{eff} (the unique element in the tangent space Λ) in the nonparametric model $\mathcal{M}$. For a detailed description of the concepts outlined here, please refer to [41, 38].

Appendix III

Supplementary Materials for Primal Fixability

Theorem 1 (Nonparametric influence function of augmented primal IPW)

Proof. The target parameter is identified via the following function of the observed data,

$$\psi_\kappa(t) = \sum_{V\setminus T} Y \times \prod_{M_i\in\mathbb{M}} p_\kappa(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \prod_{L_i\in\mathbb{L}} p_\kappa(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times p_\kappa(\mathbb{C}), \tag{III.1}$$

and according to Eq. II.3, $\frac{d}{d\kappa}\psi_\kappa(t)\Big|_{\kappa=0} = \mathbb{E}\Big[U_{\psi_t} \times S_{\eta_0}(V)\Big]$. Therefore,

$$\begin{aligned}
\frac{d}{d\kappa}\psi_\kappa(t) &= \frac{d}{d\kappa}\Big\{\sum_{V\setminus T} Y \times \prod_{M_i\in\mathbb{M}} p_\kappa(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \prod_{L_i\in\mathbb{L}\setminus T} p_\kappa(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times p_\kappa(T,\mathbb{C})\Big\} \\
&= \sum_{V\setminus T} Y \times \frac{d}{d\kappa}\Big\{\prod_{M_i\in\mathbb{M}} p_\kappa(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}\Big\} \times \sum_{T} \prod_{L_i\in\mathbb{L}\setminus T} p_\kappa(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times p_\kappa(T,\mathbb{C}) \quad \text{(1st Term)} \\
&+ \sum_{V\setminus T} Y \times \prod_{M_i\in\mathbb{M}} p_\kappa(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \frac{d}{d\kappa}\Big\{\prod_{L_i\in\mathbb{L}\setminus T} p_\kappa(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))\Big\} \times p_\kappa(T,\mathbb{C}) \quad \text{(2nd Term)} \\
&+ \sum_{V\setminus T} Y \times \prod_{M_i\in\mathbb{M}} p_\kappa(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \prod_{L_i\in\mathbb{L}\setminus T} p_\kappa(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \frac{d}{d\kappa}\Big\{p_\kappa(T,\mathbb{C})\Big\}. \quad \text{(3rd Term)}
\end{aligned}$$

First Term: The contribution of the first term to the final IF is made of individual contributions of the elements in $\mathbb{M}$. Since the derivation is similar, we only derive it for an element $M_j \in \mathbb{M}$.

$$
\begin{aligned}
&\sum_{V\setminus T} Y \times \prod_{M_i\in\{\prec M_j\}\cap\mathbb{M}} p_\kappa(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \frac{d}{d\kappa}\Big\{p_\kappa(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j))|_{T=t}\Big\} \\
&\qquad \times \prod_{M_i\in\{\succ M_j\}\cap\mathbb{M}} p_\kappa(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \prod_{L_i\in\mathbb{L}} p_\kappa(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times p_\kappa(\mathbb{C}) \\
&\overset{(1)}{=} \sum_{V\setminus\{T,\{\preceq M_j\}\}} \prod_{M_i\in\{\prec M_j\}\cap\mathbb{M}} p_\kappa(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \frac{d}{d\kappa}\Big\{p_\kappa(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j))|_{T=t}\Big\} \\
&\qquad \times \sum_{T\cup\{\succ M_j\}} Y \times \prod_{V_i\in\mathbb{L}\cup\{\{\succ M_j\}\cap\mathbb{M}\}} p_\kappa(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i\in\mathbb{M}} \times p_\kappa(\mathbb{C}) \\
&\overset{(2)}{=} \sum_{\preceq M_j} \frac{\mathbb{I}(T=t)}{\prod_{L_i\prec M_j} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times S(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j)) \times \prod_{V_i\in\{\preceq M_j\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \\
&\qquad \times \sum_{T\cup\{\succ M_j\}} Y \times \prod_{V_i\in\mathbb{L}\cup\{\{\succ M_j\}\cap\mathbb{M}\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i\in\mathbb{M}} \\
&\overset{(3)}{=} \mathbb{E}\Bigg[\frac{\mathbb{I}(T=t)}{\prod_{L_i\prec M_j} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \underbrace{\sum_{T\cup\{\succ M_j\}} Y \times \prod_{V_i\in\mathbb{L}\cup\{\{\succ M_j\}\cap\mathbb{M}\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i\in\mathbb{M}}}_{:=f(\preceq M_j)} \\
&\qquad\qquad \times S(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j))\Bigg] \\
&\overset{(4)}{=} \mathbb{E}\left[\frac{\mathbb{I}(T=t)}{\prod_{L_i\prec M_j} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \left(f(\preceq M_j) - \sum_{M_j} f(\preceq M_j) \times p(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j))\right) \times S(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j))\right] \\
&\overset{(5)}{=} \mathbb{E}\left[\frac{\mathbb{I}(T=t)}{\prod_{L_i\prec M_j} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \left(f(\preceq M_j) - \sum_{M_j} f(\preceq M_j) \times p(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j))\right) \times S(V)\right]
\end{aligned}
$$

The first equality follows from the fact that terms corresponding to $M_i \in \{\prec M_j\}$ are not functions of elements in $\{\succ M_j\}$ and of Y. The second equality follows by term grouping, the definition of conditional scores, and term cancellation. The third equality is by definition of joint expectation. The fourth and fifth equalities are implied by the fact that conditional scores have expected value of 0 (given their conditioning set). Therefore, the contribution of $M_j \in \mathbb{M}$ is the following:

$$
\begin{aligned}
&\frac{\mathbb{I}(T=t)}{\prod_{L_i\prec M_j} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \Bigg(\sum_{T\cup\{\succ M_j\}} Y \times \prod_{V_i\in\mathbb{L}\cup\{\{\succ M_j\}\cap\mathbb{M}\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i\in\mathbb{M}} \\
&\qquad\qquad - \sum_{T\cup\{\succeq M_j\}} Y \times \prod_{V_i\in\mathbb{L}\cup\{\{\succeq M_j\}\cap\mathbb{M}\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))\Big|_{T=t \text{ if } V_i\in\mathbb{M}}\Bigg).
\end{aligned}
$$

Second Term: The contribution of the second term to the final IF is made of

individual contributions of the elements in $\mathbb{L} \setminus T$. Since the derivation is similar, we only derive it for an element $L_j \in \mathbb{L} \setminus T$.

$$\begin{aligned}
&\sum_{V \setminus T} Y \times \prod_{M_i \in \mathbb{M}} p_\kappa(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \times \sum_{T} \Big\{ \prod_{L_i \in \{\prec L_j\} \cap \mathbb{L} \setminus T} p_\kappa(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \\
&\qquad \times \frac{d}{d\kappa} \Big\{ p_\kappa(L_j \mid \mathrm{mp}_{\mathcal{G}}(L_j)) \Big\} \times \prod_{L_i \in \{\succ L_j\} \cap \mathbb{L} \setminus T} p_\kappa(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \Big\} \times p_\kappa(T, \mathbb{C}) \\
&\overset{(1)}{=} \sum_{V} Y \times \prod_{V_i \in \{\succ L_j\}} p_\kappa(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \Big|_{T=t \text{ if } V_i \in \mathbb{M}} \times \frac{d}{d\kappa} \Big\{ p_\kappa(L_j \mid \mathrm{mp}_{\mathcal{G}}(L_j)) \Big\} \\
&\qquad \times \prod_{V_i \in \{\prec L_j\}} p_\kappa(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \Big|_{T=t \text{ if } V_i \in \mathbb{M}} \\
&\overset{(2)}{=} \sum_{\preceq L_j} \underbrace{\sum_{\succ L_j} Y \times \prod_{V_i \in \{\succ L_j\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \Big|_{T=t \text{ if } V_i \in \mathbb{M}}}_{f(\preceq L_j)} \times S(L_j \mid \mathrm{mp}_{\mathcal{G}}(L_j)) \\
&\qquad \times \prod_{V_i \in \{\preceq L_j\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \Big|_{T=t \text{ if } V_i \in \mathbb{M}} \\
&\overset{(3)}{=} \sum_{\preceq L_j} f(\preceq L_j) \times \frac{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \big|_{T=t}}{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times S(L_j \mid \mathrm{mp}_{\mathcal{G}}(L_j)) \times \prod_{V_i \in \{\preceq L_j\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \\
&\overset{(4)}{=} \mathbb{E}\left[\frac{\prod_{M_i \in \mathbb{M}\{\cap \prec L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \big|_{T=t}}{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times f(\preceq L_j) \times S(L_j \mid \mathrm{mp}_{\mathcal{G}}(L_j)) \right] \\
&\overset{(5)}{=} \mathbb{E}\left[\frac{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \big|_{T=t}}{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times \Big(f(\preceq L_j) - \sum_{L_j} f(\preceq L_j) \times p(L_j \mid \mathrm{mp}_{\mathcal{G}}(L_j)) \Big) \times S(L_j \mid \mathrm{mp}_{\mathcal{G}}(L_j)) \right] \\
&\overset{(6)}{=} \mathbb{E}\left[\frac{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \big|_{T=t}}{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times \Big(f(\preceq L_j) - \sum_{L_j} f(\preceq L_j) \times p(L_j \mid \mathrm{mp}_{\mathcal{G}}(L_j)) \Big) \times S(V) \right]
\end{aligned}$$

The first equality follows from the fact that terms corresponding to $M_i \in \mathbb{M}$ are not functions of T, the fact that $\mathbb{C}, \mathbb{M}, \mathbb{L}$ partition V, and term grouping. The second equality is by definition of conditional scores. The third equality is by term cancellation. The fourth is by definition of joint expectations, the fifth and sixth equalities are implied by the fact that conditional scores have expected value of 0 (given their conditioning set). Therefore, the contribution of $L_j \in \mathbb{L} \setminus T$ is the following:

$$\begin{aligned}
&\frac{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \big|_{T=t}}{\prod_{M_i \in \mathbb{M} \cap \{\prec L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times \Big(\sum_{\succ L_j} Y \times \prod_{V_i \in \{\succ L_j\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \Big|_{T=t \text{ if } V_i \in \mathbb{M}} \\
&\qquad - \sum_{\succeq L_j} Y \times \prod_{V_i \in \{\succeq L_j\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \Big|_{T=t \text{ if } V_i \in \mathbb{M}} \Big).
\end{aligned}$$

Third Term: The contribution of the last term to the final IF is as follows.

$$\sum_{V\setminus T} Y \times \prod_{M_i\in\mathbb{M}} p_\kappa(M_i \mid \text{mp}_\mathcal{G}(M_i))|_{T=t} \times \sum_{T} \prod_{L_i\in\mathbb{L}\setminus T} p_\kappa(L_i \mid \text{mp}_\mathcal{G}(L_i)) \times \frac{d}{d\kappa}\Big\{p_\kappa(T,\mathbb{C})\Big\}$$

$$\overset{(1)}{=} \sum_{T,\mathbb{C}} \Big\{ \underbrace{\sum_{V\setminus T,\mathbb{C}} Y \times \prod_{M_i\in\mathbb{M}} p_\kappa(M_i \mid \text{mp}_\mathcal{G}(M_i))|_{T=t} \times \prod_{L_i\in\mathbb{L}\setminus T} p_\kappa(L_i \mid \text{mp}_\mathcal{G}(L_i))}_{f(T,\mathbb{C})} \Big\} \times \frac{d}{d\kappa}\Big\{p_\kappa(T,\mathbb{C})\Big\}.$$

$$\overset{(2)}{=} \sum_{T,\mathbb{C}} f(T,\mathbb{C}) \times S(T,\mathbb{C}) \times p(T,\mathbb{C}) = \mathbb{E}\Big[f(T,\mathbb{C}) \times S(T,\mathbb{C})\Big]$$

$$\overset{(3)}{=} \mathbb{E}\Big[\Big(f(T,\mathbb{C}) - \sum_{T,\mathbb{C}} f(T,\mathbb{C}) \times p(T,\mathbb{C})\Big) \times S(T,\mathbb{C})\Big]$$

$$\overset{(4)}{=} \mathbb{E}\Big[\Big(f(T,\mathbb{C}) - \psi(t)\Big) \times S(V)\Big].$$

The first equality is term grouping, the second is by definition of marginal scores, the third and fourth equalities are implied by the fact that scores have expected value 0. Therefore, the contribution of the last term is the following:

$$\sum_{V\setminus\{T,\mathbb{C}\}} Y \times \prod_{M_i\in\mathbb{M}} p(M_i \mid \text{mp}_\mathcal{G}(M_i))\Big|_{T=t} \times \prod_{L_i\in\mathbb{L}\setminus T} p(L_i \mid \text{mp}_\mathcal{G}(L_i)) - \psi(t).$$

Putting all these together yields the final influence function. □

Lemma 1 (Double robustness of augmented primal IPW)

Proof. We need to show that under correct specification of conditional densities in either $\{p(M_i \mid \text{mp}_\mathcal{G}(M_i)), \forall M_i \in \mathbb{M}\}$ or $\{p(L_i \mid \text{mp}_\mathcal{G}(L_i)), \forall L_i \in \mathbb{L}\}$, the influence function in Theorem 1 remains to be mean zero. We break this down into two scenarios.

Scenario 1. Assume models in $\mathbb{L}$ are correctly specified, and let $p^*(M_i \mid \text{mp}_\mathcal{G}(M_i))$ denote the misspecified model for $p(M_i \mid \text{mp}_\mathcal{G}(M_i)), \forall M_i \in \mathbb{M}$. We note that for any $L_j \in \mathbb{L} \setminus T$, the following line in the IF evaluates to zero in expectation.

$$
\begin{aligned}
&\mathbb{E}\Big[\ \frac{\prod_{M_i \prec L_j} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))}\Big(\sum_{\succ L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_j\}} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad - \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq L_j\}} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}\Big)\ \Big] \\
&\overset{(1)}{=} \sum_{\preceq L_j} \frac{\prod_{M_i \prec L_j} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times \prod_{V_i \prec L_j} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \times p(L_j \times \mathrm{mp}_{\mathcal{G}}(L_j)) \\
&\qquad \times \Big(\sum_{\succ L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_j\}} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad\qquad - \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}\Big) \\
&\overset{(2)}{=} \sum_{\prec L_j} \frac{\prod_{M_i \prec L_j} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times \prod_{V_i \prec L_j} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \times \sum_{L_j} p(L_j \times \mathrm{mp}_{\mathcal{G}}(L_j)) \\
&\qquad \times \Big(\sum_{\succ L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_j\}} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad\qquad - \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}\Big) \\
&\overset{(3)}{=} \sum_{\prec L_i} \frac{\prod_{M_i \prec L_j} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times \prod_{V_i \prec L_j} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \\
&\qquad \times \Big(\sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad\qquad - \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_j\}} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}\Big) \\
&\overset{(4)}{=} 0.
\end{aligned}
$$

The first equality is by definition of joint expectation. The second equality is by the fact that terms associated with $\prec L_j$ are not functions of L_j. The third equality is by term grouping.

Moreover, for any $M_j, M_{j-1} \in \mathbb{M}$, the following equality holds,

$$
\begin{aligned}
&\mathbb{E}\Bigg[\frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))\ |_{T=t}\Bigg] \\
&= \mathbb{E}\Bigg[\frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_{j-1}} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \sum_{T \cup \{\succ M_{j-1}\}} Y \times \prod_{L_i \in\ \mathbb{L}} p(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ M_{j-1}\}} p^*(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))\ |_{T=t}\Bigg],
\end{aligned}
$$

since the left hand side is equal to

$$
\begin{aligned}
&\sum_{\prec M_i} p(\prec M_i) \times \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \\
&\qquad \times \Big[\sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big] \\
&\overset{(1)}{=} \sum_{\preceq M_{j-1}} p(\preceq M_{i-1}) \times \Bigg\{ \sum_{M_{j-1} \prec L_k \prec M_j} \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \times p(L_k \mid \text{mp}_{\mathcal{G}}(L_k)) \\
&\qquad \times \Big[\sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big] \Bigg\} \\
&\overset{(2)}{=} \sum_{\preceq M_{j-1}} p(\preceq M_{j-1}) \times \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_{j-1}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \\
&\qquad \times \sum_{M_{j-1} \prec L_k \prec M_j} \Big\{ \sum_{T \cup \{\succeq M_j\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big\} \\
&\overset{(3)}{=} \sum_{\preceq M_{j-1}} p(\preceq M_{j-1}) \times \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_{j-1}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \\
&\qquad \times \Big\{ \sum_{T \cup \{\succ M_{j-1}\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq M_j\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big\} \\
&\overset{(4)}{=} \mathbb{E}\Bigg[\frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_{j-1}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i))} \\
&\qquad \times \Big\{ \sum_{T \cup \{\succ M_{j-1}\}} Y \times \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \{\mathbb{M} \cap \succ M_{j-1}\}} p^*(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big\} \Bigg],
\end{aligned}
$$

which is exactly the same as the right hand side. This leaves the IF with only two terms $\psi(t)$ and β_{primal} and according to Lemma 2, $\mathbb{E}[\beta_{\text{primal}}] = \psi(t)$, provided the models in $\mathbb{L}$ are correctly specified, which was assumed. Therefore, $\mathbb{E}[U_{\psi_t}] = 0$.

