Causal and Trustworthy Machine Learning: Methods and Applications

Limor Gultchin
Keble College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Michaelmas 2023

Acknowledgements

I would like to thank my co-advisors, Prof. Matt Kusner and Prof. Varun Kanade for their wise guidance and kindness through these past few years, as well as to Prof. Ricardo Silva who has become an adopting advisor to whom I owe much of my learning through this journey. I would also like to thank the Alan Turing Institute for the kind support of this work and my studies.

I have learned and benefited from all collaborators with whom I have had the honor of working: Dr. David Watson, Caroline Yuchen Zhu, Dr. Ankur Taly, Dr. Vincent Cohen-Addad, Dr. Sophie Giffard-Roisin, Dr. Frederik Mallmann-Trenn, Siyuan Guo, Dr. Alan Malek and Dr. Silvia Chiappa. I am also grateful for the intellectual exchanges facilitated by the UCL causal inference group and the many enriching discussions it fostered.

Earlier in my research path, I am grateful to have had the support and encouragement of mentors and advisors without whom I would not have made it thus far: Dr. Adam Kalai, Dr. Ofra Amir, Prof. Barbara Grosz, Prof. Stuart Shieber, and Dr. Scott Hale.

I could not have completed this work without the warm and dedicated support of family and friends, especially my parents, Josef and Etia, and my partner, Raphael Köster, who has put up with the sacrifice of many an evening and a weekend for the completion of this work.

Abstract

This work focuses on the intersection of machine learning and causal inference and the way in which the two fields can enhance each other by sharing ideas: utilizing machine learning techniques for the computation of causal quantities, the use of ideas from causal inference for invariant predictions under unseen treatment regimes, and the exploration of topics in trustworthy machine learning, including interpretability and fairness, with a causal lens. In each one of the presented works, we grappled with the strength of assumptions needed to utilize causal inference techniques and relax portions of them when possible.

In Chapter 1, we introduce the motivation behind the works and the challenges that sparked this plan of study. Chapter 2 provides a foundation on basic topics in causal machine learning and trustworthy machine learning. In Chapter 3, we introduce a causal effect estimation method under partial causal graph knowledge. In Chapter 4, we look at causal effect estimation in complex data settings, such as images, text, and gene expression networks, and propose an invariant estimation approach utilizing crude interventions. In Chapter 5, we provide a causal perspective on explainable machine learning, unifying existing works and providing a sound and complete algorithm involving the concepts of sufficiency and necessity. Finally, in Chapters 6 and 7, we introduce methods and investigations in fair machine learning.

Contents

1 | Introduction

Machine Learning (ML) has made some significant strides in recent decades – into the public eye, as well as into the mainstream of computational and scientific research at large. The most common techniques in the traditional ML toolbox are concerned with correlations within observed datasets, aiming to fit a successful predictor for an outcome in a certain dataset. This is crucially different from finding the true, real-world mechanism that maps inputs X to outputs Y; a successful predictor only needs to sufficiently correlate X and Y according to the training data view of this relationship. However, the true mapping is needed to answer many questions researchers are most curious about – how to make accurate personal medical recommendations, how to predict outcomes of policy interventions, or how to identify conditions that give rise to natural phenomena – extend beyond the realm of mere co-occurrences. They are concerned with causes and effects.

Furthermore, growing attention and scrutiny have been given to failure modes that may turn the promise of ML into peril – focusing on finding any successful predictor for a training dataset rather than finding the true underlying function comes at a cost: Unexpected things can happen during deployment. Much of the concerns about ML failure modes can be traced back to failures to identify the true underlying function. When applied to socially critical data, models may show disparate impact toward sensitive groups in a population due to anything from data collection practices to the algorithmic amplification of existing biases. When put in the hands of practitioners and researchers alike, deep models become hard to interpret: it is not immediately easy to explain why certain inputs lead to certain predictions or outputs.

1.1 The Importance of Cause and Effect

Modern machine learning models are experts in pattern recognition. Based on historical data or access to information to train on, they can help us achieve various prediction tasks: Does an image belong to a certain class in a supervised setting? What is the most likely next word given an input sentence in a self-supervised or generative setting? What are clusters of similar information that exist in the data in an unsupervised setting? Or what is an optimal policy to maximize an outcome in an online setting?

However, these impressive tasks, which ML models are getting continuously better at performing, are still not error-free, especially in safety-critical situations with rich contexts. There are various pitfalls.

Making predictions based on historical data may not enable us to distinguish types of correlations, making predictions brittle and challenging systems' robustness. We may be misled by confounders - also termed lurking common causes - and conclude that two separate events are related due to their co-occurrence, despite the lack of an intrinsic causal connection between them. If we let ourselves be misled in this way, we may conclude that ice cream consumption *causes* an increase in drowning accidents, when, in fact, summer is just a common cause leading to increases in both consumption and accident frequency in pools and on the beach. That would pose a real risk to a system's ability to generalize or make future predictions about unseen events that do not exactly match the training dataset. In other words, it can pose a major challenge to one of machine learning's most fundamental axioms, the i.i.d. assumption (see Section 2.1.2).

Such confusion between co-occurrence and cause-and-effect relationships is a major challenge in trying to use historical, observed data, to generate future predictions and courses of action. When predictions directly involve individuals' outcomes, causes and effects are ever more important; we would like to distinguish between causes that increase the likelihood of a certain outcome in a robust manner whenever they are present, from those that were associated with the outcome historically.

That is why causality research becomes relevant in the context of machine learning fairness. We would like to ask about causes and even counterfactual outcomes, and not simply follow historical trends and expectations. Having historically few female engineers should not penalize current-day female candidates as they look for their next challenge, but often does when companies attempted training models to identify promising candidates using their historical hiring data; Similarly, historical patterns in college admissions should not penalize modern-day candidates from marginalized groups as they apply to schools.

When humans attempt to explain processes – whether in a scientific context, in trying to establish the fairness of a certain outcome, in trying to debug a faulty system, in determining motive in a legal case or deciding on a course of action in a decision-making scenario – we naturally revert to causes and effects. We want to ask what would have happened had a certain cause not been in place: Would the observed outcome have changed? What are events whose presence increases the likelihood of a certain outcome? These are the types of investigations we naturally and inevitably

run in our heads and the kinds we would perform when trying to explain the behavior of machine learning systems we build and interact with. That is why causality is an important building block in the construction of ML explainability tools, as we will later explore.

Cause and effect are also important for data efficiency. Having even partial causal knowledge about a system we try to make predictions about can focus our gaze on relevant subsets of a dataset or action space and help us cast aside redundant or unnecessary portions thereof.

1.2 Machine Learning Failure Modes

We established the centrality of causality to certain machine learning goals above, but mainly as speculation or as an intellectual exercise. However, many of them are already featured in practical failure modes and challenges facing modern machine learning systems.

Oops, it didn't happen again. Or, robustness and invariant predictions. A common failure mode of mixing up correlations with causal structure in prediction models is often referred to as "spurious correlations" or "shortcut learning". It arises from the strict reliance ML has on correlations appearing in the training data. While the i.i.d. assumption (stating variables in modeling are identically and independently distributed) that is so fundamental to ML is kept, all should be well; however, that exposes systems to vulnerabilities whenever the assumption fails. This will inevitably happen when predictions are generated on data that differ from the training data, due to temporal, geographical, or selection differences. This challenge is well-known and documented and has been studied extensively in the literature in recent years (Geirhos et al., 2020; Peters et al., 2016a; Arjovsky et al., 2020; Makar et al., 2022; Veitch et al., 2021).

A prototypical example was introduced in 2018 by Beery et al. (2018) and consists of a computer vision task: classifying images of cows. As the authors note, cows' images often include green, pasture-like backgrounds, and not beach fronts, where the presence of cows is far less common. This can become a problem for a system trained on such common images but tasked with classifying a cow when it appears on the seaside, a new environment not well captured by the training data.