Scenario 2. Assume models in $\mathbb{M}$ are correctly specified, and let $p^*(L_i \mid \text{mp}_{\mathcal{G}}(L_i))$ denote the misspecified model for $p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)), \forall L_i \in \mathbb{L}$. We note that for any $M_j \in \mathbb{M}$, the following line in the IF evaluates to zero.

$$
\begin{aligned}
&\mathbb{E}\Big[\ \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))}\Big(\sum_{T\cup\{\succ M_j\}} Y \times \prod_{L_i \in\ \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M}\cap\{\succ M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad\qquad - \sum_{T\cup\{\succeq M_j\}} Y \times \prod_{L_i \in\ \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M}\cap\{\succeq M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}\Big)\Big] \\
&\overset{(1)}{=} \sum_{\preceq M_j} \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \prod_{V_i \prec M_j} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \times p(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j)) \\
&\qquad \times \Big(\sum_{T\cup\{\succ M_j\}} Y \times \prod_{L_i \in\ \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M}\cap\{\succ M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad\quad - \sum_{T\cup\{\succeq M_j\}} Y \times \prod_{L_i \in\ \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M}\cap\{\succeq M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}\Big) \\
&\overset{(2)}{=} \sum_{\prec M_j} \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \prod_{V_i \prec M_j} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \times \sum_{M_j} p(M_j \mid \mathrm{mp}_{\mathcal{G}}(M_j)) \\
&\qquad \times \Big(\sum_{T\cup\{\succ M_j\}} Y \times \prod_{L_i \in\ \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M}\cap\{\succ M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad\quad - \sum_{T\cup\{\succeq M_j\}} Y \times \prod_{L_i \in\ \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M}\cap\{\succeq M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}\Big) \\
&\overset{(3)}{=} \sum_{\prec M_j} \frac{\mathbb{I}(T=t)}{\prod_{L_i \prec M_j} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i))} \times \prod_{V_i \prec M_j} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)) \\
&\qquad \times \Big(\sum_{T\cup\{\succeq M_j\}} Y \times \prod_{L_i \in\ \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M}\cap\{\succeq M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t} \\
&\qquad\quad - \sum_{T\cup\{\succeq M_j\}} Y \times \prod_{L_i \in\ \mathbb{L}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M}\cap\{\succeq M_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}\Big) \\
&\overset{(4)}{=} 0.
\end{aligned}
$$

Moreover, for any $L_j, L_{j-1} \in \mathbb{L}$, the following equality holds,

$$
\begin{aligned}
&\mathbb{E}\Big[\ \frac{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \\
&\qquad\qquad \times \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L}\cap\{\succeq L_j\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M}\cap\{\succeq L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))\ |_{T=t}\Big] \\
&\mathbb{E}\Big[\ \frac{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \\
&\qquad\qquad \times \sum_{\succ L_{j-1}} Y \times \prod_{L_i \in \mathbb{L}\cap\{\succ L_{j-1}\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M}\cap\{\succ L_{j-1}\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))\ |_{T=t}\Big],
\end{aligned}
$$

since the left hand side is equal to

$$
\begin{aligned}
&\sum_{\prec L_j} p(\prec L_j) \times \frac{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \\
&\qquad \times \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \\
&\overset{(1)}{=} \sum_{\preceq L_{j-1}} p(\preceq L_{j-1}) \times \Bigg\{ \sum_{L_{j-1} \prec M_k \prec L_j} \frac{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_j} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times p(M_k \mid \mathrm{mp}_{\mathcal{G}}(M_k)) \\
&\qquad \times \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Bigg\} \\
&\overset{(2)}{=} \sum_{\preceq L_{j-1}} p(\preceq L_{j-1}) \times \frac{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times \Bigg\{ \sum_{L_{j-1} \prec M_k \prec L_j} p(M_k \mid \mathrm{mp}_{\mathcal{G}}(M_k))|_{T=t} \\
&\qquad \times \sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Bigg\} \Bigg\} \\
&\overset{(3)}{=} \sum_{\preceq L_{j-1}} p(\preceq L_{j-1}) \times \frac{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \times \\
&\qquad \times \Bigg\{ \sum_{L_{j-1} \prec M_k \prec L_j} \Big[\sum_{\succeq L_j} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succeq L_j\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succeq L_j\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Big] \\
&\qquad\qquad \times p(M_k \mid \mathrm{mp}_{\mathcal{G}}(M_k))|_{T=t} \Bigg\} \\
&\overset{(4)}{=} \sum_{\preceq L_{j-1}} p(\preceq L_{j-1}) \times \frac{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \\
&\qquad \times \sum_{\succ L_{j-1}} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_{j-1}\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_{j-1}\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Bigg\} \\
&\overset{(5)}{=} \mathbb{E}\Bigg[\frac{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))|_{T=t}}{\prod_{M_i \prec L_{j-1}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i))} \\
&\qquad \times \sum_{\succ L_{j-1}} Y \times \prod_{L_i \in \mathbb{L} \cap \{\succ L_{j-1}\}} p^*(L_i \mid \mathrm{mp}_{\mathcal{G}}(L_i)) \times \prod_{M_i \in \mathbb{M} \cap \{\succ L_{j-1}\}} p(M_i \mid \mathrm{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \Bigg],
\end{aligned}
$$

which is exactly the same as the right hand side. This leaves the IF with only two terms $\psi(t)$ and β_{dual} and according to Lemma **??**, $\mathbb{E}[\beta_{\text{dual}}] = \psi(t)$. Therefore, $\mathbb{E}[U_{\psi_t}] = 0$. □

Lemma 2 (Primal and Dual IPWs)

Proof. Our goal is to demonstrate that the primal IPW formulation is equivalent to the identifying functional of the target parameter $\psi(t)$ shown in Eq. 2.2 and restated below.

$$\psi(t) = \sum_{V \setminus T} \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t} \times \sum_{T} \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y.$$

The primal IPW formulation for the target $\psi(t)$ is,

$$\mathbb{E}[\beta_{\text{primal}}(t)] \equiv \mathbb{E}\left[\frac{\mathbb{I}(T=t)}{q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T))} \times Y\right]$$

where $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T)) = \prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))$, and

$$\begin{aligned} q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T)) &= q_{D_T}(T \mid D_T \cup \text{pa}_{\mathcal{G}}(D_T) \setminus T) = \frac{q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))}{q_{D_T}(D_T \setminus T \mid \text{pa}_{\mathcal{G}}(D_T))} \\ &= \frac{q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))}{\sum_T q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))} = \frac{\prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}{\sum_T \prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))} \\ &= \frac{\prod_{V_i \in \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}{\sum_T \prod_{V_i \in \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}. \end{aligned}$$

The last equality holds because the conditional densities of $V_i \in \mathbb{C}$, does not depend on T, and they cancel out from the numerator and denominator. Therefore, product in the ratio is over the variables in $D_T \cap \{\succeq T\}$ which we have denoted by $\mathbb{L}$. Therefore,

$$\begin{aligned} \mathbb{E}[\beta_{\text{primal}}(t)] &= \mathbb{E}\left[\mathbb{I}(T=t) \times \frac{\sum_T \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))}{\prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))} \times Y\right] \\ &= \sum_V \prod_{V_i \in V} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \mathbb{I}(T=t) \times \frac{\sum_T \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))}{\prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))} \times Y \\ &= \sum_V \mathbb{I}(T=t) \times \prod_{V_i \in V \setminus \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \\ &\qquad \times \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times \frac{\sum_T \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))}{\prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))} \times Y \\ &= \sum_V \mathbb{I}(T=t) \times \prod_{V_i \in V \setminus \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \sum_T \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y. \end{aligned}$$

In the second equality, we evaluated the outer expectation with respect to the joint $p(V)$. In the third equality, we partitioned the joint into factors for the set $\mathbb{L}$ and factors for $V \setminus \mathbb{L}$. In the fourth equality, we canceled out the the factors involved in the denominator of the primal IPW with the corresponding terms in the joint.

We can then move the conditional factors of pre-treatment variables in the district of T past the summation over T as these factors are not functions of T. Finally, we evaluate the indicator function, concluding the proof. That is,

$$\psi_{\text{primal}} = \sum_{V} \mathbb{I}(T = t) \times \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \sum_{T} \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y$$

$$= \sum_{V \setminus T} \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t} \times \sum_{T} \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y = \psi(t)$$

The proof strategy is similar to the one used for the primal IPW. The dual IPW formulation for the target $\psi(t)$ is,

$$\begin{aligned}
\mathbb{E}[\beta_{\text{dual}}(t)] &= \mathbb{E}\Big[\frac{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t}}{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y\Big] \\
&= \sum_{V} \prod_{V_i \in V} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \frac{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t}}{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y \\
&= \sum_{V} \prod_{V_i \in V \setminus \text{mp}_{\mathcal{G}}^{-1}(T)} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \\
&\qquad \times \prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \times \frac{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t}}{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y \\
&= \sum_{V} \prod_{V_i \in V \setminus \text{mp}_{\mathcal{G}}^{-1}(T)} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \times Y \\
&= \sum_{V \setminus T} \prod_{V_i \in V \setminus \{\text{mp}_{\mathcal{G}}^{-1}(T) \cup D_T\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \\
&\qquad \times \sum_{T} \prod_{D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y.
\end{aligned}$$

In the above derivation, we first evaluated the outer expectation with respect to the joint $p(V)$. We then partitioned the joint into factors corresponding to $\text{mp}_{\mathcal{G}}^{-1}(T)$ and $V \setminus \text{mp}_{\mathcal{G}}^{-1}(T)$. The factors involved in the denominator of the dual IPW then canceled out with the corresponding terms in the joint. The last equality holds because by the definition of the inverse Markov pillow, $\text{mp}_{\mathcal{G}}^{-1}(T)$ contains all variables not in the district of T such that T is a member of its Markov pillow. In the above expression, factors corresponding to the inverse Markov pillow of T are evaluated at

$T = t$. Consequently, the only factors above that are still functions of T are the ones corresponding to the district of T. This allows us to push the summation over T.

Finally, since the summation over T will prevent factors within the district of T from being evaluated at $T = t$, we can simply apply the evaluation to the entire functional and merge the sets not involved in the district of T above. That is,

$$\psi_{\text{dual}} = \sum_{V \setminus T} \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \sum_{T} \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y \Big|_{T=t} = \psi(t).$$

□

Appendix IV

Supplementary Materials for Causal SDR

A. Details and Additional Results

Assume treatment is collected using p equally spaced percentages of volume. In other words, treatment is assumed to be a vector in $\mathbb{R}^p$ where the i^{th} element corresponds to the radiation dose on $q\%$ of the parotid glands. The effect of radiation on weight loss is illustrated in Fig. IV-1 by allowing p to be 10 and 20, and reducing the size of treatment to one dimension. We use IPW estimators to calculate the effects. Both plots agree with our stated conclusion in the main body of the manuscript, i.e., radiation has a negative effect on weight loss.

B. Proofs

LEMMA 3 *An estimator for β which solves (2.12) under the correct specification of $p(T \mid C)$, and either one of $\ell(g(T;\beta)) \equiv \mathbb{E}_q[Y \mid g(T;\beta)]$ or $\nu(g(T;\beta)) \equiv \mathbb{E}_q[\alpha(T) \mid g(T;\beta)]$, is consistent.*

Proof. Choosing $\phi(T, C) = 0$ in Theorem 2 yields (2.12). All elements of the orthocomplement of the nuisance tangent space are mean zero under the true distribution (we

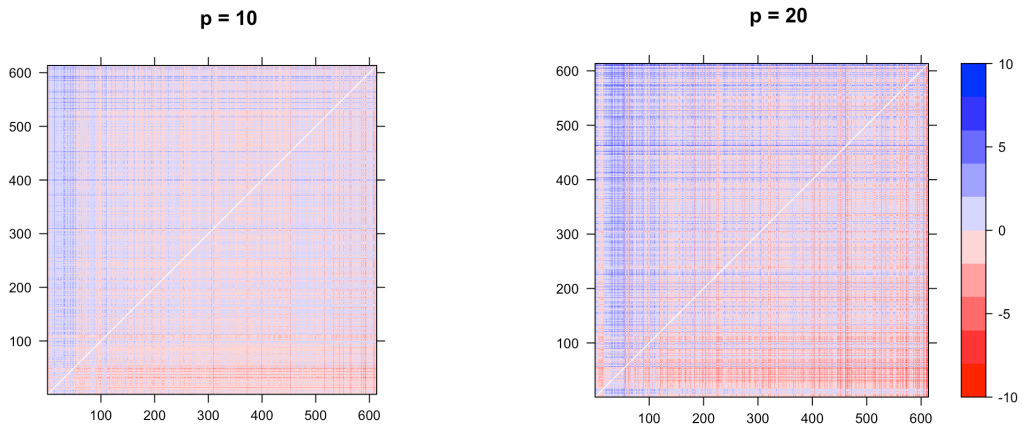


Figure IV-1: Heatmaps to illustrate the causal effect of radiation on weight loss, where effects are computed by estimating β via IPW estimator and treatment is collected using (a) 10, (b) 20 equally spaced percentages of volume in parotid glands.

give an argument for elements of $\widetilde{\Lambda}_\eta^\perp$ in Proposition IV). Since $\widetilde{U}(\beta)$ exhibits double robustness, i.e. remaining consistent if either $\ell(g(T;\beta))$ or $\nu(g(T;\beta))$ is correctly specified [59], the correct specification of $p(T \mid C)$ yields our conclusion. □

PROPOSITION IV *For all* $\tilde{U}(\beta^*) \in \widetilde{\Lambda}_\eta^\perp$, $\mathbb{E}[\widetilde{U}(\beta^*)] = 0$.

Proof. The second and third terms of $\widetilde{U}(\beta^*)$ are mean zero by construction. The first term, under truth with the property that $\mathbb{E}_q[Y \mid T] = \mathbb{E}_q[Y \mid g(T;\beta)]$, is

$$\begin{aligned}
\mathbb{E}\Big[\frac{p^*(T)}{p(T \mid C)} \times \widetilde{U}(\beta)\Big] &= \int \widetilde{U}(\beta) \times p(Y \mid T, C) \times p^*(T) \times p(C)\; d\mu_{Y,T,C} \\
&= \int \{Y - \ell(g(T;\beta))\} \times \{\alpha(T) - \nu(g(t;\beta))\} \times q(Y,T,C)\; d\mu_{Y,T,C} \\
&= \mathbb{E}_q\Big[\{Y - \ell(g(t;\beta))\} \times \{\alpha(T) - \nu(g(t;\beta))\}\Big] \\
&= \mathbb{E}_q\Big[\{\alpha(T) - \nu(g(t;\beta))\} \times \mathbb{E}_q[\{Y - \ell(g(t;\beta))\} \mid T = t]\Big] \\
&= \mathbb{E}_q\Big[\{\alpha(T) - \nu(g(t;\beta))\} \times \{\mathbb{E}_q[Y \mid T = t] - \ell(g(t;\beta))\}\Big] \\
&= 0.
\end{aligned}$$

since $\ell(g(t;\beta)) \coloneqq \mathbb{E}_q[Y \mid T = t]$. Note that even if $\ell(g(t;\beta))$ is miss-specified, the expectation will still be zero if $\nu(g(t;\beta))$ is correctly specified, shown by iterative

expectations.

□

THEOREM 2 *The orthogonal complement of the nuisance tangent space $\widetilde{\Lambda}_\eta^\perp$ for $\mathcal{M}$ contains elements of the form*

$$\widetilde{\Lambda}_\eta^\perp = \Big\{ \frac{\widetilde{U}(\beta)}{W_t(C)} - \phi(T,C) + \mathbb{E}[\phi(T,C) \mid C] \Big\},$$

where $\phi(T,C)$ is an arbitrary function of T and C, $W_t(C)$ is the IPW weight $p(T = t \mid C)/p^*(t)$ for a fixed $p^*(t)$, and $\widetilde{U}(\beta)$ is of the form

$$\widetilde{U}(\beta) = \Big\{ Y - \ell(g(t;\beta)) \Big\} \times \Big\{ \alpha(T) - \nu(g(t;\beta)) \Big\},$$

where $\ell(g(t;\beta)) \equiv \mathbb{E}_q[Y \mid g(t;\beta)]$ and $\nu(g(t;\beta)) \equiv \mathbb{E}_q[\alpha(T) \mid g(T;\beta)]$. Moreover, the most efficient estimator in this class, for any fixed $\alpha(T)$, is recovered by setting $\phi^{opt}(T,C) = \mathbb{E}\Big[\frac{\widetilde{U}(\beta)}{W_t(C)} \mid T, C \Big]$.

Proof. This is a direct consequence of Theorems 3.1 and 3.2 in [60], and results in Appendix 3 of [59]. □

LEMMA 4 *For a fixed choice of $\alpha(T)$ and normalized function $p^*(T)$, the element $\widetilde{U}(\beta^*) \in \widetilde{\Lambda}_\eta^\perp$ corresponding to the optimal choice of $\phi(T,C)$ has the form.*

$$\frac{p^*(T)}{p(T \mid C)} \times \widetilde{U}(\beta) - \frac{p^*(T)}{p(T \mid C)} \times \mathbb{E}\left[\widetilde{U}(\beta) \Big| T, C \right] + \mathbb{E}_q\left[\mathbb{E}\left[\widetilde{U}(\beta) \Big| T, C \right] \Big| C \right],$$

where $\mathbb{E}_q[.]$ is the expectation taken with respect to the density $q(Y,T,C) \equiv p(Y \mid T,C) \times p^*(T) \times p(C)$.

Proof. Plugging in the optimal $\phi(T, C)$ yields $\widetilde{U}(\beta^*)$ to be

$$\frac{p^*(T)}{p(T \mid C)} \times \widetilde{U}(\beta) - \mathbb{E}\left[\frac{p^*(T)}{p(T \mid C)}\ \widetilde{U}(\beta)\Big|T, C\right] + \mathbb{E}\left[\mathbb{E}\left[\frac{p^*(T)}{p(T \mid C)}\ \widetilde{U}(\beta)\Big|T, C\right]\Big|C\right].$$

The conclusion follows, since

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\frac{p^*(T)}{p(T \mid C)}\ \widetilde{U}(\beta)\Big|T, C\right]\Big|C\right] \\
&\qquad = \mathbb{E}\left[\frac{p^*(T)}{p(T \mid C)}\mathbb{E}\left[\ \widetilde{U}(\beta)\Big|T, C\right]\Big|C\right] \\
&\qquad = \int \frac{p^*(T)}{p(T \mid C)}\ \mathbb{E}[\widetilde{U}(\beta) \mid T, C]\ p(Y, T \mid C)\ d\mu_{Y,T} \\
&\qquad = \int \mathbb{E}[\widetilde{U}(\beta) \mid T, C]\ p(Y \mid T, C)\ p^*(T)\ d\mu_{Y,T} \\
&\qquad = \int \mathbb{E}[\widetilde{U}(\beta) \mid T, C]\ q(Y, T \mid C)\ d\mu_{Y,T} \\
&\qquad = \mathbb{E}_q\Big[\mathbb{E}\Big[\widetilde{U}(\beta) \mid T, C\Big]\Big|C\Big].
\end{aligned}$$

□

LEMMA 5 *If one of* $\Big\{p(T \mid C),\ \mathbb{E}[\widetilde{U}(\beta) \mid T, C]\Big\}$ and *one of* $\Big\{\ell(g(T;\beta)) \equiv \mathbb{E}_q[Y \mid g(T;\beta)],\ \nu(g(T;\beta)) \equiv \mathbb{E}_q[\alpha(T) \mid g(T;\beta)]\Big\}$ *is correctly specified, then the estimator for* β *based on (2.13) is consistent and asymptotically normal with mean zero and variance equal to* $\tau^{-1} \times Var\Big(\widetilde{U}(\beta^*)\Big) \times \tau^{-1'}$, *where* $\widetilde{U}(\beta^*)$ *is given in (2.13) and* τ *is defined as* $\mathbb{E}[\frac{\partial \widetilde{U}(\beta^*)}{\partial \beta}]$.

Proof. Assume either $\ell(g(T;\beta))$ or $\nu(g(T;\beta))$, and $p(T \mid C)$ are correctly specified. Consequently, the second and third terms in the expression of $\widetilde{U}(\beta^*)$ are both mean zero, even under an incorrect specification of $\mathbb{E}[\widetilde{U}(\beta) \mid T, C]$. Following the same the argument in Proposition IV, the first term is zero if either $\ell(g(T;\beta))$ or $\nu(g(T;\beta))$ is correctly specified.

Assume either $\ell(g(T;\beta))$ or $\nu(g(T;\beta))$, and $\mathbb{E}[\widetilde{U}(\beta) \mid T, C]$ are correctly specified. Consequently, the first two terms in the expression of $\widetilde{U}^*$ are both mean zero, even under an incorrect specification of $p^*(T \mid C)$. For the last term, we have:

$$
\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}_q\Big[\mathbb{E}[\widetilde{U}(\beta) \mid T, C] \;\Big|\; C\Big]\Big] \\
&\qquad = \mathbb{E}\Big[\mathbb{E}_q\Big[\int \widetilde{U}(\beta) \times p(Y \mid T, C)\; d\mu_Y \;\Big|\; C\Big]\Big] \\
&\qquad = \mathbb{E}\Big[\int \Big(\int \widetilde{U}(\beta) \times p(Y \mid T, C)\; d\mu_Y\Big) \times p^*(A)\; d\mu_A\Big] \\
&\qquad = \int \Big(\int\int \widetilde{U}(\beta) \times p(Y \mid T, C) \times p^*(T)\; d\mu_Y\; d\mu_T\Big) \times p(C)\; d\mu_C \\
&\qquad = \int \widetilde{U}(\beta) \times p(Y \mid T, C) \times p^*(T) \times p(C)\; d\mu_{Y,T,C} \\
&\qquad = \int \widetilde{U}(\beta) \times q(Y, T, C)\; d\mu_{Y,T,C} \\
&\qquad = \mathbb{E}_q[\widetilde{U}(\beta)].
\end{aligned}
$$

We conclude the proof by noting that $\mathbb{E}\Big[\widetilde{U}(\beta)\Big]$ is mean zero if either $\ell(g(T;\beta))$ or $\nu(g(T;\beta))$ is correctly specified. Note that the normalized version of $\widetilde{U}(\beta^*)$, that is $\mathbb{E}[\frac{\partial \widetilde{U}(\beta^*)}{\partial \beta}]^{-1} \times \widetilde{U}(\beta^*)$, is an influence function that lives in the orthogonal complement of the tangent space $\widetilde{\Lambda}_\eta^\perp$. Therefore, the estimator obtained by solving $\mathbb{E}[\widetilde{U}(\beta^*)] = 0$ is RAL and is consistent and asymptotically normal with mean zero and variance equal to the variance of the influence function [39, 41]. □

THEOREM 3 *Let ϕ_0 denote the influence function of the estimator β obtained from the estimating equation $\mathbb{E}[\widetilde{U}(\beta^*, \eta_0)] = 0$. If $n^{\frac{1}{4}+\epsilon}(\widehat{\eta} - \eta_0)$ is bounded in probability for some $\epsilon > 0$, then the influence function corresponding to the estimator $\widehat{\beta}$ obtained from the estimating equation $\mathbb{E}[\widetilde{U}(\beta^*, \widehat{\eta})] = 0$ is the same as ϕ_0. In other words, $\widehat{\beta}$ follows the same asymptotic properties as if we knew the true nuisance models.*

Proof. Let $\beta \in \mathbb{R}^q$ and let η be infinite dimensional. We prove this theorem for the parametric submodel in the semiparametric model of $\{p(Z;\beta,\eta)\}$. With a slight abuse

of notation, we denote $\eta \in \mathbb{R}^r$ to be the nuisance parameters within the parametric submodel. The Taylor series expansion of $\widetilde{U}(Z; \widehat{\beta}(\widehat{\eta}), \widehat{\eta})$ around β_0 is

$$\begin{aligned} 0 =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{U}(z_i; \widehat{\beta}(\widehat{\eta}), \widehat{\eta}) \\ =& \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{U}(z_i; \beta_0, \widehat{\eta})}_{(a)} + \underbrace{\frac{\partial}{\partial \beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \widetilde{U}(z_i; \beta_0, \widehat{\eta}) \right\}}_{(b)} \sqrt{n}(\widehat{\beta} - \beta_0) + o_p(1) \end{aligned} \tag{IV.1}$$

$$\begin{aligned} (a) =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{U}(z_i; \beta_0, \eta_0) \\ &+ \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \widetilde{U}(z_i; \beta_0, \eta_0)}{\partial \eta} \right)_{q \times r} \times \sqrt{n}(\widehat{\eta} - \eta_0) \\ &+ \frac{1}{2} \underbrace{n^{1/4}(\widehat{\eta} - \eta_0)'}_{1 \times 1 \times r} \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \widetilde{U}(z_i; \beta_0, \eta_0)}{\partial^2 \eta} \right)}_{r \times q \times r \text{ (tensor)}} \underbrace{n^{1/4}(\widehat{\eta} - \eta_0)}_{r \times 1 \times 1} + o_p(1) \end{aligned}$$

$$(b) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \widetilde{U}(z_i; \beta_0, \eta_0)}{\partial \beta} \right)_{q \times q}}_{(b_1)} + \frac{\partial}{\partial \beta} \Bigg\{ \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \widetilde{U}(z_i; \beta_0, \eta_0)}{\partial \eta} \right)_{q \times r}}_{(b_2)} \times (\widehat{\eta} - \eta_0)_{r \times 1} \Bigg\}$$

$$(b_1) : \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \widetilde{U}}{\partial \beta} \right)_{q \times q} \longrightarrow \mathbb{E}_{\theta_0} \left[\frac{\partial \widetilde{U}}{\partial \beta} \right]_{q \times q} = -\mathbb{E}_{\theta_0} \left[\widetilde{U}(Z; \theta_0) S'_{\beta}(Z; \theta_0) \right]$$

$$(b_2) : \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \widetilde{U}}{\partial \eta} \right)_{q \times r} \longrightarrow \mathbb{E}_{\theta_0} \left[\frac{\partial \widetilde{U}}{\partial \eta} \right]_{q \times r} = -\mathbb{E}_{\theta_0} \left[\widetilde{U}(Z; \theta_0) S'_{\eta}(Z; \theta_0) \right] = 0_{q \times r}$$

Since $n^{1/4}(\widehat{\eta} - \eta_0)$ and $\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \widetilde{U}(z_i; \beta_0, \eta_0)}{\partial \eta} \right)_{q \times r}$ both converge in probability to zero, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{U}(z_i; \beta_0, \widehat{\eta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{U}(z_i; \beta_0, \eta_0) + o_p(1).$$

Therefore, from equation IV.1

$$\sqrt{n}(\widehat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ -\mathbb{E}_{\theta_0}^{-1} \left[\frac{\partial \widetilde{U}(z_i; \beta_0, \eta_0)}{\partial \beta} \right] \widetilde{U}(z_i; \beta_0, \eta_0) \right\} + o_p(1)$$

Which concludes the proof. This procedure carries over to the case where the nuisance parameter is infinite dimensional [41]. □

LEMMA 6 *Let $U_{dim}(\psi) = Y - \widetilde{f}(T, C, \beta; \psi)$, and fix any $d(T, C)$. If either $\mathbb{E}[d(T, C) \mid g(T; \beta)]$ or $\mathbb{E}[U_{dim}(\psi) \mid g(T; \beta)]$ are correctly specified, the following estimating equations yield a consistent estimator of ψ,*

$$\mathbb{E}\Big[\{d(T, C) - \mathbb{E}[d(T, C) \mid g(T; \beta)]\} \times \{U_{dim}(\psi) - \mathbb{E}[U_{dim}(\psi) \mid g(T; \beta)]\}\Big] = 0.$$

Proof. Define $U_{dim}(\psi) = Y - \widetilde{f}(T, C, \beta; \psi)$. Therefore,

$$\mathbb{E}[U_{dim}(\psi) \mid T, C] = \ell(g(T; \beta)) = \mathbb{E}[U_{dim}(\psi) \mid g(T; \beta)].$$

This is a situation precisely isomorphic to single treatment SNMMs above, except with the roles of A and C reversed (hence this is an "inverted SNMM"). Our conclusion will then follow by results in [24, 63]. We provide a more detailed proof as follows. We have that $\widetilde{f}(t, C, \beta; \psi) = \mathbb{E}[Y \mid T = t, C] - \ell(g(t; \beta))$. Therefore,

$$\mathbb{E}[Y \mid T = t, C = C] = \ell(g(t; \beta)) + \widetilde{f}(t, C, \beta; \psi),$$

which we can rewrite as follows,

$$Y = \ell(g(t; \beta)) + \widetilde{f}(t, C, \beta; \psi) + \epsilon, \quad \text{s.t.} \quad \mathbb{E}[\epsilon \mid C, t] = 0.$$

Observed data are instances of the form $Z = (C, T, Y)$. The goal is to find semiparametric estimators for ψ in the semiparametric model $\mathcal{P} = \{p(z; \psi, \psi()), z = (c, t, y)\}$ and the truth is $p_0(z) = p(z; \psi_0, \eta_0())$. The observed data likelihood can be written as follows,

$$\begin{aligned} p(C, t, Y) &= p(C, t) \times p(Y \mid t, C) \equiv p(C, t) \times p(\epsilon \mid t, C) = \eta_1(C, t) \times \eta_2(\epsilon, t, C) \\ &= \eta_1(C, t) \times \eta_2\Big(Y - \ell(g(t; \beta)) - \widetilde{f}(t, C, \beta; \psi), t, C\Big), \end{aligned}$$

where $\epsilon = Y - \ell(g(t; \beta)) - \widetilde{f}(t, C, \beta; \psi)$, $\eta_1(C, t)$ denotes the nuisance model for $p(C, t)$, and $\eta_2(\epsilon, t, C)$ denotes the nuisance model for $p(\epsilon \mid t, C)$, which is any density such

that $\mathbb{E}[\epsilon \mid t, C] = 0$. ψ is the parameter of interest and the nuisance parameters are $\{\eta_1, \eta_2, \ell(g(t;\beta))\}$.

The nuisance tangent space of this semiparametric model, Λ, is defined as the mean-square closure of parametric submodel nuisance tangent spaces:

$$\begin{aligned}\mathcal{P}_{\psi,\zeta} &= \Big\{p(z;\psi,\psi_\zeta) = p(c,t;\zeta_1) \times p(\epsilon \mid t,c;\zeta_2)\Big\} \\ &= \Big\{p(c,t;\zeta_1) \times p\Big(y - \ell(g(t;\beta)) - \widetilde{f}(t,c,\beta;\psi) \mid t,c;\zeta_2\Big)\Big\},\end{aligned}$$

where ζ_1, ζ_2 are r_1, r_2 dimensional vectors. Thus nuisance parameters in parametric submode are finite dimensional, $\zeta = \{\zeta_1, \zeta_2, \ell(g(t;\beta))\}$.

$$\begin{aligned}\Lambda_\zeta &= \{B \times S_\zeta, \forall B\}, \\ S_\zeta &= \frac{\partial\{\text{log likelihood of the submodel evaluated at truth}\}}{\partial\zeta} \\ &= \left\{\Big(\frac{\partial \log p(z;\psi,\zeta)}{\partial\zeta_1}\Big), \Big(\frac{\partial \log p(z;\psi,\zeta)}{\partial\zeta_2}\Big), \Big(\frac{\partial \log p(z;\psi,\zeta)}{\partial\ell(g(t;\beta))}\Big)\right\}\Bigg|_{\psi_0,\zeta_0} \\ &= \Big\{S_{\zeta_1}(z;\psi_0,\zeta_0), S_{\zeta_2}(z;\psi_0,\zeta_0), S_{\ell(g(t;\beta))}(z;\psi_0,\zeta_0)\Big\}.\end{aligned}$$

Hence, $\Lambda_\zeta = \Lambda_{\zeta_1} + \Lambda_{\zeta_2} + \Lambda_{\ell(g(t;\beta))}$. S_{ζ_1} should satisfy the density conditions. In addition, S_{ζ_2} should satisfy the condition that $\mathbb{E}[\epsilon \mid t, C] = 0$. We derive each of these subspaces using theorems in [41] as a guideline.

(Theorem 4.6) $\quad \Lambda_{\zeta_1} = \{f(C,t); \mathbb{E}[f] = 0\}$

(Theorem 4.7) $\quad \Lambda_{\zeta_2} = \{f(\epsilon,t,C); \mathbb{E}[f \mid t,C] = 0, \mathbb{E}[\epsilon f \mid t,C] = 0\}$

(Lemma 4.3) $\quad \Lambda_{\zeta_1}^{\perp} = \{g(\epsilon,t,C); \mathbb{E}[g \mid t,C] = 0\}$

(Theorem 4.8) $\quad (\Lambda_{\zeta_1} + \Lambda_{\zeta_2})^{\perp} = \{g(C,t)\epsilon\}$

(Equation IV.2) $\quad \Lambda_{\ell(g(t;\beta))} = \{\frac{\psi'_{2\epsilon}(\epsilon,C,t)}{\psi_2(\epsilon,C,t)} f(g(t;\beta))\}$

In order to derive $\Lambda_{\ell(g(t;\beta))}$, we write down the corresponding score function as follows.

$$\begin{aligned}
S_{\ell(g(t;\beta))} &= \frac{\partial \log p(z;\psi,\zeta)}{\partial \ell(g(t;\beta))}\Big|_{\psi_0,\zeta_0} \\
&= \frac{\partial \log\Big(\psi_1(C,t;\zeta_{10}) \times \psi_2\Big(y-\ell(g(t;\beta))-\gamma(C,t;\psi),C,t;\zeta_{20}\Big)\Big)}{\partial \ell(g(t;\beta))} \\
&= \frac{\partial \log \psi_2\Big(y-\ell(g(t;\beta))-\gamma(C,t;\psi),l,t;\zeta_{20}\Big)}{\partial \ell(g(t;\beta))} \\
&= \frac{\partial \log \psi_2\Big(\epsilon,C,t;\zeta_{20}\Big)}{\partial \epsilon} \times \frac{\partial \epsilon}{\partial \ell(g(t;\beta))} \quad (\epsilon \text{ is a function of } \ell(g(t;\beta))) \\
&= \frac{\psi'_{2\epsilon}(\epsilon,C,t)}{\psi_2(\epsilon,C,t)} f(g(t;\beta)). \qquad \text{(IV.2)}
\end{aligned}$$

In order to derive $\Lambda_\zeta^\perp$, we proceed as follows. Since $\Lambda_\zeta = \Lambda_{\zeta_1} + \Lambda_{\zeta_2} + \Lambda_{\ell(g(t;\beta))}$ and $\Lambda_{\zeta_1} + \Lambda_{\zeta_2} \subset \Lambda_\zeta$, then $\Lambda_\zeta^\perp \subset (\Lambda_{\zeta_1} + \Lambda_{\zeta_2})^\perp = \{g(c,t)\epsilon\}$. Similarly, $\Lambda_\zeta^\perp \subset \Lambda_{\ell(g(t;\beta))}^\perp$, therefore $\Lambda_\zeta^\perp = \{(\Lambda_{\zeta_1} + \Lambda_{\zeta_2})^\perp \cap \Lambda_{\ell(g(t;\beta))}^\perp\}$.

Pick an arbitrary element in $(\Lambda_{\zeta_1} + \Lambda_{\zeta_2})^\perp$, and denote it by $d(C,t)\epsilon$. For $d(C,t)\epsilon$ to be an element in $\Lambda_\zeta^\perp$, it needs to be orthogonal to every element in $\Lambda_{\ell(g(t;\beta))}$. Pick an arbitrary element in $\Lambda_{\ell(g(t;\beta))}$ and denote it by $\frac{\psi'_{2\epsilon}}{\psi_2} h(g(t;\beta))$. We have,

$$\begin{aligned}
\forall h(g(t;\beta)) \quad 0 &=< d(C,t)\epsilon, \frac{\psi'_{2\epsilon}}{\psi_2} h(g(t;\beta)) > \\
&= \mathbb{E}\Big[d(C,t)\epsilon \frac{\psi'_{2\epsilon}}{\psi_2} h(g(t;\beta))\Big] \\
&= \mathbb{E}\Big[d(C,t) h(g(t;\beta))\Big].
\end{aligned}$$

Consequently, $\forall h(g(t;\beta))$:

$$\begin{aligned}
0 &= \mathbb{E}\Big[d(C,t) \times h(g(t;\beta))\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[d(C,t) \times h(g(t;\beta)) \;\Big|\; g(t;\beta)\Big]\Big] \\
&= \mathbb{E}\Big[h(g(t;\beta)) \times \mathbb{E}\Big[d(C,t) \;\Big|\; g(t;\beta)\Big]\Big] \\
&= \mathbb{E}\Big[h(g(t;\beta))\Big] \times \mathbb{E}\Big[d(C,t) \;\Big|\; g(t;\beta)\Big].
\end{aligned}$$

Therefore, $\mathbb{E}[\ d(C,t) \mid g(t;\beta)\ ] = 0$ and

$$\begin{aligned}
\Lambda_\zeta^\perp &= \Big\{ \Big(d(C,t) - \mathbb{E}[d(C,t) \mid g(t;\beta)] \Big) \times \epsilon \Big\} \\
&= \Big\{ \Big(d(C,t) - \mathbb{E}[d(C,t) \mid g(t;\beta)] \Big) \times \Big(Y - \gamma(C,t;\psi) - \ell(g(t;\beta)) \Big) \Big\} \\
&= \Big\{ \Big(d(C,t) - \mathbb{E}[d(C,t) \mid g(t;\beta)] \Big) \times \Big(U(\psi) - \mathbb{E}[U(\psi) \mid C,t] \Big) \Big\}.
\end{aligned}$$

Note that $\mathbb{E}[U(\psi) \mid C,t] = \mathbb{E}_q[Y \mid g(t;\beta)] = \mathbb{E}[U(\psi) \mid g(t;\beta)]$. Hence,

$$\Lambda_\zeta^\perp = \Big\{ \Big\{ d(C,t) - \mathbb{E}[d(C,t) \mid g(t;\beta)] \Big\} \times \Big\{ U(\psi) - \mathbb{E}[U(\psi) \mid g(t;\beta)] \Big\} \Big\}.$$

□

Appendix V

Supplementary Materials for Algorithmic Fairness

A. G-estimation

G-estimation applies to structural nested models, which directly model the counterfactual deviations in outcome from a reference treatment value (which we take to be $A = 0$) conditional on history, assuming all future decisions are already optimal. Specifically, for each decision point k we posit a *structural nested mean model (SNMM)* parameterized by ψ as follows:

$$\gamma_k(H_k, a_k; \psi) = \mathbb{E}[Y(\bar{a}_{k-1}, a_k, f^*_{\underline{A}_{k+1}}) - Y(\bar{a}_{k-1}, a_k = 0, f^*_{\underline{A}_{k+1}}) \mid H_k],$$

where $\underline{A}_{k+1}$ represents all treatments administered from time $k+1$ onwards. In words, γ_k is the contrast of the counterfactual mean (conditional on observed history H_k) where the past decisions are set to their observed values, the present decision is either a_k or a reference decision $a_k = 0$, and all future decisions are made optimally, $f^*_{\underline{A}_{k+1}}$.

If the true $\gamma_k(H_k, a_k; \psi)$ were known, the optimal treatment policies are those that maximize this "blip" function at each stage: $f^*_{A_k} = \arg\max_{a_k} \gamma_k(H_k, a_k; \psi)$. In order to estimate ψ using data, let

$$U(\psi, \zeta(\psi), \alpha) = \sum_{k=1}^{K} \{G_k(\psi) - \mathbb{E}\left[G_k(\psi) \mid H_k; \zeta\right]\} \times \{d_k(H_k, A_k) - E\left[d_k(H_k, A_k) \mid H_k; \alpha\right]\}, \tag{V.1}$$

where $d_k(H_k, A_k)$ is any function of H_k and A_k and $G_k(\psi)$ is defined as

$$Y - \gamma_k(H_k, a_k; \psi) + \sum_{i=k+1}^{K} \left[\gamma_i(H_i, a_i^*; \psi) - \gamma_i(H_i, a_i; \psi)\right],$$

(a_i^* is the optimal decision at ith stage). Consistent estimators of ψ can be obtained solving the estimating equations $\mathbb{E}[U(\psi, \zeta(\psi), \alpha)] = 0$, as shown in [111].

Both of the modifications discussed for Q-learning and value search must be applied when learning fair optimal policies by g-estimation. Specifically, we determine optimal polities not from the SNMM contrast $\gamma_k(H_k, a_k; \psi) = \mathbb{E}[Y(\bar{a}_{k-1}, a_k, f^*_{\underline{A}_{k+1}}) - Y(\bar{a}_{k-1}, a_k = 0, f^*_{\underline{A}_{k+1}}) \mid H_k]$ itself, but rather from a modified contrast

$$\begin{aligned}\gamma_k^*(H_k \setminus M, a_k; \psi) &= \sum_{m,s} \gamma_k(H_k, a_k; \psi) p^*(M|S, X) p^*(S|X) \\ &= \mathbb{E}[Y(\bar{a}_{k-1}, a_k, f^*_{\underline{A}_{k+1}}) - Y(\bar{a}_{k-1}, a_k = 0, f^*_{\underline{A}_{k+1}}) \mid H_k \setminus \{M, S\}]\end{aligned}$$

which does not use M and S. This is analogous to removing M and S from the Q-functions defined in Section 4 and is done for the same reason: M, S are drawn from $p(Z)$, not $p^*(Z)$.

Second, the estimating equations for ψ must use constrained models (in particular for M and S), and must be empirically solved using observations only from $p^*(Z)$. As was done with value search, we solve equation (V.1) empirically using a dataset where each row x_n, s_n, m_n is replaced by I rows of the form x_n, s^*_{ni}, m^*_{ni}, $i = 1, \ldots, I$, with s^*_{ni} and m^*_{ni} drawn from $p^*(S|x_n; \alpha_s)$ and $p^*(M|x_n, S; \alpha_m)$, respectively.