Other more pernicious examples may happen when systems involved in self-driving cars are trained on data produced in controlled environments but are required to make

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Figure 1.1: Poor generalization of visual recognition algorithms to new environments, as shown by Beery et al. (2018). The prediction outputs were produced by the authors, and show the top five labels and confidence in production produced by the algorithm on ClarifAI.com.

predictions in the real world, under varying weather conditions, in different countries, with different signage or driving cultures.

There are also common examples from medical domains, where correlations between medical conditions may be common to some countries and hospitals but not to others. Goldstein et al. explore one, pointing out that diabetes is associated with a high body mass index (BMI) in the United States, while in India and Taiwan, diabetes also frequently cooccurs with a low and average BMI (Tan et al., 2004). This represents a shift in distribution where a label in a classification task (diabetes) may have a shifting relationship with some nuisance variable (BMI) across environments. Constructing datasets and models without accounting for such shifts may lead models to exploit unstable relationships in the training data (if it comes from the US), and subsequently fail in predictions and mislead decision-makers in India and Taiwan.

What was may not always be. Or, machine learning fairness concerns. Such distribution shifts we explored above may become all the more alarming when decisions aided by ML models directly involve individuals in situations of social importance. With the advance of machine learning models into larger society, more and more examples of unfortunate failure modes of this kind began emerging. A rather short list of headlines reaching general news circulation between 2018 and 2021 includes the following:

1. Amazon trained a model that helped automate recruiting processes in the company and ultimately stopped its use because it was found to discriminate against female candidates (Dastin, 2018).

2. Many facial recognition systems were shown to misclassify Black faces at higher rates than faces from other ethnic groups (Buolamwini and Gebru, 2018; Simonite, 2019).

3. Various algorithms involved in feeding job ads to users online were shown to exclude women from their pool of relevant candidates for certain roles (Lambrecht and Tucker, 2019; Hao, 2021).

4. A system designed to judge a beauty contest picked predominantly white-skin winners, and no dark-skinned ones (Levin, 2016).

5. Uber drivers in India reported failures in the company's facial recognition software used to log into the app, costing them work (Bansal, 2022).

One particular study done by ProPublica that caught the ML academic community's attention came out in 2016, pointing to different error rates across white, black, and Hispanic defendants in the US who were assigned risk scores by the COMPAS model, aimed at predicting recidivism and aiding bail decisions in various US courts at the time (Angwin et al., 2016). We shall look at this study closer when discussing algorithmic fairness topics throughout this thesis, but it is a classic and still relevant reminder of the dangers of correlation-based systems trained on historical data used to aid future decisions with societal consequences. Thus, various researchers in the fields of causal inference and causal ML have, unsurprisingly, weighed in over the years on algorithmic fairness questions (Kusner et al., 2017a, 2019; Nabi and Shpitser, 2017, 2018; Kilbertus et al., 2017; Zhang and Bareinboim, 2018; Wu et al., 2019b; Plecko and Bareinboim, 2022; Creager et al., 2020; Makhlouf et al., 2020), and we shall do so in this work as well.

Tell me why. Or, machine learning explainability. When failure modes such as the above happen, the most basic human instinct is to try and explain such failure modes by asking – "why did the model provide a wrong prediction?". How to answer this 'Why' question has become the topic of much study and deliberation within the ML community in recent years due to the increasing complexity of properly formulating it and answering it with systems and training datasets that are becoming ever bigger and more complex.

Modern deep learning models include billions of parameters and their training data may encompass most of the content on the internet (Ventures, 2022; Simon, 2021). This is why the rise of deep learning (DL) has also been coupled with a worry about their black-box nature – the seemingly simple task of identifying bugs within them, describing failure modes, and providing explanations to users about their outputs has become rather challenging.

The answers to 'why', 'what if' or 'what would have happened' questions, which are predominantly how humans try to explain phenomena of interest – are all causal questions that can be given a mathematical formulation on the way to their quantification via the framework of causal inference. In this work, we shall also explore how causal framing can help answer such questions about the outputs of ML systems.

Cut to the chase. Or, machine learning data efficiency. Another crucial pain point for modern ML is its reliance on vast amounts of data and compute to achieve state-of-the-art results.

This is yet another area in which causal knowledge and the framework of causal inference can help machine learning models achieve even more impressive results while lightening the data and compute burden. It has been shown, for example, that in the context of online ML systems, when a model tries to find the optimal course of action or policy, to achieve a certain goal, causal knowledge can be exploited to avoid the exploration of actions that are redundant for the presented goal (Lee and Bareinboim, 2018a, 2020, 2019a; Zhang and Bareinboim, 2022).

Another way causal inference can improve data efficiency is through the use of causal feature selection. Rather than selecting features based solely on their correlation with the outcome variable, causal feature selection takes into account the causal relationships between variables to identify which features are most likely to have a causal effect on the outcome variable. By focusing on these causal features, machine learning models can be trained with less data and achieve good performance, which is more likely to generalize.

1.3 Machine Learning for Causality, Causality for Machine Learning

Keeping in mind the failure modes and challenges ML systems are facing, and also the great promise they hold for traditional causal inference investigations, this work joins in and contributes to the development of a research agenda that brings together

two communities – that of ML/AI, and that of classic causal inference in statistics, econometrics, public health, and other applied science fields. Both have much to contribute to the other: causal inference can help the development of more robust, fair, explainable, and data-efficient machine learning systems; while machine learning algorithms can be used to identify and estimate causal effects from high-dimensional and complex observational data, which is otherwise challenging with traditional statistical methods.

1.4 Contributions

In the following, we will explore topics, problems, and proposed solutions that lie between these two pillars. We will explore ML methods for the computations of causal effects, see how one can make use of causal knowledge for the discovery of pragmatic mediators that aid in the estimation of effects of unseen interventions while furthering interpretability and will explore ways to advance trustworthy ML via exploring causal and non-causal approaches to problems in fairness and explainable AI (XAI).

We will begin with the estimation of causal effects, where a single intervention may lead to an outcome via partial discovery in two different settings. My first project focused on estimating the Average Treatment Effect (ATE) of a binary intervention under highly limited knowledge of the data generation process, represented as a Directed Acyclic Graph (DAG). The work consisted of defining and optimizing a differentiable objective function to identify a valid adjustment set that would satisfy the graphical *backdoor criterion*, the conditioning on which would make the causal effect we are after identifiable from observational data. This work is explored in Chapter 3.

Next, we looked at the effect of "crude" interventions on complex, high-dimensional objects, which in turn led to a scalar outcome of interest. The motivation was the study of interventions on text, images, or networks, which usually receive little attention in the study of causality. Their limited presence in literature could be partially due to the difficulty in defining interventions over such "complex objects". While it is relatively easy to vary low dimensional covariates, such as "medicine intake" or "blood pressure level", and assign them (stochastically or deterministically) exact values, it is harder to conceptualize interventions on representations of text or images. Thus, we propose estimating the effects of crude interventions on such data modalities via a two-step procedure that uncovers mediation mechanisms from the interventions,

through the complex objects, and into the outcome of interest. This work is explored in Chapter 4.

In the realm of trustworthy ML, we looked at topics in XAI and fairness. Looking at the interpretability of models, we worked on a unifying framework that connected the somewhat spread out landscape of local explanations literature via causal concepts, which further the understanding of how such methods would differ or agree in their results. We showed that methods that belong to the main three approaches to model interpretability – feature attributions (e.g., Shapley values (Lundberg and Lee, 2017)), rule lists (e.g., Anchors (Ribeiro et al., 2018a)) and Counterfactuals (e.g., counterfactual explanations and algorithmic recourse (Wachter et al., 2018; Karimi et al., 2020b)) – can all be expressed in terms of the causal concepts of the probability of sufficiency and necessity of interventions, and show how those can be expressed and made sense of at different levels of knowledge of the underlying causal structure of the dataset. Finally, we proposed a sound and complete algorithm for the enumeration of explanatory factors. This work is presented in Chapter 5.