B. Details and Additional Results

Simulations

Here we report the precise parameter settings used in our simulation studies. The following regression models were used in our simulation study of the two-stage decision

problem:

$$
\begin{aligned}
X_1 &\sim |\mathcal{N}(0,1)| \\
(X_2, X_3) &\sim \mathcal{N}(0, \text{diag}(2)) \\
S &\sim \text{Bernoulli}(p = 0.5) \\
\text{logit}(p(M = 1)) &\sim -1 + X_1 + X_2 + X_3 + S + 3SX_1 + SX_2 + SX_3 \\
\text{logit}(p(A_1 = 1)) &\sim 1 - X_1 + X_2 + S + M - SX_1 + SX_2 + MS - 3MX_1 + 0.5MX_2 \\
\text{logit}(p(Y_1 = 1)) &\sim -2 + X_1 + X_2 + S + M + A + SX_2 + MS + AS + AM \\
\text{logit}(p(A_2 = 1)) &\sim 1 - X_1 + X_2 + M + A + W + S(1 - X_1 + X_2 + M - A) \\
&\quad - 3MX_1 + 0.5MX_2 - AX_1 - AX_2 \\
Y &= 2.5 + X_1 + X_2 + M + W + B + S(1 + X_1 + X_2 + M + A + W) \\
&\quad + A(1 + M - 2W) + MW + B(-X_1 + 2X_2 - M) + WX_1 + \mathcal{N}(0,1)
\end{aligned}
$$

For this two-stage setting we estimated the optimal policies using Q-learning and value search. In value search, we considered restricted class of polices of the form $p(A_1 = 1|X, S, M) = -1 + \alpha_x X + \alpha_s S + \alpha_m M + \alpha_{sx} SX + \alpha_{sm} SM + \alpha_{mx} MX$, and $p(A_2 = 1|X, S, M, A_1, Y_1) = -1 + \alpha_x X + \alpha_s S + \alpha_m M + \alpha_a A + \alpha_{y_1} Y_1 + \alpha_{sx} SX + \alpha_{sm} SM + \alpha_{mx} MX + \alpha_{as} AS + \alpha_{ax} AX$ where all αs range from -3 to 3 by 0.5 increments and estimated the value of policies for each combination of αs using equation (3.9).

A third method for estimating policies is to directly model the counterfactual contrasts known as *optimal blip-to-zero functions* and then learn these functions by g-estimation [111]; see Appendix A. We implemented our modified fair g-estimation for a single-stage decision problem and compared the results with Q-learning and value search. The results are provided in Table 1. The data generating process for the single-stage decision problem matches the causal model shown in Fig. 3-3(a) where X, S, M, and A were generated the same way as described above. The outcome Y was generated from a standard normal distribution with mean $-2 + X + S + M + A - 3SX_2 + MS + AS + AM + AX_2 + AX_3$. We used estimators in Theorem 5 to compute PSE^{sy} and PSE^{sa} which require using M and S models. In this synthetic data, the PSE^{sy} was 1.618 (on the mean scale) and was restricted to lie between -0.1 and 0.1. The PSE^{sa} was 0.685 (on the odds ratio scale) and was restricted to lie between 0.95 and 1.05.

Table V-I: Comparison of population outcomes $\mathbb{E}[Y]$ under policies learned by different methods. The value under the observed policy was 0.24±0.006.

	Unfair Policy	**Fair Policy**
Q-learning	1.414±0.0056	1.189±0.0059
value search	1.134±0.0245	1.056±0.0299
g-estimation	1.375±0.0099	1.312±0.0102

The COMPAS Dataset

The regression models we used in the COMPAS data analysis were specified as follows:

$$\text{logit}(p(M = 1)) \sim X_1 + X_2 + S + SX_1 + SX_2$$

$$\text{logit}(p(A = 1)) \sim X_1 + X_2 + S + M + MS + (M + S)(X_1 + X_2)$$

$$Y \sim X_1 + X_2 + S + M + A + AS + AM + MS + (S + M + A)(X_1 + X_2)$$

For estimating the PSEs which we constrain, we used the same IPW estimators described in the main chapter and reproduced in the theorem below. We constrained the PSEs to lie between -0.05 and 0.05 and 0.95 and 1.05, respectively.

In Fig. V-1, we compare the overall incarceration rates recommended by the optimal fair and unconstrained policies on the COMPAS data, as a function of the utility parameter θ. For low values of θ the incarceration rate is zero, and becomes higher as θ increases, but differentially for the fair and unconstrained optimal policies. The difference between the policies depends crucially on the utility function. For some values of the utility parameter, the unfair and fair policies coincide, but for other values we would expect significantly different overall incarceration rates as well as different disparities between racial groups (see result in the main chapter).

In Fig. V-2, we show the relative utility achieved by the optimal fair and unconstrained policies, as well as the utility of the observed decision pattern, as a function of θ. As expected, choosing an optimal policy improves on the observed policy, with the unfair (unconstrained) choice being higher utility than the fair (constrained) choice;

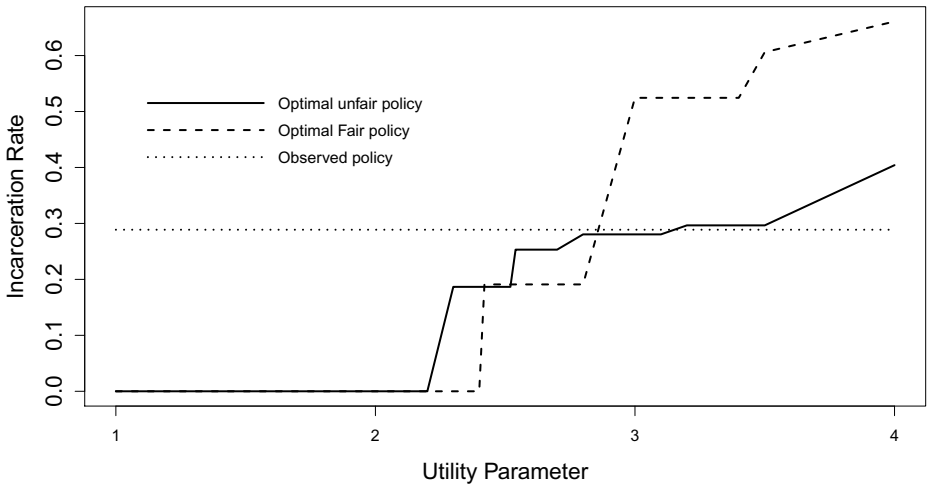


Figure V-1: Overall incarceration rates for the COMPAS data as a function of the utility parameter θ.

we sacrifice some optimality to satisfy the fairness constraints. However, the difference depends on the utility parameter and for a range of parameter values the fair and unfair policies are nearly the same in terms of optimality (even when they may disagree on the resulting incarceration rate, around $\theta = 2.6$). The fair and unfair policies drift far apart in terms of utility around $\theta = 3$, when the policies recommend an incarceration rate comparable to or higher than the observed rate.

C. Multiple Sets of Mediators

In the main chapter, we discussed a K-stage decision problem with one set of permissible mediators, M. Here, we extend those results to the setting where we have multiple sets of mediators $M_1, \ldots, M_K$, i.e., a DAG with topological ordering $X, S, M_1, A_1, Y_1, \ldots, M_K, A_K, Y_K$. In this case, we consider the following paths impermissible: PSE^{sy}, representing the effect of S on Y along all paths *other than* the paths of the form $S \to M_k \to \ldots \to Y$ $(\forall k)$; and PSE^{sa_k}, representing the effect of S on A_k along all paths *other than* the paths of the form $S \to M_j \to \ldots \to A_k$

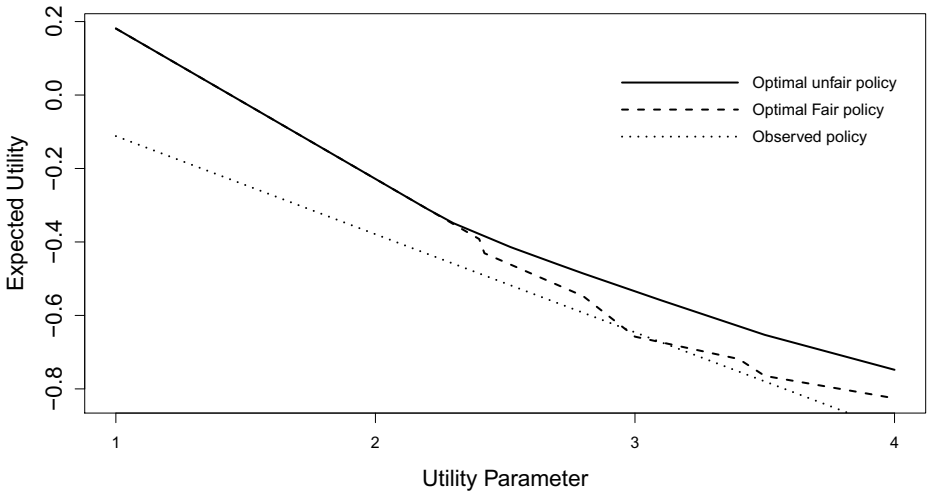


Figure V-2: The relative utility of policies for the COMPAS data as a function of the utility parameter θ.

($\forall j \leq k$). That is, we consider *only* pathways connecting S and A_k or Y through the allowed mediators $M_1, \ldots, M_K$ to be fair. In this case, the PSEs are identified by a modification of the previous formula given in Section 3.2.2.

$$
\begin{aligned}
\text{PSE}^{sy} &= \mathbb{E}[Y(s, M_1(s'), \ldots, M_K(s'))] - \mathbb{E}[Y(s')] \\
&= \sum_{x, \overline{m}_K, \overline{a}_{K-1}, \overline{y}_{K-1},} \{\mathbb{E}[Y|s, \overline{M}_K, \overline{A}_{K-1}, \overline{Y}_{K-1}, X] \\
&\qquad - \mathbb{E}[Y|s', \overline{M}_K, \overline{A}_{K-1}, \overline{Y}_{K-1}, X]\} \prod_{k=1}^{K} p(M_k|s', \overline{A}_{k-1}, \overline{Y}_{k-1}, X) \\
&\qquad \times \prod_{k=1}^{K-1} p(A_k|s, \overline{M}_k, \overline{A}_{k-1}, \overline{Y}_k, X) p(Y_k|s, \overline{M}_k, \overline{A}_k, \overline{Y}_{k-1}, X) p(X),
\end{aligned}
$$

$$
\begin{aligned}
\text{PSE}^{sa_k} &= \mathbb{E}[A_k(s, M_1(s'), \ldots, M_K(s'))] - \mathbb{E}[A_k(s')] \\
&= \sum_{x, \overline{m}_k, \overline{a}_{k-1}, \overline{y}_{k-1},} \{\mathbb{E}[A_k|s, \overline{M}_k, \overline{A}_{k-1}, \overline{Y}_{k-1}, X] \\
&\qquad - \mathbb{E}[A_k|s', \overline{M}_k, \overline{A}_{k-1}, \overline{Y}_{k-1}, X]\} \prod_{k=1}^{K} p(M_k|s', \overline{A}_{k-1}, \overline{Y}_{k-1}, X) \\
&\qquad \times \prod_{j=1}^{k-1} p(A_j|s, \overline{M}_j, \overline{A}_{j-1}, \overline{Y}_j, X) p(Y_j|s, \overline{M}_j, \overline{A}_j, \overline{Y}_{j-1}, X) p(X).
\end{aligned}
$$

With these definitions, we can replace the estimators in Theorem 1 with:

$$
\widehat{g}^{sy}(Z) = \frac{1}{N} \sum_{n=1}^{N} \left\{ \frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)} \prod_{k=1}^{K} \frac{p(M_{k,n}|s', \overline{A}_{k-1,n}, \overline{Y}_{k-1,n}, X_n)}{p(M_{k,n}|s, \overline{A}_{k-1,n}, \overline{Y}_{k-1,n}, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)} \right\} Y_n
$$

$$
\widehat{g}^{sa_k}(Z) = \frac{1}{N} \sum_{n=1}^{N} \left\{ \frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)} \prod_{k=1}^{K} \frac{p(M_{k,n}|s', \overline{A}_{k-1,n}, \overline{Y}_{k-1,n}, X_n)}{p(M_{k,n}|s, \overline{A}_{k-1,n}, \overline{Y}_{k-1,n}, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)} \right\} A_{kn}
$$

Then, in Theorem 2 we analogously define $\tilde{p}(Z)$ as follows:

$$
\tilde{p}(Z) \equiv p(X) p^*(S|X; \alpha_s) \prod_{k=1}^{K} \Big\{ p^*(M_k|S, \overline{A}_{k-1}, \overline{Y}_{k-1}, X; \alpha_m) \times p(A_k|H_k) p(Y_k|A_k, H_k) \Big\}.
$$

In this case we constrain the S and M_k models $\forall k$, the rest of the procedure remaining the same. Aside from the form of the identifying functional, the proofs of modified versions of the theorems are analogous.

D. Proofs

THEOREM 4 *Assume the observed data distribution $p(Z)$ is induced by a causal model where $Z = \{Y, C, S, M\}$ includes baseline measures C, binary sensitive feature S, and a set of mediators M, between S and Y. Let $p(Y(\pi, s, s'))$ denote the potential outcome distribution that corresponds to the effect of S on Y along unfair causal paths in π, where π includes the direct edge $S \to Y$, and let $p(Y_0(\pi, s, s'))$ denote the identifying functional for $p(Y(\pi, s, s'))$ obtained from the edge g-formula, where the term $p(Y \mid Z)$ is evaluated at $\{Z \setminus S\} = 0$. Then $\mathbb{E}[Y \mid Z]$ can be written as follows:*

$$\mathbb{E}[Y \mid Z] = f(Z) - \Big(\mathbb{E}[Y(\pi, s, s')] - \mathbb{E}[Y_0(\pi, s, s')]\Big) \; + \; \phi(S),$$

where $f(Z) := \mathbb{E}[Y \mid Z] - \mathbb{E}[Y \mid S, \{Z \setminus S\} = 0]$ and $\phi(S) = w_0 + w_s S$. Furthermore, w_s corresponds to π-specific effect of S on Y.

Proof. By letting $\phi(A = a) = \mathbb{E}[Y(\pi, a, a')]$, it suffices to show that $\mathbb{E}[Y_0(\pi, a, a')] = \mathbb{E}[Y|A, \{Z \setminus A\} = 0]$. Given the identification result for edge-consistent counterfactuals in [91], we can write the identification functional as follows.

$$\mathbb{E}[Y_0(\pi, a, a')] = \sum_{V \in \mathfrak{X}_V \setminus \{A, Y\}} \mathbb{E}[Y|A = a, \{Z \setminus A\} = 0] \times h(V \in \mathfrak{X}_V \setminus Y),$$

where $h(V \in \mathfrak{X}_V \setminus Y)$ is a function of all variables excluding Y. Note that h, does not include any density where A appears on the LHS of the conditioning bar. Therefore, we have:

$$\begin{aligned}\mathbb{E}[Y_0(\pi, s, s')] &= \mathbb{E}[Y \mid S = s, \{Z \setminus S\} = 0] \times \sum_{V \in \mathfrak{X}_V \setminus \{S, Y\}} h(V \in \mathfrak{X}_V \setminus Y) \\ &= \mathbb{E}[Y \mid S = s, \{Z \setminus S\} = 0].\end{aligned}$$

□

THEOREM 5 *Assume S is binary. Under the causal model above, the followings are consistent estimators of PSE^{sy} and PSE^{sa_k}, assuming all models are correctly specified:*

$$\widehat{g}^{sy}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)} \times \frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} \times Y_n,$$

$$\widehat{g}^{sa_k}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)} \times \frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} \times A_{kn}.$$

Proof. The latent projection [155] of any K stage DAG onto X, S, M, A, Y suffices to identify and estimate the two path-specific effects in question, and this latent projection is the complete DAG with topological ordering X, S, M, A, Y. The consistency of the estimators above then follows directly from derivations in [156]. As an example, we have the following derivation for the first term of $g^{sy}(Z)$:

$$\begin{aligned}
&\sum_{X,M} \mathbb{E}\left[Y \mid s, M, X\right] \times p(M \mid S = s', X) \times p(X) \\
&\quad = \sum_{X,M,A,Y} Y \times p(Y \mid S = s, M, A, X) \times p(A \mid S = s, M, X) \times p(M \mid S = s', X) \times p(X) \\
&\quad = \sum_{X,S,M,A,Y} \frac{\mathbb{I}(S = s)}{p(S \mid X)} \times \frac{p(M \mid S = s', X)}{p(M \mid S = s, X)} \times Y \times p(Y, S, M, A, X) \\
&\quad = \mathbb{E}\left[\frac{\mathbb{I}(S = s)}{p(S \mid X)} \times \frac{p(M \mid S = s', X)}{p(M \mid S = s, X)} Y\right],
\end{aligned}$$

which is precisely the identifying functional for the first term of the PSE we are interested in. That the above estimator is consistent for this functional is a standard result. □

THEOREM 6 *Consider the K-stage decision problem described by the DAG in Fig. 3-3(c). Let* $p^*(M \mid S,, X; \alpha_m)$ *and* $p^*(S \mid X; \alpha_s)$ *be the constrained models chosen to satisfy* $PSE^{sy} = 0$ *and* $PSE^{sa_k} = 0$. *Let* $\widetilde{p}(Z)$ *be the joint distribution induced by* $p^*(M \mid S, X; \alpha_m)$ *and* $p^*(S \mid X; \alpha_s)$, *and where all other distributions in the factorization are unrestricted. That is,*

$$\widetilde{p}(Z) \equiv p(X) \times p^*(S \mid X; \alpha_s) \times p^*(M \mid S, X; \alpha_m) \times \prod_{k=1}^{K} p(A_k \mid H_k) \times p(Y_k \mid A_k, H_k).$$

Then the functionals PSE^{sy} *and* PSE^{sa_i} *taken w.r.t.* $\widetilde{p}(Z)$ *are also zero.*

Proof. Let $Y \equiv Y_K$. Because M preceeds all A_k, Y_k for $k = 1, \ldots K$, it suffices to consider the latent projection with only variables X, S, M, A, Y without affecting identifiability considerations. Then we have the following:

$$\begin{aligned}
\widetilde{\mathrm{PSE}}^{sy} &= \widetilde{\mathbb{E}}[\, Y(s, M(s')) \,] - \widetilde{\mathbb{E}}[\, Y(s') \,] \\
&= \sum_{X,M} \{\widetilde{\mathbb{E}}[Y \mid s, M, X] - \widetilde{\mathbb{E}}[Y \mid s', M, X]\} \times p^*(M \mid s', X; \alpha_m) \times p(X) \\
&= \sum_{X,M} \{\mathbb{E}[Y \mid s, M, X] - \mathbb{E}[Y \mid s', M, X]\} \times p^*(M \mid s', X; \alpha_m) \times p(X) \\
&= \sum_{X,M,Y} Y \times \{p(Y \mid s, M, X) - p(Y \mid s', M, X)\} \times p^*(M \mid s', X; \alpha_m) \times p(X) \\
&= \sum_{X,S,M,Y} Y \times \Big\{ \frac{\mathbb{I}(S = s)}{p^*(S \mid X; \alpha_s)} \times \frac{p^*(M \mid s', X; \alpha_m)}{p^*(M \mid s, X; \alpha_m)} - \frac{\mathbb{I}(S = s')}{p^*(S \mid X; \alpha_s)} \Big\} \\
&\qquad \times p(Y \mid M, S, X) \times p^*(M \mid S, X; \alpha_m) \times p^*(S \mid X; \alpha_s) \times p(X) \\
&= 0,
\end{aligned}$$

by choice of $p^*(M \mid S, X; \alpha_m)$ and $p^*(S \mid X; \alpha_s)$. The proof is structurally the same for $\widetilde{\mathrm{PSE}}^{sa_k}$. □

Appendix VI

Supplementary Materials for Missing Data

A. Parameterization of Missing Data ADMGs

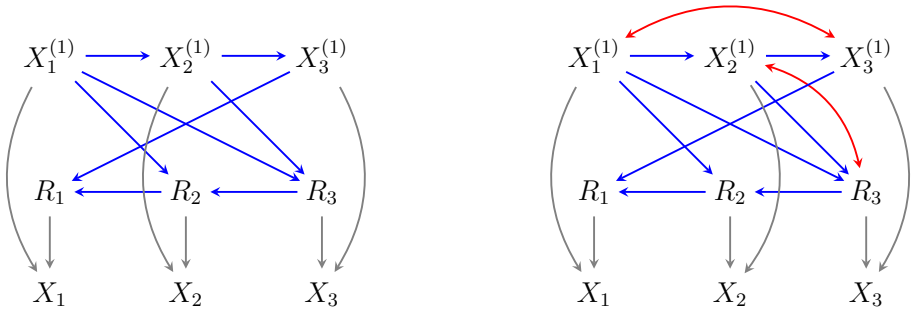
We summarize the necessary concepts required in order to explain our proof of completeness for identification of the full law in missing data acyclic directed mixed graphs (ADMGs). These concepts draw on the binary parameterization of nested Markov models of an ADMG, described in Appendix I. It is shown in [27] that the nested Markov model [29] of an ADMG $\mathcal{G}(V)$ is a smooth super model with fixed dimension, of the underlying latent variable model, that captures all equality constraints and avoids non-regular asymptotics arising from singularities in the parameter space [26, 27]. We use this fact in order to justify the use of nested Markov models of a missing data ADMG in order to describe full laws that are Markov relative to a missing data DAG with hidden variables. That is, the nested Markov model of a missing data ADMG $\mathcal{G}(V)$, where $V = \{O, X^{(1)}, R, X\}$, is a smooth super model of the missing data DAG model $\mathcal{G}(V \cup U)$. We also utilize nested Markov models of an ADMG $\mathcal{G}(V \setminus X^{(1)})$, corresponding to projection of the missing data ADMG $\mathcal{G}(V)$ onto variables that are fully observable. While such a model does not capture all equality constraints in the true observed law, it is still a smooth super model of it, thus providing an *upper bound* on the model dimension of the observed law.

We use the Moebius parameterization [150] in order to count the number of parameters required to parameterize the full law of a missing data ADMG and its corresponding observed law. We then use this to reason that if the number of parameters in the full law exceeds those in the observed law, it is impossible to establish a map from the observed law to the full law. This in turn implies that such a full law is not identified.

The binary parameterization of the **full law** of a missing data ADMG $\mathcal{G}(X^{(1)}, O, R, X)$ is exactly the same as that of an ordinary ADMG, except that the deterministic factors $p(X_i \mid R_i, X_i^{(1)})$, can be ignored, as $X_i = X_i^{(1)}$ with probability one when $R_i = 1$, and $X_i =?$ with probability one when $R_i = 0$.

The **observed law** is parameterized as follows. First, variables in $X^{(1)}$ are treated as completely unobserved, and an observed law ADMG $\mathcal{G}(X, O, R)$ is obtained by applying the latent projection operator to $\mathcal{G}(X^{(1)}, O, R, X)$. The Moebius parameters are then derived in a similar manner as before, with the additional constraint that if $X_i \in X$ appears in the head of a Moebius parameter, and the corresponding missingness indicator R_i appears in the tail, then the kernel must be restricted to cases where $R_i = 1$. This is because when $R_i = 0$, the probability of the head taking on any value, aside from those where $X_i =?$, is deterministically defined to be 0.

Note that parameterizing the observed law by treating variables in $X^{(1)}$ as fully unobserved does not quite capture all equality constraints that may be detectable in the observed law, as these variables are, in fact, sometimes observable when their corresponding missingness indicators are set to one. Indeed, a smooth parameterization of the observed law of missing data models that captures all constraints implied by the model, is still an open problem. Nevertheless, parameterizing an observed law ADMG, such as the one mentioned earlier, provides an *upper bound* on the number of parameters required to parameterize the true observed law. This suffices for our purposes, as demonstrating that the upper bound on the number of parameters in



(a) $\mathcal{G}_a$ (b) $\mathcal{G}_b$

Figure VI-1: (a) The missing data DAG model used in Scenario 2. (b) the missing data ADMG model used in Scenario 3.

the observed law is less than the number of parameters in the full law, is sufficient to prove that the full law is not identified.

B. Example: Odds Ratio Parameterization

To build up a more concrete intuition for Theorems 7 and 9, we provide an example of the odds ratio parameterization for the missing data models used in Scenarios 2 and 3 of the main chapter, reproduced here in Figs. VI-1(a, b). Utilizing the order R_1, R_2, R_3 on the missingness indicators, the odds ratio parameterization of the missing data process for both models is as follows.

$$\frac{1}{Z} \times \left(\prod_{k=1}^{3} p(R_i \mid R_{-i} = 1, X^{(1)}) \right) \times \text{OR}(R_1, R_2, \mid R_3 = 1, X^{(1)}) \times \text{OR}(R_3, (R_1, R_2) \mid X^{(1)}). \tag{VI.1}$$

We now argue that each piece in Eq. VI.1 is identified. Note that, in the missing data DAG shown in Fig. VI-1(a), $R_i \perp\!\!\!\perp X_i^{(1)} \mid R_{-i}, X_{-i}^{(1)}$ by d-separation. The same is true for the missing data ADMG in Fig. VI-1(b) by m-separation. Thus, in both cases, the product over conditional pieces of each R_i given the remaining variables is not a function $X_i^{(1)}$, and is thus a function of observed data. We now show that $\text{OR}(R_1, R_2 \mid R_3 = 1, X^{(1)})$ is not a function of $X_1^{(1)}, X_2^{(1)}$ by utilizing the symmetry

property of the odds ratio.

$$\begin{aligned}
\mathrm{OR}(R_1, R_2 \mid R_3 = 1, X^{(1)}) &= \frac{p(R_1 \mid R_2, R_3 = 1, X_2^{(1)}, X_3^{(1)})}{p(R_1 = 1 \mid R_2, R_3 = 1, X_2^{(1)}, X_3^{(1)})} \\
&\times \frac{p(R_1 = 1 \mid R_2 = 1, R_3 = 1, X_2^{(1)}, X_3^{(1)})}{p(R_1 \mid R_2 = 1, R_3 = 1, X_2^{(1)}, X_3^{(1)})} \\
\mathrm{OR}(R_2, R_1 \mid R_3 = 1, X^{(1)}) &= \frac{p(R_2 \mid R_1, R_3 = 1, X_1^{(1)}, X_3^{(1)})}{p(R_2 = 1 \mid R_1, R_3 = 1, X_1^{(1)}, X_3^{(1)})} \\
&\times \frac{p(R_2 = 1 \mid R_1 = 1, R_3 = 1, X_1^{(1)}, X_3^{(1)})}{p(R_2 \mid R_1 = 1, R_3 = 1, X_1^{(1)}, X_3^{(1)})}.
\end{aligned}$$

Thus, from the first equality, the odds ratio is not a function of $X_2^{(1)}$ as $R_1 \perp\!\!\!\perp X_1^{(1)} \mid R_{-1}, X_{-1}^{(1)}$ by d-separation in Fig. VI-1(a) and by m-separation in Fig. VI-1(b). A symmetric argument holds for $X_2^{(1)}$ and R_2 as seen in the second and third equalities. Hence, the odds ratio is only a function of $X_3^{(1)}$, which is observable, as the function is evaluated at $R_3 = 1$.

We now utilize an identity from [157] in order to simplify the final term in Eq. VI.1. That is,

$$\begin{aligned}
\mathrm{OR}(R_3, (R_1, R_2) \mid X^{(1)}) &= \mathrm{OR}(R_3, R_2 \mid R_1 = 1, X^{(1)})\ \mathrm{OR}(R_3, R_1 \mid R_2, X^{(1)}) \\
&= \mathrm{OR}(R_3, R_2 \mid R_1 = 1, X^{(1)}) \times \mathrm{OR}(R_3, R_1 \mid R_2 = 1, X^{(1)}) \\
&\times \underbrace{\frac{\mathrm{OR}(R_3, R_1 \mid R_2, X^{(1)})}{\mathrm{OR}(R_3, R_1 \mid R_2 = 1, X^{(1)})}}_{f(R_1, R_2, R_3 \mid X^{(1)})}.
\end{aligned}$$

The first two pairwise odds ratio terms are functions of observed data using an analogous argument that draws on the symmetry property of the odds ratio and the conditional independence $R_i \perp\!\!\!\perp X_i \mid R_{-i}, X_{-i}^{(1)}$, as before. The final term $f(R_1, R_2, R_3 \mid X^{(1)})$, is a three-way interaction term on the odds ratio scale and can be expressed in three different ways as follows [157],

$$\frac{\mathrm{OR}(R_3, R_1 \mid R_2, X^{(1)})}{\mathrm{OR}(R_3, R_1 \mid R_2 = 1, X^{(1)})} = \frac{\mathrm{OR}(R_2, R_3 \mid R_1, X^{(1)})}{\mathrm{OR}(R_2, R_3 \mid R_1 = 1, X^{(1)})} = \frac{\mathrm{OR}(R_1, R_2 \mid R_3, X^{(1)})}{\mathrm{OR}(R_1, R_2 \mid R_3 = 1, X^{(1)})}.$$

From the first equality, we note by symmetry of the odds ratio and conditional independence that f is not a function of $X_1^{(1)}, X_3^{(1)}$. Similarly, from the second

equality, we note that f is not a function of $X_2^{(1)}, X_3^{(1)}$. Finally, from the third equality, we note that f is not a function of $X_1^{(1)}, X_2^{(1)}$. Therefore, f is not a function of $X_1^{(1)}, X_2^{(1)}, X_3^{(1)}$ and is identified.

The normalizing function Z, is a function of all the pieces that we have already shown to be identified, and is therefore also identified. Thus, the missing data mechanisms $p(R \mid X^{(1)})$, and consequently, the full laws corresponding to the missing data graphs shown in Figs. VI-1(a,b) are identified by Remark 2.

C. Proofs

We first prove Lemmas 7 and 8 as we use them in the course of proving Theorems 7 and 9. We start with Lemma 8, as the proof for Lemma 7 simplifies to a special case.

LEMMA 8 *A missing data model of an ADMG $\mathcal{G}$ that contains no colluding paths is a submodel of the itemwise conditionally independent nonresponse model described in [123, 124].*

Proof. The *complete Markov blanket* of a vertex V_i in an ADMG $\mathcal{G}$, denoted $\mathrm{mb}_{\mathcal{G}}^{c}(V_i)$ is the set of vertices such that $V_i \perp\!\!\!\perp V_{-i} \setminus \mathrm{mb}_{\mathcal{G}}^{c}(V_i) \mid \mathrm{mb}_{\mathcal{G}}^{c}(V_i)$ [2, 31]. In ADMGs, this set corresponds to the Markov blanket of V_i, its children, and the Markov blanket of its children. That is,

$$\mathrm{mb}_{\mathcal{G}}^{c}(V_i) \equiv \mathrm{mb}_{\mathcal{G}}(V_i) \cup \left(\bigcup_{V_j \in \mathrm{ch}_{\mathcal{G}}(V_i)} V_j \cup \mathrm{mb}_{\mathcal{G}}(V_j) \right) \setminus \{V_i\}.$$

Without loss of generality, we ignore the part of the graph involving the deterministic factors $p(X \mid X^{(1)}, R)$ and the corresponding deterministic edges, in the construction of the Markov blanket and complete Markov blanket of variables in a missing data graph $\mathcal{G}(X^{(1)}, O, R)$. We now show that the absence of non-deterministic colluder paths between a pair $X_i^{(1)}$ and R_i in $\mathcal{G}$ implies that $X_i^{(1)} \notin \mathrm{mb}_{\mathcal{G}}^{c}(R_i)$.

- $X_i^{(1)}$ is not a parent of R_i, as $X_i^{(1)} \to R_i$ is trivially a colluder path.
- $X_i^{(1)}$ is not in the district of R_i, as $X_i^{(1)} \leftrightarrow \cdots \leftrightarrow R_i$ is also a colluder path.

These two points together imply that $X_i^{(1)} \notin \text{mb}_{\mathcal{G}}(R_i)$. We now show that the union over children of R_i and their Markov blankets also exclude $X_i^{(1)}$.

- $X_i^{(1)}$ is not a child of R_i, as directed edges from R_i to variables in $X^{(1)}$ are ruled out by construction in missing data graphs.
- $X_i^{(1)}$ is also not in the district of any children of R_i, as $R_i \to \cdots \leftrightarrow X_i^{(1)}$ is a colluding path.
- $X_i^{(1)}$ is also not a parent of the district of any children of R_i, as $R_i \to \cdots \leftarrow X_i^{(1)}$ is a colluding path.

These three points together rule out the possibility that $X_i^{(1)}$ is present in the union over children and Markov blankets of children of R_i. Thus, we have shown that $X_i^{(1)} \notin \text{mb}_{\mathcal{G}}^c(R_i)$. This implies the following,

$$R_i \perp\!\!\!\perp V \setminus \{R_i, \text{mb}_{\mathcal{G}}^c(R_i)\} \mid \text{mb}_{\mathcal{G}}^c(R_i) \implies R_i \perp\!\!\!\perp X_i^{(1)} \mid \text{mb}_{\mathcal{G}}^c(R_i).$$

By semi-graphoid axioms (see for example, [32, 2]) this yields the conditional independence $R_i \perp\!\!\!\perp X_i^{(1)} \mid R_{-i}, X_{-i}^{(1)}, O$.

The same line of reasoning detailed above can be used for all $R_i \in R$, which then gives us the set of conditional independences implied by the no self-censoring model. That is,

$$R_i \perp\!\!\!\perp X_i^{(1)} \mid R_{-i}, X_{-i}^{(1)}, O, \quad \forall R_i \in R.$$

□

LEMMA 7 *A missing data model of a DAG $\mathcal{G}$ that contains no self-censoring edges and no colluders, is a submodel of the itemwise conditionally independent nonresponse model described in [123, 124].*

Proof. A DAG is simply a special case of an ADMG with no bidirected edges. Consequently the only two types of colluding paths, are self-censoring edges ($X_i^{(1)} \to R_i$) and colluder structures ($X_i^{(1)} \to R_j \leftarrow R_i$). Thus, the absence of these two structures in a missing data DAG $\mathcal{G}$, rules out all possible colluding paths. The rest of the proof then carries over straightforwardly from Lemma 8. □

THEOREM 7 *A full law $p(R, X^{(1)}, O)$ that is Markov relative to a missing data DAG $\mathcal{G}$ is identified if $\mathcal{G}$ does not contain edges of the form $X_i^{(1)} \to R_i$ (no self-censoring) and structures of the form $X_j^{(1)} \to R_i \leftarrow R_j$ (no colluders), and the stated positivity assumption holds. Moreover, the resulting identifying functional for the missingness mechanism $p(R \mid X^{(1)}, O)$ is given by the odds ratio parameterization provided in Eq. 4.2 of the main draft, and the identifying functionals for the target law and full law are given by Remarks 1 and 2.*

Proof. Given Eq. (4.2), we know that

$$p(R \mid X^{(1)}, O) = \frac{1}{Z} \times \prod_{k=1}^{K} p(R_k \mid R_{-k} = 1, X^{(1)}, O) \times \prod_{k=2}^{K} \mathrm{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X^{(1)}, O),$$

where $R_{-k} = R \setminus R_k, R_{\prec k} = \{R_1, \ldots, R_{k-1}\}, R_{\succ k} = \{R_{k+1}, \ldots, R_K\}$, and

$$\begin{aligned} &\mathrm{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X^{(1)}, O) \\ &= \frac{p(R_k \mid R_{\succ k} = 1, R_{\prec k}, X^{(1)}, O)}{p(R_k = 1 \mid R_{\succ k} = 1, R_{\prec k}, X^{(1)}, O)} \times \frac{p(R_k = 1 \mid R_{-k} = 1, X^{(1)}, O)}{p(R_k \mid R_{-k} = 1, X^{(1)}, O)}, \end{aligned}$$

and Z is the normalizing term and is equal to $\sum_r \{\prod_{k=1}^{K} p(r_k \mid R_{-k} = 1, X^{(1)}, O) \times \prod_{k=2}^{K} \mathrm{OR}(r_k, r_{\prec k} \mid R_{\succ k} = 1, X^{(1)}, O)\}$. If we can prove that all the pieces in this

factorization are identified, then the missingness process is identified and so is the full law. We provide the proof in two steps. Our proof is similar to the identification proof of the no self-censoring model given in [135].

For each $k \in 3, \ldots, K$, we can apply the following expansion to the odds ratio term. Without loss of generality we drop fully observed random variables O for brevity,

$$\begin{aligned}&\mathrm{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X^{(1)}) = \mathrm{OR}(R_k, R_{k-1} \mid R_{-(k,k-1)} = 1, X^{(1)}) \\ &\times \mathrm{OR}(R_k, R_{\prec k-2} \mid R_{\succ k} = 1, R_{k-1}, X^{(1)}).\end{aligned} \tag{VI.2}$$

This expansion can be applied inductively to the second term in the above product until $\mathrm{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X^{(1)})$ is expressed as a function of pairwise odds ratios and higher-order interaction terms. Applying the inductive expansion to each odds ratio term in $\prod_{k=2}^{K} \mathrm{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X^{(1)})$ we can re-express the identifying functional as,

$$\begin{aligned}p(R \mid X^{(1)}) = \frac{1}{Z} \times &\prod_{k=1}^{K} p(R_k \mid R_{-k} = 1, X^{(1)}) \\ &\times \prod_{R_k, R_l \in R} \mathrm{OR}(R_k, R_l \mid R_{-(k,l)} = 1, X^{(1)}) \\ &\times \prod_{R_k, R_l, R_m \in R} f(R_k, R_l, R_m \mid R_{-(k,l,m)} = 1, X^{(1)}) \\ &\times \prod_{R_k, R_l, R_m, R_n \in R} f(R_k, R_l, R_m, R_n \mid R_{-(k,l,m,n)} = 1, X^{(1)}) \times \cdots \times f(R_1, \ldots, R_K \mid X^{(1)}),\end{aligned} \tag{VI.3}$$

$$\tag{VI.4}$$

where Z is the normalizing constant as before, and each $f(\cdot \mid \cdot, X^{(1)})$ are 3-way, 4-way, up to K-way interaction terms. These interaction terms are defined as follows.

$$f(R_i, R_j, R_k \mid R_{-(i,j,k)} = 1, X^{(1)}) = \frac{\mathrm{OR}(R_i, R_j | R_k, R_{-(i,j,k)} = 1, X^{(1)})}{\mathrm{OR}(R_i, R_j | R_k = 1, R_{-(i,j,k)} = 1, X^{(1)})},$$

and

$$\begin{aligned}&f(R_i, R_j, R_k, R_l | R_{-(i,j,k,l)} = 1, X^{(1)}) \\ &\quad = \frac{\mathrm{OR}(R_i, R_j | R_k, R_l, R_{-(i,j,k,l)} = 1, X^{(1)})}{\mathrm{OR}(R_i, R_j | R_k = 1, R_l, R_{-(i,j,k,l)} = 1, X^{(1)})} \times \frac{\mathrm{OR}(R_i, R_j | R_k = 1, R_l = 1, R_{-(i,j,k,l)} = 1, X^{(1)})}{\mathrm{OR}(R_i, R_j | R_k, R_l = 1, R_{-(i,j,k,l)} = 1, X^{(1)})},\end{aligned}$$

and so on, up to

$$f(R_1,...,R_K \mid X^{(1)}) = \text{OR}(R_i,R_j|R_{-(i,j)},X^{(1)})\times$$
$$\times \frac{\prod\limits_{R_k,R_l\in R} \text{OR}(R_i,R_j|R_{(k,l)}=1,R_{-(i,j,k,l)},X^{(1)}) \prod\limits_{R_k,R_l,R_m,R_n\in R} \text{OR}(R_i,R_j|R_{(k,l,m,n)}=1,R_{-(i,j,k,l,m,n)},X^{(1)})\times\cdots}{\prod\limits_{R_k\in R} \text{OR}(R_i,R_j|R_k=1,R_{-(i,j,k)},X^{(1)}) \prod\limits_{R_k,R_l,R_m\in R} \text{OR}(R_i,R_j|R_{(k,l,m)}=1,R_{-(i,j,k,l,m)},X^{(1)})\times\cdots}.$$

Readers familiar with the clique potential factorization of Markov random fields may treat these interaction terms analogously [135]. We now show that each term in the above factorization is identified.