Within the algorithmic fairness literature, as already discussed briefly, a great contributor to the renewed interest in algorithmic fairness since 2016 was an investigative data piece by ProPublica, which looked at the fairness of COMPAS, a predictive recidivism system used in courts in certain counties in the US (Angwin et al., 2016). In response to Angwin et al.'s finding that showed the system's error rates were largely unequal across groups – showing twice as large false positive rate (FPR) error for black defendants compared to white defendants, and the reverse for false negative rate (FNR) – a famous impossibility result was proposed by multiple authors in the academic community. The impossibility result showed that in settings where labels are not spread equally across groups, exact equal error rates and exact calibration (related to the notions of predictive positive value (PPV), false discovery rate (FDR), etc., which alongside error rates complete the classification confusion matrix), cannot be simultaneously achieved unless we already have a perfect classifier (Chouldechova, 2017; Kleinberg et al., 2017). We took another look at the original result, offered a contra-positive view and a refinement of the connection to accuracy, and used it to propose looking for a balance between error rate and calibration-style violations of fairness, rather than abandoning one or the other as previously proposed in the literature. We propose to achieve such a better balance with a minimization of the upper bound of all violations reflected in a confusion matrix. This work is introduced in Chapter 6.

Finally, we will expand our view on fairness to policy optimization settings, beyond supervised prediction systems. There, we studied what can be done with a causal perspective in mind, but beyond counterfactual fairness (Kusner et al., 2017a) and its path-specific variants (e.g., Chiappa 2019). Most causal approaches to fairness focused on the manipulation of the sensitive attribute, and equating counterfactual quantities in response across groups. However, such an objective often relies on full knowledge of the causal graph, and on complex and restrictive counterfactual computations. Instead, we propose a framework that optimizes a policy for maximum expected utility overall, while mitigating disparity by focusing on what we can control – the policy itself. We propose a framework we name pragmatic fairness, consisting of two approaches to disparity control. One aims to not introduce new differences in effects across groups by the introduction of the new policy (compared to a baseline); the other aims to reduce existing disparities by minimizing the interaction of chosen actions and the sensitive group membership, making sure groups are treated similarly to the extent we can control with the policy. This work is presented in Chapter 7.

2 | Background

This chapter is designed to provide background on the problems introduced in the rest of the work. We will survey relevant concepts and basic literature for each of the directions, and explain how it informs and ties into the work presented in the following chapters.

2.1 Basic Toolkit

2.1.1 Independence: Marginal and Conditional

Statistical independence is an important concept in probability theory and statistics, which refers to the lack of a relationship between two or more random variables. Marginal independence and conditional independence are two common types of statistical independence, which will become building blocks for much of what follows. Marginal independence refers to the lack of a relationship between two random variables, without considering the effect of any other variables. Mathematically, two random variables X and Y are marginally independent ($X \perp\!\!\!\perp Y$) if their joint probability distribution can be expressed as the product of their marginal probability distributions, i.e.,

$$P(X, Y) = P(X)P(Y). \tag{2.1}$$

On the other hand, conditional independence refers to the lack of a relationship between two random variables, given the value of one or more other variables. Mathematically, two random variables X and Y are conditionally independent given a variable Z ($X \perp\!\!\!\perp Y \mid Z$) if their conditional probability distribution satisfies the following equation:

$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z). \tag{2.2}$$

Conditional independence is a powerful concept in probability theory and statistics and is often used to simplify complex models and computations (Koller and Friedman, 2009, p.5).

Since the concepts of marginal and conditional independence are so fundamental to our discussion of relations between variables – those that co-occur and those that do not vary together – they are key to our discussion of causation. Our aim is often to understand what would happen to certain variables under variations in others, i.e., what would happen in a system of variables under response to interventions changing the values of some subset of them. Independencies and dependencies between variables therefore become a necessary ground for this discussion, in both passively observed systems and those under active interventions.

The fundamental assumptions of the two most popular causal inference frameworks, potential outcomes (Rubin, 2005) and structural causal models (SCMs) (Pearl, 2009b; Peters et al., 2017), are grounded in statistical independence. We will survey these further in Section 2.2.1.

2.1.2 The i.i.d. Assumption in Machine Learning and Causal Inference

The i.i.d. (independent and identically distributed) assumption is a fundamental concept in machine learning, and it is critical for ensuring the validity and generalizability of machine learning models. In the context of supervised learning, the i.i.d. assumption assumes that the training and test data are independently drawn from the same underlying probability distribution and that the samples in the dataset, $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, are all identically distributed, i.e., $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \sim P_X$. This assumption enables the use of statistical tools and techniques, such as maximum likelihood estimation and hypothesis testing, which are based on the assumption of independent and identically distributed data.

The i.i.d. assumption is also closely related to causal inference, which is concerned with identifying the causal effects of interventions in complex systems. In particular, the i.i.d. assumption is often violated in real-world problems. Data may be dependent and non-identically distributed due to distribution shifts across geography or time, sampling practices, or the presence of confounding variables and selection bias. These issues can be tackled using causal inference methods that explicitly account for the causal relationships between variables and the underlying structure of the data-generating process.

Crucially, the i.i.d. assumption is also often violated in common machine learning use cases, such as prediction tasks where inference is done under distribution shifts, due to time lags, changes in location or geography, changing cultural norms, or many

other factors. We have begun exploring the implications of such shifts and the failure modes that could arise in Section 1.2.

In recent years, such examples have sparked a growing interest in developing machine learning methods that are explicitly designed to handle causal inference problems (Kaddour et al., 2022). These methods, which are part of the "causal inference for ML" category, aim to incorporate causal assumptions into the modeling process. By leveraging insights and tools from causal inference, these methods can help address some of the key challenges in machine learning, such as confounding bias, selection bias, and data heterogeneity, and enable more accurate and reliable predictions and decision-making across time, geography, and data collection practices.

2.2 Causal Inference for Machine Learning

2.2.1 Different Approaches to Causality

Potential Outcomes. The potential outcomes framework (Rubin, 2005), also known as the Neyman-Rubin model or the counterfactual framework, is a statistical approach to causal inference. It has become a standard method in many fields, including medicine, social sciences, economics, or any other setting in which we are interested in measuring the effect of an intervention, program, or policy. A classic example is assessing the efficacy of a new drug: before releasing a new medical treatment to the market, it is necessary to determine its effectiveness compared to a placebo or existing treatment. The gold standard setting for estimating such an effect is experimental: perform a Randomized Control Trial (RCT). However, RCTs are costly, lengthy and not always feasible.

The framework of potential outcomes is one way, tightly linked with an RCT formulation, that may fill the gap when an RCT is infeasible (e.g., under a pandemic caused by a little known virus), or when we hope to make trials more focused and efficient. It is based on the idea that each unit in a population (e.g., patient) has a potential outcome under each possible treatment condition (e.g., drug taking/placebo/current treatment). The most common target of estimation is then the causal effect of a treatment (e.g., the efficacy of the new drug): the difference between the potential outcomes (or a summary thereof) for the treated and control units (the resulting medical condition of the patients who took the new drug vs. those who did not).

Formally, let Y_1 and Y_0 be the potential outcomes under treatment and control, respectively. The causal effect of the treatment, denoted by the Average Treatment

Effect (ATE), is defined as:

$$ATE = \mathbb{E}[Y_1 - Y_0]. \tag{2.3}$$

We will introduce ATE once again later, under an alternative notation and framework. The key takeaway at this stage is that the potential outcomes notation is fundamental to the potential outcomes framework, as well as the consideration of each outcome-treatment combination as its own random variable.