Step 1.

We start off by looking at the conditional pieces $p(R_k \mid R_{-k}=1,X^{(1)},O)$. Given Lemma. 7, we know that $R_k \perp\!\!\!\perp X_k^{(1)} \mid R_{-k},X_{-k}^{(1)},O$. Therefore, $p(R_k \mid R_{-k}=1,X^{(1)},O)=p(R_k \mid R_{-k}=1,X_{-k}^{(1)},O),\forall k$, is identified for all $R_k\in R$.

Step 2.

We now show that for any $R_k,R_l\in R$, the pairwise odds ratio $\text{OR}(R_k,R_l \mid R_{\{-(k,l)\}}=1,X^{(1)})$ given in Eq. (VI.4) is identified. We know that

$$\text{OR}(R_k,R_l \mid R_{-(k,l)}=1,X^{(1)}) = \text{OR}(R_k,R_l \mid R_{-(k,l)}=1,X_{-(k,l)},X_k^{(1)},X_l^{(1)}).$$

Consequently, if we can show that the odds ratio is neither a function of $X_k^{(1)}$ nor $X_l^{(1)}$, then we can safely claim that the odds ratio is only a function of observed data and hence is identified. We get to this conclusion by exploiting the symmetric notion in odds ratios.

$$\begin{aligned}\text{OR}(R_k,R_l \mid R_{-(k,l)}=1,X^{(1)}) &= \frac{p(R_k \mid R_l,R_{-(k,l)}=1,X^{(1)})}{p(R_k=1 \mid R_l,R_{-(k,l)}=1,X^{(1)})}\times\frac{p(R_k=1 \mid R_{-k}=1,X^{(1)})}{p(R_k \mid R_{-k}=1,X^{(1)})}\\ &= \frac{p(R_l \mid R_k,R_{-(k,l)}=1,X^{(1)})}{p(R_l=1 \mid R_k,R_{-(k,l)}=1,X^{(1)})}\times\frac{p(R_l=1 \mid R_{-l}=1,X^{(1)})}{p(R_l \mid R_{-l}=1,X^{(1)})}\end{aligned}$$

In the first equality, we can see that the odds ratio is not a function of $X_k^{(1)}$ since $R_k \perp\!\!\!\perp X_k^{(1)} \mid R_{-k},X_{-k}^{(1)}$. Similarly, from the second equality, we can see that the odds

ratio is not a function of $X_l^{(1)}$ since $R_l \perp\!\!\!\perp X_l^{(1)} \mid R_{-l}, X_{-l}^{(1)}$. Therefore, the pairwise odds ratios are all identified.

Finally we show that each of the higher-order interaction terms are identified. For each of these terms we need to show that they are not a function of missing variables with indices corresponding to indicators to the left of the conditioning bar. That is, we need to show that the 3-way interaction terms $f(R_k, R_l, R_m \mid R_{-(k,l,m)} = 1, X^{(1)})$ are not functions of $X_{(k,l,m)}^{(1)}$, the 4-way interaction terms $f(R_k, R_l, R_m, R_n \mid R_{-(k,l,m,n)} = 1, X^{(1)})$ are not functions of $X_{(k,l,m,n)}^{(1)}$, and so on until finally the K-way interaction term $f(R_1, \ldots, R_K \mid X^{(1)})$ is not a function of $X^{(1)}$.

Because of the way the odds ratio is defined, each $f(\cdot \mid \cdot, X^{(1)})$ is symmetric in the k arguments appearing to the left of the conditioning bar and can be rewritten in multiple equivalent ways. In particular, each k-way interaction term can be rewritten in $\binom{k}{2}$ ways for any choice of indices i, j of the missingness indicators that appear to the left of the conditioning bar. Each such representation allows us to conclude that $f(\cdot \mid \cdot, X^{(1)})$ is not a function of $X_i^{(1)}, X_j^{(1)}$. Combining all these together allows us to conclude that the k-way interaction term $f(\cdot \mid \cdot, X^{(1)})$ is not a function of the missing variables corresponding to the indicators appearing on the left of the conditioning bar.

As a concrete example, consider the 3-way interaction $f(R_1, R_2, R_3 \mid R_{-(1,2,3)} = 1, X^{(1)})$. We can write it down in three different ways as follows.

$$\begin{aligned} f(R_i, R_j, R_k \mid R_{-(1,2,3)} = 1, X^{(1)}) &= \frac{\mathrm{OR}(R_1, R_2 \mid R_{-(1,2,3)} = 1, R_3, X^{(1)})}{\mathrm{OR}(R_1, R_2 \mid R_{-(1,2,3)} = 1, R_3 = 1, X^{(1)})} \\ &= \frac{\mathrm{OR}(R_1, R_3 \mid R_{-(1,2,3)} = 1, R_2, X^{(1)})}{\mathrm{OR}(R_1, R_3 \mid R_{-(1,2,3)} = 1, R_2 = 1, X^{(1)})} \\ &= \frac{\mathrm{OR}(R_2, R_3 \mid R_{-(1,2,3)} = 1, R_1, X^{(1)})}{\mathrm{OR}(R_2, R_3 \mid R_{-(1,2,3)} = 1, R_1 = 1, X^{(1)})} \end{aligned}$$

From the first equality, we note that f is not a function of $X_1^{(1)}, X_2^{(1)}$. From the second equality, we note that f is not a function of $X_1^{(1)}, X_3^{(1)}$. From the third equality, we note that f is not a function of $X_2^{(1)}, X_3^{(1)}$. Therefore, f is not a function of $X_1^{(1)}, X_2^{(1)}, X_3^{(1)}$

and is identified. □

THEOREM 8 *The graphical condition of no self-censoring and no colluders, put forward in Theorem 7, is sound and complete for the identification of full laws $p(R, O, X^{(1)})$ that are Markov relative to a missing data DAG $\mathcal{G}$.*

Proof. Soundness is a direct consequence of Theorem 7. To prove completeness, it needs to be shown that in the presence of a self-censoring edge, or a colluder structure, the full law is no longer (non-parametrically) identified. A proof by counterexample of both these facts was provided in [131]. However, this can also be seen from the fact that self-censoring edges and colluders are special cases of the colluding paths that we prove results in non-identification of the full law in Lemma 9. □

THEOREM 9 *A full law $p(R, X^{(1)}, O)$ that is Markov relative to a missing data ADMG $\mathcal{G}$ is identified if $\mathcal{G}$ does not contain any colluding paths and the stated positivity assumption in Section 4.3 holds. Moreover, the resulting identifying functional for the missingness mechanism $p(R \mid X^{(1)}, O)$ is given by the odds ratio parametrization provided in Eq. 4.2 of the main draft.*

Proof. The proof strategy is nearly identical to the one utilized in Theorem 7, except the conditional independences $R_k \perp\!\!\!\perp X_k^{(1)} \mid R_{-k}, X_{-k}^{(1)}, O$ come from Lemma 8 instead of Lemma 7. □

LEMMA 9 *A full law $p(R, X^{(1)}, O)$ that is Markov relative to a missing data ADMG $\mathcal{G}$ containing a colluding path between any pair $X_i^{(1)} \in X^{(1)}$ and $R_i \in R$, is not identified.*

Proof. Proving the non-identifiability of missing data models of an ADMG $\mathcal{G}$ that contains a colluding path can be shown by providing two models $\mathcal{M}_1$ and $\mathcal{M}_2$ that

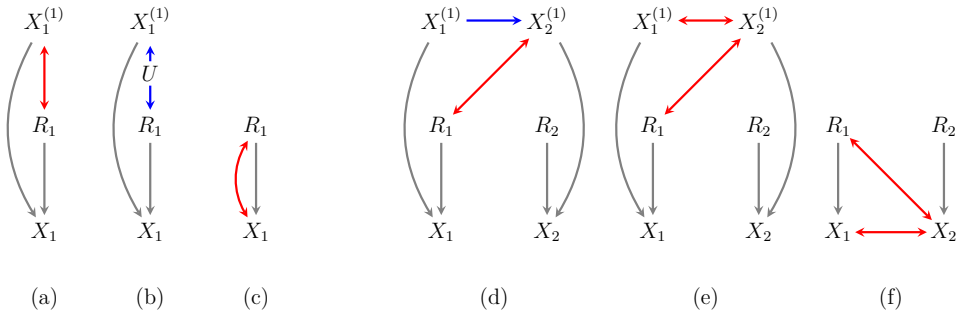


Figure VI-2: (a, d, e) Examples of colluding paths in missing data models of ADMGs. (b) A DAG with hidden variable U that is Markov equivalent to (a). (c) Projecting out $X_1^{(1)}$ from (a), (f) Projecting out $X_1^{(1)}$ and $X_2^{(1)}$ from (d) and (e).

disagree on the full law but agree on the observed law. Coming up with a single example of such a pair of models is sufficient for arguing against non-parametric identification of the full law. Therefore, for simplicity, we restrict our attention to binary random variables. We first provide an example of such a pair of models on the simplest form of a colluding path, a bidirected edge $X_i^{(1)} \leftrightarrow R_i$ as shown in Fig. VI-2(a). According to Table VI-I, in order for the observed laws to agree, the only requirement is that the quantity $ab + (1-a)c$ remain equal in both models; hence we can come up with infinitely many counterexamples of full laws that are not the same but map to the same observed law.

Constructing explicit counterexamples are not necessary to prove non-identification as long as it can be shown that there exist at least two distinct functions that map two different full laws onto the exact same observed law. For instance, if the number of parameters in the full law is strictly larger than the number of parameters in the observed law, then there would exist infinitely many such functions. Consequently, we rely on a parameter counting argument to prove the completeness of our results. Since we are considering missing data models of ADMGs, we use the Moebius parameterization of binary nested Markov models of an ADMG described in Appendix A.

U	$p(U)$
0	a
1	$1-a$

R_1	U	$p(R_1\|U)$
0	0	b
1	0	$1-b$
0	1	c
1	1	$1-c$

$X_1^{(1)}$	U	$p(X_1^{(1)}\|U)$
0	0	d
1	0	$1-d$
0	1	e
1	1	$1-e$

R_1	$X_1^{(1)}$	U	$p(R_1, X_1^{(1)}, U)$
0	0	0	$a*b*d$
0	0	1	$(1-a)*c*e$
0	1	0	$a*b*(1-d)$
0	1	1	$(1-a)*c*(1-e)$
1	0	0	$a*(1-b)*d$
1	0	1	$(1-a)*(1-c)*e$
1	1	0	$a*(1-b)*(1-d)$
1	1	1	$(1-a)*(1-c)*(1-e)$

R_1	$X_1^{(1)}$	p(Full Law)	X_1	p(Observed Law)
0	0	$a*b*d+(1-a)*c*e$	?	$a*b+(1-a)*c$
	1	$a*b*(1-d)+(1-a)*c*(1-e)$		
1	0	$a*(1-b)*d+(1-a)*(1-c)*e$	0	$a*(1-b)+(1-a)*(1-c)$
	1	$a*(1-b)*(1-d)+(1-a)*(1-c)*(1-e)$	1	

Table VI-I: Construction of counterexamples for non-identifiablity of the full law in Fig. VI-2(a) using the DAG with hidden variable U in Fig. VI-2(b) that is Markov equivalent to (a).

The nested Markov model of a missing data ADMG $\mathcal{G}(V)$, where $V = \{O, X^{(1)}, R, X\}$, is a smooth super model of the missing data DAG model $\mathcal{G}(V \cup U)$, and has the same model dimension as the latent variable model [27]. We also utilize nested Markov models of an ADMG $\mathcal{G}(V \setminus X^{(1)})$, corresponding to projection of the missing data ADMG $\mathcal{G}(V)$ onto variables that are fully observable. While such a model does not capture all equality constraints in the true observed law, it is still a smooth super model of it, thus providing an *upper bound* on the model dimension of the observed law. This suffices for our purposes, as demonstrating that the upper bound on the number of parameters in the observed law is less than the number of parameters in the full law, is sufficient to prove that the full law is not identified. We first walk the reader through a few examples to demonstrate this proof strategy, and then provide the general argument.

Moebius Parameterization of the Full Law in Fig. VI-2(d)			
Districts	Intrinsic Head/Tail	Moebius Parameters	Counts
$\{X_1^{(1)}\}$	$\{X_1^{(1)}\}, \{\}$	$q(X_1^{(1)} = 0)$	1
$\{R_2\}$	$\{R_2\}, \{\}$	$q(R_2 = 0)$	1
$\{R_1, X_2^{(1)}\}$	$\{R_1\}, \{\}$	$q(R_1 = 0)$	1
	$\{X_2^{(1)}\}, \{X_1^{(1)}\}$	$q(X_2^{(1)} = 0 \mid X_1^{(1)})$	2
	$\{R_1, X_2^{(1)}\}, \{X_1^{(1)}\}$	$q(R_1 = 0, X_2^{(1)} = 0 \mid X_1^{(1)})$	2
		Total	7
Moebius Parameterization of the Full Law in Fig. VI-2(e)			
Districts	Intrinsic Head/Tail	Moebius Parameters	Counts
$\{R_2\}$	$\{R_2\}, \{\}$	$q(R_2 = 0)$	1
$\{R_1, X_1^{(1)}, X_2^{(1)}\}$	$\{R_1\}, \{\}$	$q(R_1 = 0)$	1
	$\{X_1^{(1)}\}, \{\}$	$q(X_1^{(1)} = 0)$	1
	$\{X_2^{(1)}\}, \{\}$	$q(X_2^{(1)} = 0)$	1
	$\{R_1, X_2^{(1)}\}, \{\}$	$q(R_1 = 0, X_2^{(1)} = 0)$	1
	$\{X_1^{(1)}, X_2^{(1)}\}, \{\}$	$q(X_1^{(1)} = 0, X_2^{(1)} = 0)$	1
	$\{R_1, X_1^{(1)}, X_2^{(1)}\}, \{\}$	$q(R_1 = 0, X_1^{(1)} = 0, X_2^{(1)} = 0)$	1
		Total	7
Moebius Parameterization of the Observed Law in Fig. VI-2(f)			
Districts	Intrinsic Head/Tail	Moebius Parameters	Counts
R_2	$\{R_2\}, \{\}$	$q(R_2 = 0)$	1
$\{R_1, X_1, X_2\}$	$\{R_1\}, \{\}$	$q(R_1 = 0)$	1
	$\{X_1\}, \{R_1\}$	$q(X_1 = 0 \mid R_1)$	1
	$\{X_2\}, \{R_2\}$	$q(X_2 = 0 \mid R_2)$	1
	$\{R_1, X_2\}, \{R_2\}$	$q(R_1 = 0, X_2 = 0 \mid R_2)$	1
	$\{X_1, X_2\}, \{R_1, R_2\}$	$q(X_1 = 0, X_2 = 0 \mid R_1, R_2)$	1
		Total	6

Table VI-II: Moebius Parameterization of the Full and Observed Laws of missing data ADMGs

Self-censoring through unmeasured confounding:

We start by reanalyzing the colluding path given in Fig. VI-2(a) and the corresponding projection given in Fig. VI-2(c). The Moebius parameters associated with the full law are $q(X_1^{(1)} = 0), q(R_1 = 0), q(X_1^{(1)} = 0, R_1 = 1)$, for a total of 3 parameters. The Moebius parameters associated with the observed law in Fig VI-2(c) are $q(R_1 = 0), q(X_1^{(1)} = 0 \mid R_1 = 0)$, for a total of only 2 parameters. Since $2 < 3$, we can construct infinitely many mappings, as it was shown in Table VI-I.

Simple colluding paths:

Consider the colluding paths given in Fig. VI-2(d, e) and the corresponding projection (which are identical in both cases) given in Fig. VI-2(f). The Moebius parameters associated with the full laws and observed law are shown in Table VI-II. Once again, since the number of parameters in the observed law is less than the number in the full law ($6 < 7$), we can construct infinitely many mappings.

A general argument:

In order to generalize our argument, we first provide a more precise representation (that does not use dashed edges) in Figs. VI-3(a-d), of all possible colluding paths between $X_i^{(1)}$ and R_i. Without loss of generality, assume that there are K variables in $X^{(1)}$ and there are S variables that lie on the collider path between $X_i^{(1)}$ and R_i, $S \in \{0, 1, \ldots, 2 * (K-1)\}$. We denote the sth variable on the collider path by V_s; $V_s \in \{X^{(1)} \setminus X_i^{(1)}, R \setminus R_i\}$. Note that V_S in Figs. VI-3(c, d) can only belong to $\{R \setminus R_i\}$ by convention. Fig. VI-3(e) illustrates the corresponding projections of figures (a) and (b), and Fig. VI-3(f) illustrates the corresponding projections of figures (c) and (d). In the projections shown in Figs. VI-3(e, f), $V^* \in \{X \setminus X_i^{(1)}, R \setminus R_i\}$.

We now go over each of these colluding paths and their corresponding latent projections, as if they appear in a larger graph that is otherwise completely disconnected. We count the number of Moebius parameters as a function of S, and show that the full law always has one more parameter than the observed law. One can then imagine placing these colluding paths in a larger graph with arbitrary connectivity, and arguing that the full law is still not identified as a consequence of the parameter discrepancy arising from the colluding path alone. That is, if we show a fully disconnected graph containing a single colluding path is not identified, then it is also the case that any edge super graph (super model) is also not identified.

In the following proof we heavily rely on the following fact. Given a bidirected

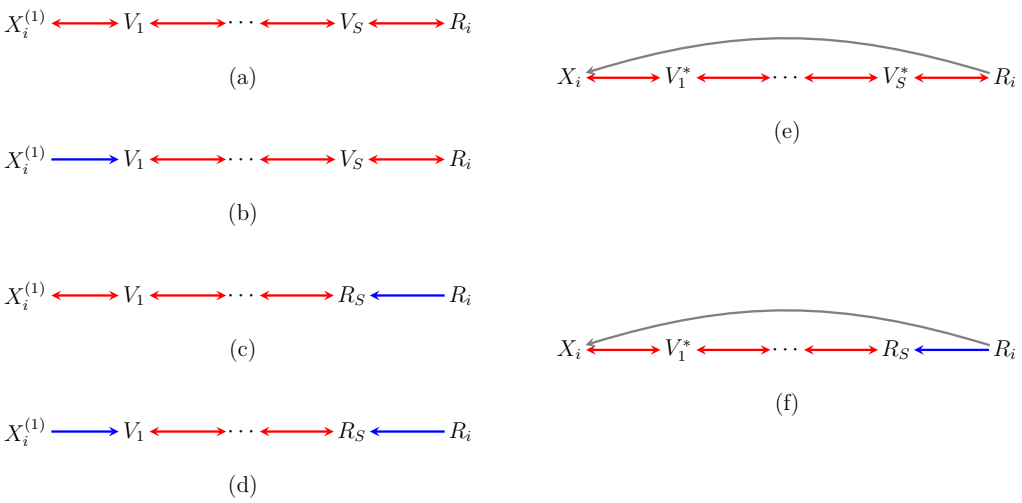


Figure VI-3: (a) Colluding paths (b) Projecting out $X^{(1)}$

chain of length $V_1 \leftrightarrow, \cdots, \leftrightarrow V_K$, of length K, the number of Moebius parameters required to parameterize this chain is given by the sum of natural numbers 1 to K, i.e., $\frac{K(K+1)}{2}$. This can be seen from the fact that the corresponding Moebius parameters are given by the series,

- $q(V_1 = 0), q(V_1 = 0, V_2 = 0), \ldots, q(V_1 = 0, \ldots, V_K = 0)$ corresponding to K parameters.
- $q(V_2 = 0), q(V_2, V_3 = 0), \ldots, q(V_2 = 0, \ldots, V_K = 0)$ corresponding to $K - 1$ parameters.
- $\ldots$
- $q(V_K = 0)$ corresponding to 1 parameter.

In counting the number of parameters for a disconnected graph (with the exception of the colluding path), we can also exclude the singleton (disconnected) nodes from the counting argument since they account for the same number of parameters in both the full law and observed law. In the full law they are either $q(R_s = 0)$ or $q(X_s^{(1)} = 0)$ and the corresponding parameters in the observed law are $q(R_s = 0)$ or $q(X_s = 0 \mid R_s = 1)$.

The Moebius parameter counts for each of the colluding paths in Figs. VI-3(a-d) and their corresponding latent projections in Figs. VI-3(e,f) are as follows.

Figures a, b, and e

1. Number of Moebius parameters in Fig. VI-3(a) is $\frac{(S+2)(S+3)}{2}$
   - A bidirected chain $X_i^{(1)} \leftrightarrow, \cdots, \leftrightarrow R_i$ of length $S+2$, i.e., $(S+2)*(S+3)/2$ parameters.
2. Number of Moebius parameters in Fig. VI-3(b) is $\frac{(S+2)(S+3)}{2}$
   - $q(X_i^{(1)} = 0)$, i.e. 1 parameter,
   - A bidirected chain $V_2 \leftrightarrow \cdots \leftrightarrow R_i$ of length S, i.e. $S*(S+1)/2$ parameters,
   - Intrinsic sets involving V_1, i.e., $q(V_1 = 0 \mid X_i^{(1)}), q(V_1 = 0, V_2 = 0 \mid X_i^{(1)}), q(V_1 = 0, \ldots, R_i = 0 \mid X_i^{(1)})$ corresponding to $2*(S+1)$ parameters.
3. Number of Moebius parameters in Fig. VI-3(e) is $\frac{(S+2)(S+3)}{2} - 1$
   - Note that even though each proxy X_s that may appear in the bidirected chain has a directed edge from R_s pointing into it, the corresponding intrinsic head tail pair that involves both variables, will always have $R_i = 1$. Hence, we may ignore these deterministic edges and count the parameters as if it were a bidirected chain $V_1^* \leftrightarrow \cdots \leftrightarrow R_i$ of length $S+1$, corresponding to $(S+1)*(S+2)/2$ parameters,
   - When enumerating intrinsic sets involving X_i, we note that $\{X_i, V_1^*, \ldots V_S^*\}$ is not intrinsic as R_i is not fixable (due to the bidirected path between R_i and X_i and the edge $R_i \to X_i$). Thus, as there is one less intrinsic set involving X_i, the number of parameters required to parameterize all intrinsic sets involving X_i is one fewer, i.e., $S+1$ (instead of $S+2$) parameters.

Figures c, d, and f

1. Number of Moebius parameters in Fig. VI-3(c) is $\frac{(S+2)(S+3)}{2}$

   - $q(R_i = 0)$, i.e. 1 parameter,
   - A bidirected chain $X_i^{(1)} \leftrightarrow \cdots \leftrightarrow V_{S-1}$ of length S, i.e. $S * (S+1)/2$ parameters,
   - Intrinsic sets involving R_S, i.e., $q(R_S = 0 \mid R_i), q(R_S = 0, V_{S-1} = 0 \mid R_i), \ldots, q(R_S = 0, V_{S-1} = 0 \ldots, X_i^{(1)} \mid R_i)$, corresponding to $2 * (S+1)$ parameters.

2. Number of Moebius parameters in Fig. VI-3(d) is $\frac{(S+2)(S+3)}{2}$

   - $q(X_i^{(1)} = 0), q(R_i = 0)$, i.e. 2 parameters,
   - A bidirected chain $V_2 \leftrightarrow \cdots \leftrightarrow V_{S-2}$ of length $S-2$, i.e. $(S-2) * (S-1)/2$ parameters,
   - Intrinsic sets involving V_1 and not R_S, i.e., $q(V_1 = 0 \mid X_i^{(1)}), q(V_1 = 0, V_2 = 0 \mid X_i^{(1)}), \ldots, q(V_1 = 0, V_2 = 0, \ldots, V_{S-1} \mid X_i^{(1)})$, corresponding to $2 * (S-1)$ parameters,
   - Intrinsic sets involving R_S and not V_1, i.e., $q(R_S = 0 \mid R_i), q(R_S = 0, V_{S-1} = 0 \mid R_i), \ldots, q(R_S = 0, V_{S-1} = 0, \ldots, V_2 \mid R_i)$ corresponding to $2 * (S-1)$ parameters.
   - The intrinsic set involving both V_1 and R_S, i.e., $q(V_1 = 0, V_2 = 0, \ldots, R_S = 0 \mid X_i^{(1)}, R_i)$, corresponding to 4 parameters.

3. Number of Moebius parameters in Fig. VI-3(f) is $\frac{(S+2)(S+3)}{2} - 1$

   - $q(R_i = 0)$, i.e. 1 parameter,
   - By the same argument as before, deterministic tails can be ignored. Hence, we have a bidirected chain $X_i \leftrightarrow \cdots \leftrightarrow V_{S-1}$ of length S, i.e. $S * (S+1)/2$ parameters,

- Intrinsic sets involving R_S, i.e.,

  $q(R_S = 0 \mid R_i), q(R_S = 0, V_{S-1} \mid R_i), \ldots, q(R_S, V_{S-1}, \ldots, V_1 \mid R_i)$, corresponding to $2 * S$ parameters, and the special intrinsic set which results in the observed law having one less parameter $q(R_S, V_{S-1}, \ldots, V_1, X_i \mid R_i = 1)$ corresponding to just 1 parameter instead of 2 due to the presence of the proxy X_i in the head and the corresponding R_i in the tail.

□

THEOREM 10 *The graphical condition of the absence of colluding paths, put forward in Theorem 9, is sound and complete for the identification of full laws $p(R, O, X^{(1)})$ that are Markov relative to a missing data ADMG $\mathcal{G}$.*

Proof. Soundness is a direct consequence of Theorem 9 and completeness is a direct consequence of Lemma. 9. □

Bibliography

[1] Jerzy Neyman. Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in English. *Statistical Science*, 5:463–472, 1923.

[2] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[3] James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[4] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[5] Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. *arXiv preprint arXiv:1912.00306*, 2019.

[6] James M Robins. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pages 113–159, 1989.

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

[8] Iván Díaz and Mark J van der Laan. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The international journal of biostatistics*, 9(2):149–160, 2013.

[9] Matteo Bonvini and Edward H Kennedy. Sensitivity analysis via the proportion of unmeasured confounding. *arXiv preprint arXiv:1912.02793*, 2019.

[10] Noam Finkelstein and Ilya Shpitser. Deriving bounds and inequality constraints using logical relations among counterfactuals. In *Conference on Uncertainty in Artificial Intelligence*, pages 1348–1357. PMLR, 2020.

[11] Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

[12] James M Robins, Andrea Rotnitzky, and Daniel O Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

[13] Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

[14] Vincent Dorie, Masataka Harada, Nicole Bohme Carnegie, and Jennifer Hill. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*, 35(20):3453–3470, 2016.

[15] Bo Zhang and Eric J Tchetgen Tchetgen. A semiparametric approach to model-based sensitivity analysis in observational studies. *arXiv preprint arXiv:1910.14130*, 2019.

[16] AlexanderM Franks, Alexander D?Amour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, pages 1–33, 2019.

[17] Weiwei Liu, S Janet Kuramoto, and Elizabeth A Stuart. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science*, 14(6):570–580, 2013.

[18] Amy Richardson, Michael G Hudgens, Peter B Gilbert, and Jason P Fine. Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):596, 2014.

[19] Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

[20] Ryo Okui, Dylan S. Small, Zhiqiang Tan, and James M. Robins. Doubly robust instrumental variable regression. *Statistica Sinica*, pages 173–205, 2012.

[21] Linbo Wang and Eric J. Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society Series B, Statistical Methodology*, 80(3):531, 2018.

[22] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[23] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

[24] James M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, The Environment, and Clinical Trials*, pages 95–133. Springer, 2000.

[25] Mark J. van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media, 2011.

[26] Mathias Drton. Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009.

[27] Robin J. Evans. Margins of discrete Bayesian networks. *The Annals of Statistics*, 46(6A):2623–2656, 2018.

[28] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, 1990.

[29] Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.

[30] Robin J. Evans and Thomas S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 25(2):848–876, 2019.

[31] Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[32] Steffen L. Lauritzen. *Graphical Models.* Oxford, U.K.: Clarendon, 1996.

[33] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pages 567–573. American Association for Artificial Intelligence, 2002.

[34] Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. arXiv: 2003.12659, 2020.

[35] Peter L. Spirtes, Clark N. Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas S. Richardson. *Causation, prediction, and search*. MIT press, 2000.

[36] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.

[37] Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 13–16, 2006.

[38] Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

[39] Aad W. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

[40] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[41] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[42] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[43] Isabel R. Fulcher, Ilya Shpitser, Stella Marealle, and Eric J. Tchetgen Tchetgen. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):199–214, 2020.

[44] Constantine E. Frangakis, Tianchen Qian, Zhenke Wu, and Iván Díaz. Deductive derivation and Turing-computerization of semiparametric efficient estimation. *Biometrics*, 71(4):867–874, 2015.

[45] Marco Carone, Alexander R. Luedtke, and Mark J. van der Laan. Toward computerized efficient estimation in infinite-dimensional models. *Journal of the American Statistical Association*, 114(527):1174–1190, 2019.

[46] M. Gentzkow, B. Kelley, and M. Taddy. Text as data. *Journal of Economic Literature*, 57:535–574, 2019.

[47] Joseph D Ramsey, Stephen José Hanson, Catherine Hanson, Yaroslav O Halchenko, Russell A Poldrack, and Clark Glymour. Six problems for causal inference from fmri. *neuroimage*, 49(2):1545–1558, 2010.

[48] Mara Mather, John T Cacioppo, and Nancy Kanwisher. How fmri can inform cognitive theories. *Perspectives on Psychological Science*, 8(1):108–113, 2013.

[49] S.P. Robertson, H. Quon, A. P. Kiess, J. A. Moore, W. Yang, Z. Cheng, S. Afonso, M. Allen, M. Richardson, A. Choflet, A. Sharabi, and T. R. McNutt. A data-mining framework for large scale analysis of dose-outcome relationships in a database of irradiated head and neck cancer patients. *Med Phys*, pages 4329–4337, 2015.

[50] Razieh Nabi, Todd McNutt, and Ilya Shpitser. Semiparametric causal sufficient dimension reduction of high dimensional treatments. arXiv: 1710.06727, 2020.

[51] KC. Li. Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86:316–342, 1991.

[52] L. Li. Sparse sufficient dimension reduction. *Biometrika*, 94:603–613, 2007.

[53] RD. Cook and S. Weisberg. Discussion of sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:28–33, 1991.

[54] B. Li and S. Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102:997–1008, 2007.

[55] LX. Zhu and KT. Fang. Asymptotics for kernel estimation of sliced inverse regression. *The Annals of Statistics*, 3:1053–1068, 1996.

[56] W. Hardle and TM. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84:986–995, 1989.

[57] H. Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58:71–120, 1993.

[58] RD. Cook and B. Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30:455–474, 2002.

[59] Y. Ma and L. Zhu. A semiparametric approach to dimension reduction. *Journal of American Statistical Association*, 107:168–179, 2012.

[60] J. M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*. NY: Springer-Verlag, 1999.

[61] A. Fisher and E. H. Kennedy. Visually communicating and teaching intuition for influence functions. *arxiv preprint: 1810.03260*, 2018.

[62] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals ofStatistics*, 10:1040–1053, 1982.

[63] S. Vansteelandt and M. Joffe. Structural nested models and g-estimation: The partially realized promise. *Statistical Science*, 29(4):707–731, 2014.

[64] Y. Dong and B. Li. Dimension reduction for non-elliptically distributed predictors: Second-order moments. *Biometrika*, 97:279–294, 2010.

[65] Z. Ye and R. E. Weiss. Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98:968–979, 2003.

[66] Catherine A Johnston, Thomas J Keane, and Susan M Prudo. Weight loss in patients receiving radical radiation therapy for head and neck cancer: a prospective study. *Journal of Parenteral and Enteral Nutrition*, 6(5):399–402, 1982.

[67] Jon Cacicedo, Francisco Casquero, Lorea Martinez-Indart, Olga Del Hoyo, Alfonso Gomez de Iturriaga, Arturo Navarro, and Pedro Bilbao. A prospective analysis of factors that influence weight loss in patients undergoing radiotherapy. *Chinese journal of cancer*, 33(4):204, 2014.

[68] Joseph O Deasy, Vitali Moiseenko, Lawrence Marks, KS Clifford Chao, Jiho Nam, and Avraham Eisbruch. Radiotherapy dose–volume effects on salivary gland function. *International Journal of Radiation Oncology* Biology* Physics*, 76(3):S58–S63, 2010.

[69] Serena Monti, Giuseppe Palma, Vittoria D'Avino, Marianna Gerardi, Giulia Marvaso, Delia Ciardo, Roberto Pacelli, Barbara A Jereczek-Fossa, Daniela Alterio, and Laura Cella. Voxel-based analysis unveils regional dose differences associated with radiation-induced morbidity in head and neck cancer patients. *Scientific Reports*, 7(1):1–8, 2017.

[70] Numair Sani, Jaron Lee, Razieh Nabi, and Ilya Shpitser. A semiparametric approach to interpretable machine learning. arXiv: 2006.04732, 2020.

[71] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. ProPublica, Machine bias: risk assessments in criminal sentencing, 2016.

[72] A. Barry-Jester, B. Casselman, and D. Goldstein. The Marshall Project, The new science of sentencing, 2015.

[73] D. Roberts. *Shattered Bonds: The color of child welfare*. Civitas Books, 2002.

[74] J. Gaulding. Race, sex, and genetic discrimination in insurance: What's fair. *Cornell Law Review*, 80, 1995.

[75] K. Petrasic, B. Saul, and M. Bornfreund. Algorithms and bias: What lenders need to know, 2017.

[76] C. Miller. The New York Times, Can an algorithm hire better than a human?, 2015.

[77] Kristian L. and William I. To predict and serve? *Significance*, 13(5):14–19, 2016.

[78] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the Conference on Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[79] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Optimal training of fair predictive models. arXiv: 1910.04109, 2020.

[80] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

[81] Walter L. Perry, Brian McInnis, Carter C. Price, Susan Smith, and John S. Hollywood. Predictive policing: The role of crime forecasting in law enforce-

ment operations. RAND Corporation, http://www.rand.org/pubs/research_reports/RR233.html, 2013.

[82] Teus H Kappen, Wilton A van Klei, Leo van Wolfswinkel, Cor J Kalkman, Yvonne Vergouwe, and Karel GM Moons. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic and Prognostic Research*, 2(1):11, 2018.

[83] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine learning algorithms. *Journal of Banking & Finance*, 34:2767–2787, 2010.

[84] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3929–3935, 2017.

[85] M. Bertrand and S. Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American Economic Review*, 94:991–1013, 2004.

[86] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 2017.

[87] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making – the causal explanation formula. In *Proceedings of the Thirty-Second AAAI Conference on Association for the Advancement of Artificial Intelligence*, 2018.

[88] Shira Mitchell, Eric Potash, and Solon Barocas. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

[89] Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420, 2001.

[90] Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science (Rumelhart special issue)*, 37:1011–1035, 2013.

[91] Ilya Shpitser and Eric J. Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of Statistics*, 44(6):2433–2466, 2016.

[92] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[93] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances In Neural Information Processing Systems*, pages 3315–3323, 2016.

[94] 7th Circuit Court. Carson vs Bethlehem Steel Corp., 1996. 70 FEP cases 921.

[95] Caleb Miles, Phyllis Kanki, Seema Meloni, and Eric Tchetgen Tchetgen. On partial identification of the pure direct effect. *Journal of Causal Inference*, 2016.

[96] Art Owen. *Empirical Likelihood*. Chapman & Hall, 2001.

[97] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[98] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266–298, 2010.

[99] Alan E. Gelfand, Adrian F. M. Smith, and Tai-Ming Lee. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, 87(418):523–532, 1992.

[100] M. Lichman. UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/adult, 2013.

[101] Tyler J. VanderWeele and Stijn Vansteelandt. Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, 172:1339–1348, 2010.

[102] Bibbas Chakraborty and Erica E. Moodie. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. New York: Springer-Verlag, 2013.

[103] Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT press, 1998.

[104] Dimitri P. Bertsekas and John Tsitsiklis. *Neuro-Dynamic Programming*. Athena Publishing, 1996.

[105] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568, 2008.

[106] Faisal Kamiran, Indre Zliobaite, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3):613–644, 2013.

[107] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 797–806, 2017.

[108] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, , and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of International Conference on Machine Learning*, 2017.

[109] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

[110] Dan Hurley. Can an algorithm tell when kids are in danger? The New York Times, 2018.

[111] James M. Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*, pages 189–326, 2004.

[112] Thomas S. Richardson and Jamie M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Preprint: http://www.csss.washington.edu/Papers/wp128.pdf, 2013.

[113] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[114] Philip E. Cheng. Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89(425):81–87, 1994.

[115] James M. Robins, Andrea Rotnitzky, and Lue P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994.

[116] Shona Fielding, Peter M. Fayers, Alison McDonald, Gladys McPherson, Marion K. Campbell, et al. Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. *Health and Quality of Life Outcomes*, 6(1):57, 2008.

[117] Louise Marston, James R. Carpenter, Kate R. Walters, Richard W. Morris, Irwin Nazareth, and Irene Petersen. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiology and Drug Safety*, 19(6):618–626, 2010.

[118] James M. Robins and Richard D. Gill. Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, 16(1):39–56, 1997.

[119] Stijn Vansteelandt, Andrea Rotnitzky, and James M. Robins. Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, 94(4):841–860, 2007.

[120] Giampiero Marra, Rosalba Radice, Till Bärnighausen, Simon N. Wood, and Mark E. McGovern. A simultaneous equation approach to estimating HIV prevalence with nonignorable missing responses. *Journal of the American Statistical Association*, 112(518):484–496, 2017.

[121] Eric J Tchetgen Tchetgen, Linbo Wang, and BaoLuo Sun. Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28(4):2069–2088, 2018.

[122] Yan Zhou, Roderick J. A. Little, and John D. Kalbfleisch. Block-conditional missing at random models for missing data. *Statistical Science*, 25(4):517–532, 2010.

[123] Ilya Shpitser. Consistent estimation of functions of data missing non-monotonically and not at random. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*. 2016.

[124] Mauricio Sadinle and Jerome P. Reiter. Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104(1):207–220, 2017.

[125] Rhian M. Daniel, Michael G. Kenward, Simon N. Cousens, and Bianca L. De Stavola. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256, 2012.