Since only one of the potential outcomes can be observed for each unit (the one that actually took place when the measurement was taken), the ATE cannot be estimated directly by naively comparing means. Instead, various methods have been developed to estimate the ATE from observed data, such as propensity score matching, instrumental variables, and regression adjustment (Rosenbaum and Rubin, 1983; Hernán and Robins, 2019).

The potential outcomes framework relies on a set of assumptions, known as the "no unmeasured confounding" or "ignorability" assumptions, to ensure that the estimated ATE is unbiased. These assumptions require that the treatment assignment itself is independent of the potential outcomes, given a set of observed covariates. In other words, the treatment assignment must be "as good as random" within each level of the covariates, akin to a coin flip – just like in an RCT design.

The potential outcomes framework has been applied to a wide range of research questions, such as the effects of medications, educational interventions, and social policies. While the framework has limitations and assumptions that must be carefully considered in each application, it has become a valuable tool for researchers seeking to make causal inferences from observational data.

Structural Causal Models (SCMs). The structural causal models (SCMs) framework is based on the idea that causal relationships can be represented as a set of directed edges between variables, where each variable is either a cause, an effect, or both of the others.

Formally, let $\mathcal{G} = (V, E)$ be a directed causal graph (DAG), where V is a set of variables and E is a set of directed edges. The graph represents the causal relationships between the variables, where an arrow from variable X to variable Y denotes a direct causal effect of X on Y.

Framework Differences. The potential outcomes framework and SCMs are two approaches to causal inference that share some similarities but also have some key differences. The potential outcomes framework is a statistical approach to causal

inference that focuses on comparing the outcomes that would have been observed under different treatment conditions. The starting point of the framework is that each unit in a population has a potential outcome under each possible treatment condition and that only one of these potential outcomes can be observed for each unit. The causal effect of a treatment is then defined as the difference between the potential outcomes for the treated and control units.

SCMs, in turn, are a graphical approach to causal inference that focuses on identifying the causal relationships among variables in a system. This approach uses directed acyclic graphs (DAGs) to represent the causal relationships among variables and employs formal rules to determine which variables can be considered causes or effects in a given system.

Despite these differences, both the potential outcomes framework and SCMs often aim to estimate causal effects from observational data, and both rely on assumptions to make causal inferences; to do so, both approaches aim to account for sources of confounding and other forms of bias that might affect the estimation of causal effects.

Thus, the potential outcomes framework and SCMs can be seen as complementary approaches to causal inference that can be used to address different research questions and use cases. The potential outcomes framework is closer in terminology to randomized controlled trials and experimental design, while SCMs can help analyze more complex systems with an increasing number of variables and multiple types of relationships.

In the following, we will mainly focus on SCMs as the framework that the presented works will be situated in. However, it is worth noting SCMs[1] come with a relatively strong set of cross-world independent assumptions which are not strictly necessary for the works presented in the following chapters. The same methods and results can be achieved under the single-world intervention graphs (SWIGs) framework, which relies on weaker assumptions (Robins and Richardson, 2010; Richardson and Robins, 2013b,a).

2.2.2 Structural Causal Models

Structural causal models (SCMs) are a powerful framework for representing and reasoning about causal relationships between variables in a system. SCMs provide a way to specify the underlying mechanisms that generate the observed data, and thus give us the "full story" about how it came to be, and how variables take their value in

[1]SCMs can also be described as non-parametric SEMs with independent errors, NPSEM-IEs.

relation to other ones in the dataset. This, of course, can become quite handy when trying to make causal and counterfactual calculations.

Formally, an SCM is defined as a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{F})$, where $\mathcal{U}$ is a set of exogenous random variables (i.e., variables whose values are solely determined by factors *outside* of the model), $\mathcal{V}$ is a set of endogenous random variables (whose values are determined by factors *within* the model), and $\mathcal{F}$ is a set of structural equations f that define how any $V_i \in \mathcal{V}$ depends on the exogenous variable(s) $U_i \in \mathcal{U}$ and possibly other endogenous variable(s), $V_j \in \mathcal{V} \setminus V_i$. The structural equations take the form:

$$v_i = f_i(pa_i, u_i), \tag{2.4}$$

where v_i is the value of the ith endogenous variable V_i, u_i is the value of the ith exogenous variable U_i and f_i is a deterministic function that maps the values of its parent variables pa_i (i.e., variables pa_i cause V_i) and U_i to the value of V_i.

Among their other uses, SCMs can allow us to make counterfactual predictions, i.e., predictions about what would have happened if we intervened on one or more variables in the system, potentially in contrast to events that actually took place during data collection. Note, however, that some of these quantities can be computed without full knowledge of an SCM and the graphical model $\mathcal{G}$ associated with it, as we will explore later in this work.

Given an SCM, we can define an intervention on a variable X as setting its value to a specific value x, denoted by $do(X = x)$. The resulting system is called the intervened SCM, denoted by $(\mathcal{U}, \mathcal{V}, \mathcal{F}[X = x])$. By comparing the distribution of the intervened SCM to the original SCM, we can make causal and potentially counterfactual predictions about the effect of the intervention on other variables in the system, in relation to observed events we can condition on at the same time.

SCMs, alongside their associated DAGs, can facilitate various tasks in causal inference, as we're about to explore next. One important property we gain by considering graphical models - beyond their intuitive nature and how they enable the development of graphical criteria to achieve tasks - is that we get to inherit their rule of product decomposition. We can specify a joint distribution over a system of variables in terms of a product of conditional factors involving each node and its parents.

Rule of Product Decomposition. For any DAG, the joint distribution of its variables factorizes as

$$P(X_1, X_2, ..., X_n) = \prod_i P(X_i \mid pa(X_i)), \tag{2.5}$$

where $pa(X_i)$ are the parents of each variable X_i in the model. This fact already highlights the equivalence between a graph structure and the conditional independence relations between its nodes/variables. We will continue exploring this connection in the next section.

2.2.3 d-separation and Formalization of Dependencies

In Section 2.1.1, we began to explore the debt causal inference owes to the fundamentals of marginal and conditional in/dependencies. Now that we have introduced the concept of a causal graph, we can make use of it to formalize and systematize the study of relations between variables.

Mathematically, we can use the concept of d-separation to determine whether two nodes are marginally or conditionally independent in a causal graph. To do so, we need to introduce the notion of a blocked path, where a *path* is a sequence of nodes and edges connecting two variables (or nodes in the graph) with each other, regardless of orientation. This notion will, in turn, depend on the three types of relationships possible between any set of three variables in a DAG: collision, confounding, and mediation. The following definitions will also require us to define what are *ancestor* and *descendant* nodes in a causal graph. The first node in a directed path is the ancestor of all following nodes, while every node on the path is a descendant of the first node.

d-separation, Intuition. A *path* between two nodes X and Y can be *blocked* in the following ways:

- *without conditioning* by a *collider* Z (a node Z with two incoming edges from neighboring nodes, i.e., $X \rightarrow Z \leftarrow Y$), or
- *by conditioning on a set of nodes* which contains one of the following Z patterns:
  - a *confounder* (a common cause, or a node with outgoing arrows into two other nodes $X \leftarrow Z \rightarrow Y$), or
  - a *mediator*, lying on a directed path from X to Y ($X \rightarrow Z \rightarrow Y$).

A set of nodes Z d-separates two nodes X and Y if every path between X and Y is blocked by Z. If X and Y are d-separated by Z, then they are conditionally independent given Z. Similarly, if X and Y are d-separated by the empty set, then they are marginally independent.

d-separation, Formal Definition. (Pearl et al., 2016, p. 46-7) d-separation can be defined as follows. Let $\mathcal{G} = (V, E)$ be a directed acyclic graph (DAG), and let X, Y, and $Z \subseteq V$. A path p from X to Y is said to be blocked by Z if and only if:

- p contains a chain of nodes, i.e. a mediation path $X \rightarrow C \rightarrow Y$ or a fork, i.e., confounding $X \leftarrow C \rightarrow Y$ such that the middle node $C \in Z$ (i.e., C is conditioned on), or
- p contains a collider $X \rightarrow C \leftarrow Y$ such that the collision node $C \notin Z$, and no descendant of $C \in Z$.