[126] Felix Thoemmes and Norman Rose. Selection of auxiliary variables in missing data problems: Not all auxiliary variables are created equal. Technical report, R-002, Cornell University, 2013.

[127] Fernando Martel García. Definition and diagnosis of problematic attrition in randomized controlled experiments. *Working paper. Available at SSRN 2302735*, 2013.

[128] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In *Proceedings of the 27th Conference on Advances in Neural Information Processing Systems*, pages 1277–1285. 2013.

[129] Karthika Mohan and Judea Pearl. Graphical models for recovering probabilistic and causal queries from missing data. In *Proceedings of the 28th Conference on Advances in Neural Information Processing Systems*, pages 1520–1528. 2014.

[130] Mojdeh Saadati and Jin Tian. Adjustment criteria for recovering causal effects from missing data. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2019.

[131] Rohit Bhattacharya, Razieh Nabi, Ilya Shpitser, and James Robins. Identification in missing data models represented by directed acyclic graphs. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2019.

[132] Ilya Shpitser, Karthika Mohan, and Judea Pearl. Missing data as a causal and probabilistic problem. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 802–811. AUAI Press, 2015.

[133] Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full law identification in graphical models of missing data: Completeness results. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[134] Hua Yun Chen. A semiparametric odds ratio model for measuring association. *Biometrics*, 63:413–421, 2007.

[135] Daniel Malinsky, Ilya Shpitser, and Eric J. Tchetgen Tchetgen. Semiparametric inference for non-monotone missing-not-at-random data: the no self-censoring model. *arXiv preprint arXiv:1909.01848*, 2019.

[136] Eric V. Strobl, Shyam Visweswaran, and Peter L. Spirtes. Fast causal inference with non-random missingness by test-wise deletion. *International Journal of Data Science and Analytics*, 6(1):47–62, 2018.

[137] Alex Gain and Ilya Shpitser. Structure learning under missing data. In *Proceedings of the 9th International Conference on Probabilistic Graphical Models*, pages 121–132, 2018.

[138] Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1762–1770, 2019.

[139] David M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

[140] Joseph D. Ramsey. Scaling up greedy causal search for continuous variables. *arXiv preprint arXiv:1507.07749*, 2015.

[141] Juan M. Ogarrio, Peter L. Spirtes, and Joseph D. Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of the 8th International Conference on Probabilistic Graphical Models*, pages 368–379, 2016.

[142] Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

[143] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[144] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483, 2018.

[145] Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. *arXiv preprint arXiv:2010.06978*, 2020.

[146] Razieh Nabi, Phyllis Kanki, and Ilya Shpitser. Estimation of personalized effects associated with causal pathways. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[147] Razieh Nabi, Joel Pfeiffer, Murat Ali Bayir, Denis Charles, and Emre Kıcıman. Causal inference in the presence of interference in sponsored search advertising. *arXiv preprint arXiv:2010.07458*, 2020.

[148] Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2019. NIH Public Access, 2019.

[149] Eli Sherman and Ilya Shpitser. Identification and estimation of causal effects from dependent data. In *Advances in Neural Information Processing Systems*, pages 9424–9435, 2018.

[150] Robin J. Evans and Thomas S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *The Annals of Statistics*, pages 1452–1482, 2014.

[151] Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

[152] Jaroslav Hájek. A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(4):323–330, 1970.

[153] Whitney K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990.

[154] David G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, 1997.

[155] Thomas S. Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles, 1990.

[156] Eric J. Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, 2012.

[157] Hua Yun Chen, Daniel E. Rader, and Mingyao Li. Likelihood inferences on semiparametric odds ratio model. *Journal of the American Statistical Association*, 110(511):1125–1135, 2015.

[158] Razieh Nabi and Xiaogang Su. coxphMIC: An R package for sparse estimation of cox proportional hazards models. *The R Journal*, 9:229–238, 2017.

[159] Razieh Nabi-Abdolyousefi. Conversion rate prediction in search engine marketing. Masters thesis, 2015.

[160] Razieh Nabi-Abdolyousefi and Afshin Banazadeh. 3D offline path planning for a surveillance aerial vehicle using b-splines. In *Proceedings of the 2013 International Conference on Advanced Mechatronic Systems*, pages 306–311. IEEE, 2013.

Razieh Nabi — Curriculum Vitae

rnabi@jhu.edu https://www.cs.jhu.edu/~rnabi

Research Interests

Causal Inference, Algorithmic Fairness, Missing Data, Semiparametric Statistics, Graphical Models, Machine Learning

Education

Johns Hopkins University, Baltimore, MD, USA — Sept. 2016 - Spring 2021
Ph.D. in Computer Science
Advisor: Ilya Shpitser
Thesis: *Causal Inference Methods for Bias Correction in Data Analyses*

Harvard University, Boston, MA, USA — Sept. 2019 - Oct. 2019
Visiting Scholar, Department of Epidemiology, School of Public Health
Host: James Robins

The University of Texas at El Paso, Texas, USA — Jan. 2015 - Aug. 2016
M.Sc. in Statistics
Advisor: Xiaogang Su
Thesis: *coxphMIC: R Package for Sparse Estimation of Cox Proportional Hazards Models*

Istanbul Sehir University, Istanbul, Turkey — Sept. 2013 - Jan. 2015
M.Sc. in Electronics and Computer Engineering
Advisor: Ahmet Bulut
Thesis: *Conversion Rate Prediction in Search Engine Marketing*

Sharif University of Technology, Tehran, Iran — Sept. 2007 - July 2012
B.Sc. in Aerospace Engineering
Advisor: Afshin Banazadeh
Thesis: *Trajectory Planning for Multiple Unmanned Aerial Vehicles in Urban Environment*

Research Experience

Johns Hopkins University — **Research Assistant**
Department of Computer Science, Baltimore, MD, USA — Sept. 2016 - Present

Microsoft Research — **Research Intern**
Information and Data Sciences Group, WA, USA — June 2020 - Aug. 2020

Center of Institutional Evaluation, Research and Planning — **Research Intern**
University of Texas at El Paso, TX, USA — June 2015 - Sept. 2015

Data Science Lab — **Research Assistant**
Istanbul Sehir University, Istanbul, Turkey — Feb. 2014 - Jan. 2015

Publications

Razieh Nabi*,[1] Rohit Bhattacharya*, and Ilya Shpitser, "Full Law Identification In Graphical Models Of Missing Data: Completeness Results," In *Proceedings of the Thirty Seventh International Conference on Machine Learning (ICML),* PMLR 119: 2352-2362, 2020.

Razieh Nabi*, Rohit Bhattacharya*, Ilya Shpitser, and James Robins, "Identification In Missing Data Models Represented By Directed Acyclic Graphs," In *Proceedings of the Thirty Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, AUAI Press, 2019.
Recipient of the **Tom Ten Have award** at Atlantic Causal Inference Conference.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser, "Learning Optimal Fair Policies." In *Proceedings of the Thirty Sixth International Conference on Machine Learning (ICML)*, PMLR 97: 4674-4682, 2019.

Razieh Nabi, Phyllis Kanki, and Ilya Shpitser, "Estimation of Personalized Effects Associated With Causal Pathways," In *Proceedings of the Thirty Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, AUAI Press, 2018.

Razieh Nabi and Ilya Shpitser, "Fair Inference on Outcomes," In *Proceedings of the Thirty Second Conference on Association for the Advancement of Artificial Intelligence (AAAI)*, AAAI Press, 2018.

Razieh Nabi and Xiaogang Su, "coxphMIC: An R Package for Sparse Estimation of Cox Proportional Hazards Models via Approximated Information Criteria," *The R Journal*, 9(1): 229 - 238, 2017.

Razieh Nabi and Afshin Banazadeh, "3D Offline Path Planning for Surveillance Aerial Vehicles using B-splines," *International Conference on Advanced Mechatronic Systems,* 2013.

Under Review:

Razieh Nabi*, Rohit Bhattacharya*, and Ilya Shpitser, "Semiparametric Inference For Causal Effects In Graphical Models With Hidden Variables," revised manuscript in review at *Journal of Machine Learning Research (JMLR)*, arXiv: 2003.12659, 2020.

Razieh Nabi, Todd McNutt, and Ilya Shpitser, "Semiparametric Causal Sufficient Dimension Reduction of High Dimensional Treatments," under review at *Journal of the Royal Statistical Society (JRSS): Series B,* arXiv: 1710.06727, 2020.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser, "Optimal Training of Fair Predictive Models," under review at *Journal of Knowledge and Information Systems (KAIS),* arXiv: 1910.04109, 2020.

Numair Sani, Jaron Lee, **Razieh Nabi**, and Ilya Shpitser, "A Semiparametric Approach to Interpretable Machine Learning," arXiv: 2006.04732, under revision, 2020.

[1]* Indicates equal contribution.

Razieh Nabi, Joel Pfeiffer, Murat Ali Bayir, Denis Charles, and Emre Kıcıman, "Causal Inference In The Presence Of Interference In Sponsored Search Advertising," *under review at NeurIPS workshop on Causal Discovery & Causality-Inspired ML,* arXiv: 2010.07458, 2020.

In Preparation:

Razieh Nabi, Eric Tchetgen Tchetgen, and Ilya Shpitser, "Semiparametric Estimation Theory for Causal Mediation Analysis in Longitudinal Settings with Multiple Mediators."

Razieh Nabi, Edward Kennedy, Ming-Yueh Huang, Matteo Bonvini, and Daniel Scharfstein, "Semiparametric Sensitivity Analysis: Unmeasured Confounding in Observational Studies."

Razieh Nabi, Rohit Bhattacharya, James Robins, Ilya Shpitser, "Graphical Methods for Identification in Missing Data Problems."

Rohit Bhattacharya, Daniel Malinsky, Jaron Lee, **Razieh Nabi**, and Ilya Shpitser, "Graphical Structure Learning from Data Missing-Not-At-Random."

Rohit Bhattacharya, Jaron Lee, **Razieh Nabi**, and Ilya Shpitser, "*Ananke-causal*: A Python Package for Causal Inference with Graphical Models."

Daniel Malinsky, **Razieh Nabi**, and Ilya Shpitser, "Algorithmic Fairness and Data-Driven Decisions: An Approach Based on Causal Constraints."

Razieh Nabi et al, "Racial Disparities in Cardiac Surgery Outcomes," Joint work with the Malone Center for Engineering in Healthcare and Cardiovascular Surgical Intensive Care Unit, Johns Hopkins University.

Professional Activities

Conference Reviewer

- Conference on Machine Learning for Healthcare, MLHC 2020
- Conference on Neural Information Processing Systems, NeurIPS 2020
- Conference on Uncertainty in Artificial Intelligence, UAI 2020
- International Conference on Machine Learning, ICML 2020
- NeurIPS Reproducibility Challenge, 2019
- Conference on Neural Information Processing Systems, NeurIPS 2019
- International Conference on Machine Learning, ICML 2019
- International Conference on Artificial Intelligence and Statistics, AISTATS 2019
- Conference on Neural Information Processing Systems, NeurIPS 2018

Journal Reviewer

- Journal of the American Statistical Association (JASA)

- Journal of Machine Learning Research (JMLR)
- Journal of Statistical Software (JSS)
- Jounral of Data Mining and Knowledge Discovery
- Journal of Experimental and Theoretical AI

Program Committee Member

- Workshop on Causal Discovery and Causality-Inspired Machine Learning, 2020
  Held as part of the conference on Neural Information Processing Systems (NeurIPS)
- Workshop on Algorithmic Fairness through the Lens of Causality & Interpretability, 2020
  Appointed as reviewer for submissions to both the *Papers track* and *Breakout sessions*
  Held as part of the conference on Neural Information Processing Systems (NeurIPS)
- Workshop on Consequential Decision Making in Dynamic Environments, 2020
  Held as part of the conference on Neural Information Processing Systems (NeurIPS)
- Workshop on Algorithmic Bias in Search and Recommendation, 2020
  Held as part of the European Conference on Information Retrieval (ECIR)
- Workshop on Knowledge Discovery in Healthcare Data, 2016
  Held as part of the International Joint Conference on Artificial Intelligence (IJCAI)

Teaching Experience

Instructor

- Fairness in Data Science: Criteria, Algorithms and Open Problems (upcoming) 2021
  A day-long course on developed methodologies for "fairness-aware" algorithms
  To be held in *Statistics in Epidemiology* session at Joint Statistical Meetings (JSM)
- Should Susan Smoke: An Introduction to Causal Inference Intersession 2020
  A month-long course on the statistical and philosophical foundations of causality
  The course was co-instructed and featured at Johns Hopkins Hub magazine
- Pre-College Math, University of Texas at El Paso Summer 2015
  A summer-long course on basic calculus concepts, algebra, trigonometry, and geometry

Head Course Assistant

- Machine Learning: Data to Models, Johns Hopkins University Spring 2019

Teaching Assistant

- Causal Inference, Johns Hopkins University Fall 2018
- Probability and Statistics, University of Texas at El Paso Spring 2016
- Elementary Statistical Methods, University of Texas at El Paso Fall 2015
- Calculus I/II, University of Texas at El Paso Spring 2015
- Physics I/II and Laboratory, Istanbul Sehir University Sept. 2013 - Jan. 2015

Invited Talks

American Causal Inference Conference, TX, USA *(upcoming)* May 2021[2]
Title: *Identification In Missing Data Models Represented By Directed Acyclic Graphs*
(presenting as part of the Thomas Ten Have award)

European Consortium for Informatics and Mathematics Dec 2020
ERCIM Working Group on Computational and Methodological Statistics (CMStatistics)
Title: *Semiparametric Inference For Causal Effects In Graphical Models With Hidden Variables*

Amazon Research, Tuebingen, Germany Dec 2020
Title: *Algorithmic Fairness via Causal Mediation Analysis*

Cornell University, NY, USA Oct 2020
Title: *Semiparametric Inference For Causal Effects In Graphical Models With Hidden Variables*

Microsoft Bing Ads and Microsoft Research, WA, USA Aug 2020
Title: *Causal Inference With Interference In Ad Placement*

Microsoft Research, AI and Society Seminar Series, WA, USA July 2020
Title: *Learning Optimal Fair Policies*

University College London, London, UK July 2020
Title: *Full Law Identification In Graphical Models Of Missing Data: Completeness Results*

Netflix Inc, CA, USA June 2020
Title: *Full Law Identification In Graphical Models Of Missing Data: Completeness Results*

University of Oxford and DeepMind, AI Safety Teams, UK June 2020
Title: *Learning Optimal Fair Policies*

Ecole Polytechnique, INRIA Saclay, and Google Brain, France May 2020
Title: *Full Law Identification In Graphical Models Of Missing Data: Completeness Results*

Bloomberg School of Public Health, Johns Hopkins University, MD, USA April 2020
Title: *Semiparametric Inference For Causal Effects In Graphical Models With Hidden Variables*

Harvard University, Kolokotrones Circle, MA, USA Oct 2019
Title: *Learning Optimal Fair Policies*

Reading Group at Institute for Quantitative Social Science, USA Oct 2019
Title: *Identification In Missing Data Models Represented By Directed Acyclic Graphs*

International Conference on Machine Learning, CA, USA June 2019
Title: *Learning Optimal Fair Policies*
(Plenary Talk)

[2]Plenary talks postponed from 2020 to 2021 due to COVID-19.

Caltech, Decisions, Games, and Logic Workshop, CA, USA June 2019
Title: *Learning Optimal Fair Policies*
(Plenary Talk)

Grad Council Student Seminar, Johns Hopkins University, MD, USA June 2019
Title: *Learning Optimal Fair Policies*

Guest lecturer at Causal Inference course, Johns Hopkins University Dec 2018
Title: *Fair Regressions and Fair Policies*

Association for the Advancement of Artificial Intelligence, LA, USA Jan 2018
Title: *Fair Inference On Outcomes*
(Plenary Talk)

Join Meeting of Statistical Genetics and Causal Inference groups, JHU Oct 2017
Title: presenting work on *Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome*

University of Washington, Aerospace Engineering Department, WA, USA Feb 2015
Title: *Dynamic Machine Learning and Big Data Analysis*

Poster Presentations

Workshop on Causal Discovery and Causality-Inspired ML, NeurIPS 2020

- Causal Inference in the Presence of Interference in Sponsored Search Advertising

Statistical and Applied Mathematical Sciences Institute, Duke University 2019

- Identification in Missing Data Models Represented by Directed Acyclic Graphs, and
- Estimation of Personalized Effects Associated with Causal Pathways

International Conference on Machine Learning, Long Beach, CA 2019

- Learning Optimal Fair Policies

Uncertainty in Artificial Intelligence, Causal Inference Workshop, Monterey, CA 2018

- Learning Optimal Fair Policies, and
- Semiparametric Causal Sufficient Dimension Reduction of High Dimensional Treatment

Atlantic Causal Inference Conference, Pittsburgh, PA 2018

- Estimation of Optimal Path-Specific Policies,
- Fair Inference on Outcomes, and
- Semiparametric Causal Sufficient Dimension Reduction of High Dimensional Treatment

Computing Research Association, Grad Cohort for Women, San Francisco, CA 2018

- Fair Inference on Outcomes

Neural Information Processing Systems, Causal Inference Workshop, CA 2017

- Fair Inference on Outcomes, and
- Semiparametric Causal Sufficient Dimension Reduction of High Dimensional Treatment

Computing Community Consortium Symposium, Washington DC 2017

- Fairness Through Causality

Honors and Awards

- Reviewer Award, Conference on Neural Information Processing Systems, NeurIPS 2020
  Awarded to top 10% of high-scoring reviewers at NeurIPS
- Grace Hopper Celebration (GHC) Student Scholarship, 2020
  Organized by AnitaB.org for celebration of women in computing
- Travel Award, Statistical and Applied Mathematical Sciences Institute (SAMSI), Dec. 2019
  Causal Inference Program Opening Workshop, Duke University
- Thomas R. Ten Have award, ACIC, May 2019
  Awarded for best poster at the Atlantic Causal Inference Conference, Montreal, Canada
- Summer Institute Scholarship, University of Washington, June 2018
  Program: Summer Institute In Statistics and Modeling in Infectious Diseases (SISMID)
- Travel Award, Grad Cohort for Women, April 2018
  Awarded by Computing Research Association
- Travel Award, Computing Community Consortium, Oct. 2017
- Distinguished Bachelor Dissertation Award, Tehran, Iran, 2012

Extracurricular

- Volunteer Translator, Coursera Global Translator Community
- Volunteer Data Scientist, Open Justice Baltimore
- Volunteer Mentor, STEM Achievement in Baltimore Elementary Schools
- Volunteer, Ronald McDonald House Charities and Generosity Global, Baltimore, MD
- Marathon Runner

Biographical Sketch

Razieh is a PhD candidate in the Department of Computer Science at Johns Hopkins University. She is joining Emory University as Rollins Assistant Professor in the Department of Biostatistics and Bioinformatics at Rollins School of Public Health starting July 2021. Her current research is situated at the intersection of machine learning and statistics, focusing on causal inference and its applications in healthcare and social justice. Her work spans problems in algorithmic fairness, missing data, personalized medicine, dependent data, mediation analysis, semiparametric inference, and causal graphical models. Prior to her PhD, Razieh received her M.Sc. in Statistics from the University of Texas at El Paso. Her thesis was titled "coxphMIC: R Package for Sparse Estimation of Cox Proportional Hazards Models" [158]. Prior to that, she earned her M.Sc. in Electronics and Computer Engineering from the Istanbul Sehir University, Istanbul, Turkey. Her thesis was titled "Conversion Rate Prediction in Search Engine Marketing" [159]. She earned her B.Sc. in Aerospace Engineering from Sharif University of Technology, Tehran, Iran. Her Bachelor's thesis, titled "Trajectory Planning for Multiple Unmanned Aerial Vehicles in Urban Environment," won the Distinguished Bachelor Dissertation Award by the Aerospace Research Institute [160].