If Z blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on Z, and thus are independent conditional on Z, i.e. $X \perp\!\!\!\perp Y \mid Z$.

By using causal graphs and d-separation, we can thus determine whether sets of variables are marginally or conditionally independent, and use this information to infer causal relationships.

The next section will discuss some of the prototypical estimands we are after in causal inference – causal effects. Identifying those from observational data requires an identification strategy that will correctly account for challenges such as confounding bias. In certain settings, this can be achieved by identifying the correct covariates to involve in a given estimation, while leaving others out of it. In such cases, d-separation can be viewed as a unifying principle guiding the selection of suitable sets of covariates for a given estimation task.

2.2.4 Causal Effects: ATE, CATE, ITE

Causal effect estimation, under various assumptions, is one of the main areas of study in the causal inference literature. Many problems in causal inference involve questions of the "what if" form, e.g., "What if patient z took medicine x? What value y would they attain?". Traditional statistics considers observational probability distributions to answer such questions, e.g. $P(Y = y \mid X = x, Z = z)$. However, such quantities involve correlative information. They tell us something about the co-occurrence of such events in a certain sample we study. In contrast, causal inference aims to identify *causal* quantities from observational data – to answer some future or hypothetical question of what values would random variables attain (or would have attained) in response to an intervention, i.e. assigning or controlling the values of some variables in a model, and estimating some response. Mathematically, and using Pearl's do-notation,

we can represent such quantities via the do-operator, $P(Y = y \mid do(X = x), Z = z)$ (Pearl et al., 2016, pp. 55-56).

Different summaries of $P(Y = y \mid do(X = x), Z = z)$ might be considered when estimating causal effects, and we introduce the canonical ones below.

Formally, the causal effect of a treatment X on an outcome Y is defined as the difference in the expected outcome between the treatment group and the control group. This can be represented mathematically as:

$$\tau = \mathbb{E}[Y \mid do(X = 1)] - \mathbb{E}[Y \mid do(X = 0)], \tag{2.6}$$

where Y is the outcome variable, X is the treatment variable, and $do(X)$ denotes the intervention that sets X to a particular value, e.g. 1 (treatment) or 0 (control). It is also possible to consider other values to assign for the treatment variable, as the use case calls for. The causal effect, sometimes denoted by τ, represents the average difference (across the population) in the outcome between the treatment group and the control group, after adjusting for confounding variables.

We therefore call the basic quantity above the **Average Treatment Effect (ATE)**, which measures the average effect of a treatment on the population as a whole. It is defined as the difference in the expected outcome between the treatment group and the control group, averaged over the entire population.

The ATE is a popular target in Randomized Control Trials (RCTs). But these are often expensive, difficult to execute, morally problematic, or simply infeasible. Instead, a common goal in causal inference is to try and recover this quantity from observational information and some knowledge of the underlying data generation process. The main difficulty in achieving this task is overcoming bias introduced by confounders (conditioning on Z in a pattern such as $X \leftarrow Z \rightarrow Y$), while not d-separating X from Y by closing informative mediation paths or opening already separated paths via conditioning on colliders. While a naive approach might be to simply control for all covariates in the dataset which are not X and Y, it may also be doomed to fail, as we might unnecessarily block mediation paths (if Z as a facilitating mechanism of the causal effect, i.e. $X \rightarrow Z \rightarrow Y$) or introduce bias by conditioning on colliders (similar to Berkson's paradox, $X \rightarrow Z \leftarrow Y$).

Certain tasks require more specialized estimands than simply averaging across a population. The **Conditional Average Treatment Effect (CATE)** measures the effect of a treatment on a specific subgroup of the population, defined by a set of conditioning variables. The CATE is the difference in the expected outcome between

the treatment group and the control group, averaged over the subgroup. Formally, the CATE is given by:

$$CATE(z) = \mathbb{E}[Y \mid do(X = 1), Z = z] - \mathbb{E}[Y \mid do(X = 0), Z = z], \tag{2.7}$$

where Z is a set of conditioning variables.

Even more specifically, the **Individual Treatment Effect (ITE)** measures the effect of a treatment on an individual unit, such as a patient or a customer. It is defined as the difference in the outcome that an individual would have under the treatment and control conditions. Formally, the ITE is given by:

$$ITE_i = Y^i_{X=1} - Y^i_{X=0}, \tag{2.8}$$

where $Y^i_{X=1}$ and $Y^i_{X=0}$ are the outcomes of individual i under the treatment and control conditions, respectively.

2.2.5 The do-calculus

Using *d-separation*, a general machinery for the identification of causal effects, where possible, is the *do-calculus* (Pearl, 2009c, sec. 3.4). The do-calculus is a set of rules developed for reasoning about causal relationships between variables, and for deriving causal effects from observational and interventional data. In other words, its goal is to translate causal estimands where possible to mere statistical quantities by removing do-operators from expressions. It uses DAGs and graphical relationships within them to systematically perform such manipulations – the exchange of observational data for actions, or the removal of certain actions or variables altogether.

The three rules of do-calculus are:

1. **Insertion/deletion of observations:**

$$P(Y|do(X), Z, W) = P(Y|do(X), W) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{\mathcal{G}_{\overline{X}}} \tag{2.9}$$

   Here, the independence statement must hold in graph $\mathcal{G}_{\overline{X}}$, which is derived from the original causal graph $\mathcal{G}$ by removing all incoming arrows into X.

2. **Action/observation exchange:** We can amend this rule to have two interventions on the left-hand side and one on the right-hand side. This could be stated as follows:

$$P(Y|do(X), do(Z), W) = P(Y|do(X), Z, W) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{\mathcal{G}_{\overline{X},\underline{Z}}} \quad (2.10)$$

Here, the independence statement must hold in graph $\mathcal{G}_{\overline{X},\underline{Z}}$, which is obtained from the original causal graph $\mathcal{G}$ by removing all arrows incoming into X and outgoing from Z.

3. **Insertion/deletion of actions:** Finally, this rule can be stated as follows:

$$P(Y|do(X), do(Z), W) = P(Y|do(X), W) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{\mathcal{G}_{\overline{X},\overline{Z^*}}} \quad (2.11)$$

This time, the independence statement must hold in graph $\mathcal{G}_{\overline{X},\overline{Z^*}}$, which is derived from the original causal graph $\mathcal{G}$ by removing all arrows into X, and then all arrows pointing into Z^* (Z^* is the subset of nodes in Z, which are not ancestors of any variable in $\mathcal{G}_{\overline{X}}$).

The do-calculus was shown to be complete (Shpitser and Pearl, 2006a; Huang and Valtorta, 2006; Shpitser and Pearl, 2006b), in the sense that if a sequential application of the rules of do-calculus cannot identify a causal effect, no other method can provide non-parametric identification for the same set and strength of assumptions.

Finding the sequence of rule applications does not immediately follow from the do-calculus, but over time the ID algorithm was developed, which is capable of finding these steps and exiting with failure whenever the answer does not exist (non-identifiability) (Tian and Pearl, 2002; Shpitser and Pearl, 2006a).

2.2.6 Canonical Identification Strategies

Special cases of the do-calculus for specific graphical settings exist, including the *backdoor* and *frontdoor* criteria (Pearl et al., 2016, p. 61-9). One – the backdoor criterion – in fact preceded the do-calculus (Pearl, 1993, 1995) and is closely related to conditional unconfoundedness from the potential outcomes framework. The other – the frontdoor criterion – was one of the first outcomes of the do-calculus.

Chapter 3 offers a differentiable optimization procedure that can lead to the discovery of a valid adjustment set as defined by the *backdoor criterion* if one exists. The backdoor criterion is a graphical criterion, which states,

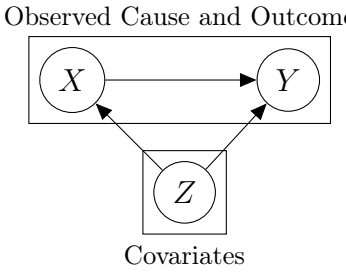


Figure 2.1: A DAG with circles representing variables and arrows representing causal relationships between them. The rectangles represent the observed variables and the covariates according to the Backdoor Criterion.

The Backdoor Criterion. Given an ordered pair of variables (X, Y) in a directed acyclic graph $\mathcal{G}$, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.

Next, given a valid adjustment Z defined as above, we could apply the adjustment formula to recover the causal effect of X on Y[2]

$$P(Y = y \mid do(X = x)) = \sum_{\mathbf{z}} P(Y = y \mid X = x, Z = z)P(Z = z). \tag{2.12}$$

The backdoor criterion can help us achieve our goals in various settings, but it relies on a graphical representation, in the form of a DAG, of pre-existing assumptions about the data generation process. In the lack of such clear understanding – which endows the available dataset with its causal interpretation by formalizing an interventional probability distribution – the backdoor criterion cannot be directly applied.

The Frontdoor Criterion. In the criterion we just introduced, all variables were observed. The front-door criterion is a method for estimating the causal effect of a treatment variable on an outcome variable, when there is an unobserved confounding variable that affects both the treatment and the outcome variables. The front-door criterion provides a way to bypass the confounding variable, by identifying a set of variables that satisfy certain conditions.

The front-door criterion enables that (as long as there is a suitable mediator M) by decomposing the effect identification into two steps. For M to be suitable, the following conditions need to be satisfied.

[2]Notice we define the backdoor criterion and the proceeding frontdoor criterion via categorical variables, which involves a marginalization via summation. For continuous variables, we simply replace summation with integration.

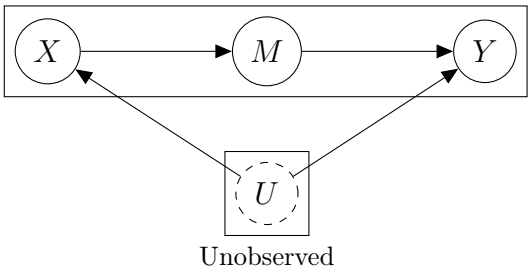


Figure 2.2: A graph with circles representing variables and arrows representing causal relationships between them. The rectangles represent the observed (including the mediator) and unobserved confounder variables characteristic of the Frontdoor Criterion.

1. M has to intercept all directed paths from X to Y. In other words, all directed paths from X to Y need to include/pass through M.

2. There could be no unblocked backdoor paths from X to M.

3. All backdoor paths from M to Y need to be blocked by X.

If we have such a suitable M that satisfies the criteria above we can identify the causal effect of X on Y in two steps: first via getting the effect of X on M, and then at the effect of M on Y, while controlling for X. This can be expressed mathematically as:

1. By condition 2,

$$P(M = m \mid do(X = x)) = P(M = m \mid X = x) \tag{2.13}$$

2. By condition 3,

$$P(Y = y \mid do(M = m)) = \sum_{x} P(Y = y \mid M = m, X = x)P(X = x) \tag{2.14}$$

Keeping in mind condition 1, we can now string it all together:

$$\begin{aligned} P(Y = y \mid do(X = x)) = \\ \sum_{m} P(Y = y \mid do(M = m))P(M = m \mid do(X = x)) = \\ \sum_{m} \sum_{x'} P(Y = y \mid M = m, X = x')P(X = x')P(M = m \mid X = x). \end{aligned} \tag{2.15}$$

We can finally get the estimated causal effect of X on Y from $P(Y = y \mid do(X = x))$, or whichever summary of it we're after, such as the ATE.

2.2.7 Pearl's Causal Hierarchy

We have considered the estimands and identification strategies enabled by the SCM causal framework. We can summarize the types of queries we now have tools to answer with a hierarchy, going from ordinary statistical, observational estimands all the way to counterfactuals, strengthening the required assumptions at every step.

1. **Association (Observational) Level:** The lowest level of the hierarchy represents questions about associations in observed data, the kind of questions answered by traditional statistics and machine learning models. At this level, we can make statements like "Y is correlated with X". For example, in a given dataset, we might investigate whether there's an association between smoking (X) and lung cancer (Y).

   Example estimand:

   $$\mathbb{E}[Y \mid X = x] \tag{2.16}$$

2. **Intervention (Causal) Level:** This level goes beyond mere association and begins to answer questions about the effect of interventions or actions. Here we can ask questions like "If we assign (i.e., do) $X = x$, what Y will occur?" For instance, we might assess the effect of assigning treatment with a new drug ($X = x$) on recovery (Y).

   Example estimand:

   $$\mathbb{E}[Y \mid do(X = x)] \tag{2.17}$$

3. **Counterfactual (Retrospective) Level:** The highest level of the hierarchy involves counterfactuals, which are retrospective questions about alternative possibilities, or nearby "worlds". At this level, we can ask questions like "What would have happened to Y if we had done X differently?". For instance, we might ponder whether a patient (Y) would have recovered faster if they had been given a different drug ($X = x'$).

   Example estimand:

   $$\mathbb{E}[Y_{x'} \mid X = x, Y = y] \tag{2.18}$$

   Notice the subscript x' in this last expression is a potential outcome notation, which is equivalent to an intervention or a do operation $do(X = x')$. We move to this notation in counterfactual worlds colliding with the one that actually took place $X = x$ to avoid confusion on the right-hand side of the conditioning bar.

Level	Questions	Example Estimand
Association	What if I see $X = x$?	$\mathbb{E}[Y \mid X = x]$
Intervention	What if I do $X = x$?	$\mathbb{E}[Y \mid do(X = x)]$
Counterfactual	What if I had done $X = x'$?	$\mathbb{E}[Y_{x'} \mid X = x, Y = y]$

Table 2.1: Summary of the Causal Hierarchy within the SCM framework.

Each level in the hierarchy requires its own set of assumptions to answer its particular class of queries. Importantly, the assumptions at each level also encompass those at the lower levels.

2.2.8 Estimation Approaches

Causal inference aims to estimate the effect of an intervention on a target outcome variable, given a set of observed variables. So far, we have mainly focused on identification strategies for causal effects, translating causal quantities into observational quantities. But once such a strategy is found – how do we proceed to estimate the actual causal effect? We review two relevant tools in this context: gradient-based optimization, which is one way of approaching elements needed for causal effect estimation after identification, and inverse probability weighting, which is an alternative view of identification.

Response Function Estimation. One way to estimate the causal effect of an intervention is to use response function estimation methods to learn a function that maps the observed variables to the target outcome variable. Specifically, we can formulate the causal effect as a function of the observed and intervention variable(s), and then learn the parameters of this function. This approach is based on the assumption that we have access to some observed variables (adjustment set) which are sufficient for estimating the causal effect and that the effect can be expressed as a function of these variables.

Formally, let Y be the target outcome variable, Z be the observed variables, and X be the intervention variable. We want to estimate the causal effect of X on Y (sometimes denoted as $\tau(X)$). We can define a function $f(X, Z; \theta)$ that maps the observed variables and the intervention variable to the target outcome variable, where θ are the parameters of the function. Then, we can learn the parameters θ that minimize the difference between the predicted outcome and the observed outcome:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}[(Y - f(X, Z; \boldsymbol{\theta}))^2]. \tag{2.19}$$

Once we have learned the function f, we can estimate the causal effect of the intervention as:

$$\tau(X) = \mathbb{E}[f(X = 1, Z; \theta_1)] - \mathbb{E}_Y[f(X = 0, Z; \theta_0)], \tag{2.20}$$

where θ_1 and θ_0 are the learned parameters for the intervention and the control group, respectively. We will make use of estimation approaches of this kind in Chapters 3, 4, 5, and 7.

Inverse Probability Weighting. Another approach for estimating causal effects is inverse probability weighting (IPW), which is based on the idea of re-weighting the observed data to balance the distribution of the intervention variable between the treatment and control groups.

Overall, it would still involve finding a relevant set of adjustment covariates, akin to the backdoor criterion (2.2.6). Crucially, however, it offers another view on the adjustment formula proposed earlier. Instead of going through a straightforward but costly marginalization of covariates as the previous one necessitates, especially as covariates get high dimensional in limited-sample size regimes, IPW offers an alternative that would only need to consider covariate values present in the available samples, rather than all possible values the covariates can take on. Thus, it can provide a more efficient route to the computation of causal effects.

The basic idea is to assign weights to each observation based on the probability of receiving the intervention given the observed covariates (often called propensity score, and computed via a logistic regression model), and then use these weights to estimate the average treatment effect.

Formally, let Y be the target outcome variable, Z be the observed variables, and X be the intervention variable. We are after an estimate of the causal effect of X on Y. We can define the probability of receiving the intervention given the observed variables:

$$e(Z) = P(X = 1 \mid Z) \tag{2.21}$$

which is the probability of receiving the treatment (we assume a binary treatment variable for simplicity of presentation), given covariates Z.

Then, given a dataset $\{(X_i, Z_i, Y_i)\}_{i=1}^N$ we can estimate the ATE (and with minor adjustments other causal effects) with the following:

$$ATE = \tau = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{Y_i X_i}{e(Z_i)} - \frac{Y_i(1 - X_i)}{1 - e(Z_i)} \right]. \tag{2.22}$$

X_i in this case is an indicator of whether treatment was received.

Inverse probability weighting is a widely used method for estimating causal effects and has been shown to be effective in various applications. However, it requires the estimation of the propensity score $e(Z)$, the quality of which depends on the completeness of the available covariates data.

IPW is also related to propensity score matching, another popular causal effect estimation method that relies on propensity scores. While IPW weighs each individual sample in the study by the inverse of their propensity score, giving rise to a re-weighted pseudo-population, propensity score matching involves the matching of pairs of groups of individuals with similar scores across the treatment and control groups, which allows the comparison of the outcome variable level between the two groups. IPW may be better suited when covariate distributions are imbalanced or where there are many covariates to control for; propensity score matching may be preferred under small sample size regimes. We will make use of the IPW estimator (albeit not directly for an ATE computation) as part of Chapter 7.

2.3 Expanding the Toolkit

2.3.1 Causal Discovery and Its Limitations

Many of the topics and methods discussed in this chapter so far rely on an understanding or fulfillment of assumptions about the data generation process underlying a dataset. This knowledge, within the SCM framework, is expressed via a DAG. But what happens when one does not know the causal graph behind a system of study? How may we proceed then?

One natural approach to address the lack of a pre-known DAG is to try and recover it from observational data using a causal discovery method, such as the constraint-based PC and FCI algorithms (Spirtes et al., 2000b), and in the absence of unmeasured confounding, the score-based Greedy Equivalence Search (GES) (Chickering, 2002; Glymour et al., 2019). However, such approaches in general suffer from some of the following difficulties. They:

- Require solving expensive combinatorial optimization problems,
- Involve numerous independence tests, which diminish in power as dimensionality grows, or

- Only find a Markov-equivalence class of graphs which all imply the same conditional independencies. Such an equivalence class would be insufficient when searching for an adjustment set since changes in edge directionality or the presence or absence of certain edges could lead to an erroneous choice of covariates for adjustment.

Fortunately, motivated by the task of estimating a specific causal effect of interest, we *do not need to recover the full causal graph.* We can do with partial knowledge thereof, as shown by recent works for covariate selection for the identification of a specific causal effect (VanderWeele and Shpitser, 2011b; Entner et al., 2013b; Witte and Didelez, 2019).

Entner et al. (2013b) already showed that using two testable independence and dependence conditions, a valid adjustment set can be found. Crucially, their method relies on relatively weak assumptions:

1. The existence of an auxiliary observed variable W (a weaker type of instrument we will further discuss in Chapter 3,

2. Partial ordering between the variables should be such that

   2.1. The outcome Y cannot be an ancestor of the treatment X, and

   2.2. The treatment X and outcome Y are not ancestors of any of the covariates Z and

3. The faithfulness assumption holds, which means all independencies in the joint distribution over all variables $P(V)$ are due to the structure of the graph (e.g., and not due to "chance" cancellation of paths) (Spirtes et al., 2000b).

Assuming these hold, the criteria Entner et al. (2013b) suggest can identify a valid adjustment set $Z^\star$,

$$\begin{aligned} &W \perp\!\!\!\perp Y \mid Z^\star \cup \{X\}, \\ &W \not\perp\!\!\!\perp Y \mid Z^\star. \end{aligned} \tag{2.23}$$

W in this case is an auxiliary "witnessing" variable which aids in the validation of $Z^\star$ as a valid adjustment set. These conditions ensure that all paths from W into Y are mediated by X. The original suggestion to operationalize the above conditions is to perform a greedy or random search.

In Chapter 3 We propose a more efficient procedure in the style of modern ML differentiable objective function optimization. This makes our method highly amenable

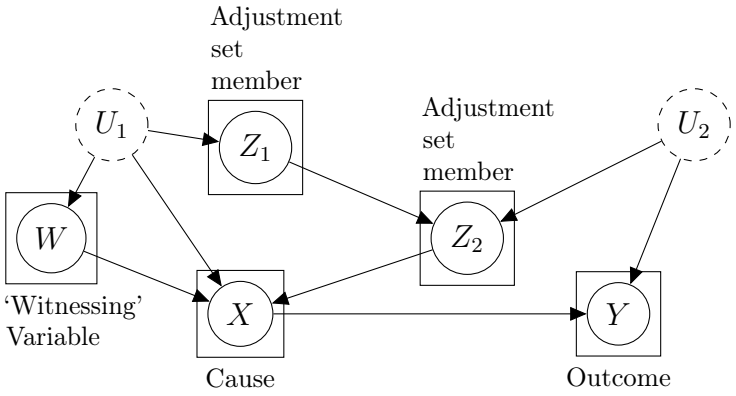


Figure 2.3: A DAG taken from Entner et al. (2013a) describing a setting where criteria 2.23 can be used to identify a causal effect of X on Y, using W and $Z^\star = \{Z_1, Z_2\}$ as the valid adjustment set.

to differentiable programming tools that contemporary software and hardware are designed for. This continuous optimization formulation shares similarity with Zheng et al., but unlike our targeted approach, tailored to a specific causal effect of X on Y, NOTEARS still aims to recover full DAGs (Zheng et al., 2018b). We argue that this is a harder problem than needs to be solved for most causal effect estimation tasks. In this sense, we see this work as an effort to connect the extensive literature on causal effect estimation with the no less illustrious body of work on causal discovery, into a single, targeted "supervised partial discovery" approach.

2.3.2 Effects of Crude Interventions and Mediation Discovery

Most related literature - and our background section so far - considers hard atomic interventions. A hard atomic intervention fixes the value of the treatment variable to some scalar (or vector if the variable is multi-dimensional) while keeping all other variables in the system at their original values. However, causes can be *complex* such that we cannot directly and entirely intervene upon them. This problem formulation, which we grapple with in Chapter 4, occurs in many domains such as those involving images, text, and complex networks (e.g. gene expression).

To define interventions in these settings we formalize a *crude intervention*. That is, since we cannot atomically intervene on a complex cause X, we consider "pressing a button" or "pulling a lever" in the form of atomic interventions on an *actionable* set of variables W, that are parents of X (see Figure 2.4). More precisely, we are looking to estimate the causal effect of intervention w, in the presence of covariates Z, on an

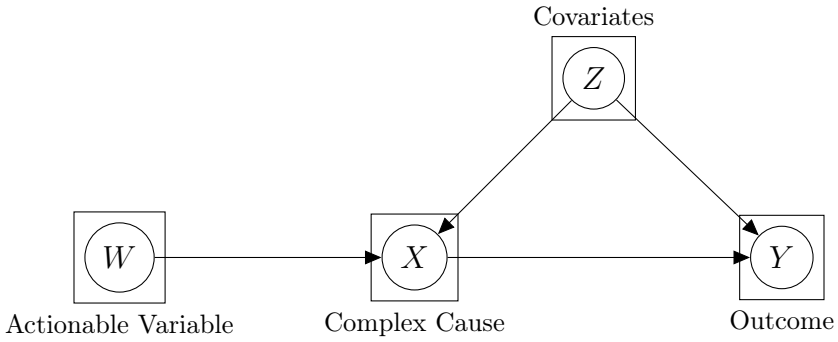


Figure 2.4: DAG describing the setting of the work presented in Chapter 4, looking to estimate causal effects in systems with complex causes (X) that one cannot atomically intervene on.

outcome Y,

$$\mathbb{E}[Y \mid Z, do(W = w)].$$

Beyond Hard, Atomic Interventions. Previous works have made the point that real-life problems often do not involve such perfect forms of interventions, and suggested relaxations to this idealized description. The most obvious form of relaxation moves from considering deterministically fixing intervention variables to a value, to a stochastic form, that only assigns a value to the treatment variable with a certain probability (Kocaoglu et al., 2019; Correa and Bareinboim, 2020; Huang and Valtorta, 2006). Crucially, such interventions do not operate with the same cleanliness on the causal graph. This led Correa and Bareinboim (2020) to suggest an extension of the do-calculus to handle such settings, which they named the σ-calculus. A clear distinction between our problem setup and such *soft* or *imperfect* interventions is that, while the control over the fixing of value as represented by the do-operator is no longer deterministic, those actions still affect specific target variables, which in principle can still be directly manipulated. In contrast our *crude* interventions of Chapter 4 target complex objects that cannot be fixed to a single value or be directly manipulated; furthermore, one way to think of our setting is that the crude intervention would be some lower-dimensional, or simpler intervention, applied to a high-dimensional complex cause with a possibly intricate internal structure. In that sense, we can only set a *process* running *through* the complex cause with subsequent influence on the outcome Y, and the intervention has "lower resolution" in comparison to the complex cause.

Unknown and Uncertain Interventions. Another extension to the vocabulary of interventions was offered in works on *uncertain* or *unknown* interventions. Those describe cases where the targets of interventions are unknown, in the sense that when an intervention is applied it might not affect a single target of outcome, but instead will affect a latent variable and in turn propagate effects through the causal graph (Eaton and Murphy, 2007; Jaber et al., 2020). Both works consider a motivating example from molecular biology for their problem formulation: when chemicals are added to a cell, no specific variables are set to specific values. Instead, some intervention nodes can be introduced into the graph, that might affect some latent variables and in turn, propagate effects into observed variables. While this setting sounds similar to ours, it focuses on full networks with edges connecting different variables affected by the unknown target interventions. In our case, instead, we consider a single complex cause that is being affected by the *crude* intervention, where the inner connections of elements within the complex cause might be unknown or simply hard to conceptualize: it is hard to discuss pixels in an image as causes of each other, and likewise with possible dimensions of text representation, e.g., dimensions of a word-embedding vector. Furthermore, while the goal of works such as Eaton and Murphy and Jaber et al. is to perform causal structure learning in the presence of added intervention nodes with such unknown targets, we are focused on *causal effect estimation*, such that mechanisms of mediation are discovered, and crucially, such that was can share parameters across *crude* intervention regimes, and thus predict the effects of rare or yet unseen treatments.

Invariant Causal Predictions. The notion of intervention regimes defining different environments, across which we would like to make stable predictions, gets us closer to recent literature on invariant risk minimization or invariant causal predictions, the motivation for which we briefly touched upon in Section 1.2 (Peters et al., 2016b; Heinze-Deml et al., 2018; Arjovsky et al., 2020). This line of work aims to exploit causal notions of invariance, motivated by the intuitive assumption that causal mechanisms should not change across datasets or environments of deployment: we expect the causal relations between height, pressure, and temperature to hold and operate similarly regardless of the specific country in which we carry out measurements; similarly, we expect the features that make a cow a cow or a camel a camel do not differ based on the background or conditions under which they are photographed, which should imply that we should encourage invariance in machine learning classifiers' predictions with respect to environmental conditions of the dataset (Arjovsky et al., 2020). Following this

logic, Peters et al. suggest exploiting invariance across conditions, and in particular interventional regimes, in order to perform causal discovery and offer confidence bounds for predicted graphical makeup. Arjovsky et al. in turn focus on utilizing the same idea for the estimation of causal predictors with special promise for generalization out of distribution. In particular, they suggest that enforcing such invariances of predictions with respect to background conditions that are not fundamental to targeted causal relations can be used to tell apart spurious associations discovered as correlations in datasets collected under the different background conditions, from the stable causal relationships that are not just incidental to a given setting. Our approach certainly bares similarity to the invariant causality literature – we too consider regimes, defined by the value of the *crude* intervention, and share parameters across them, exploiting stable causal relationships. However, we crucially try not only to identify parent nodes of outcome Y or use crude intervention W to report a stable prediction of $\mathbb{E}[Y \mid do(X = x)]$ (potentially additionally conditioning on covariates Z). Instead, we do not consider direct intervention on the entirety of X feasible or the main point of interest. We are focused on estimating $\mathbb{E}[Y \mid do(W = w)]$ (potentially conditioning on covariates Z), while also discovering mediators of the effect from W to Y.

Causal Abstraction. Another related body of works is the one focusing on causal abstraction (Chalupka et al., 2015, 2016a,b; Rubenstein et al., 2017; Beckers and Halpern, 2019; Beckers et al., 2020; Locatello et al., 2019). The understanding that observation level is not always indicative of a more high-level phenomenon that is the real target of analysis has been made clear in this line of work. The series of papers from Chalupka et al. showed how such a mismatch between granularity of observations and desired level of analysis plays out in image examples, as well as in climate data, such as wind and temperature measurements which underlay a more abstract weather phenomenon such as the "El-Niño" effect. The first paper proposed visual feature learning via a coarsening algorithm which led from low-level observations to more high-level feature creation and could lead to more stable causal predictions from a classifier trained on top of the coarsened learned features. The second paper applied a similar idea to the "El-Niño" case, and the last paper in this series studied more general multi-level cause-effect relations via the *Fundamental Causal Coarsening Theorem*. In particular, in that last paper, they considered a discrete setting unlike their previous works, where a macro-level effect is not specified. The works of Rubenstein et al. and Beckers and Halpern advanced the works of Chalupka et al. by taking a more theoretical point of view and defining a framework for the transition between causal

models defined at different levels of abstraction. In chronological order, this general theoretical work defined a notion of exact abstraction transformations and later relaxed those to approximate abstractions, where the high-level system is an approximate description of the low-level system. This notion of different resolutions of observations and analyses plays a key role in our setup description as well. Our description of a complex cause that we cannot manipulate directly and fully is inspired by the abstraction notion discussed above. However, it should be evident that in our approach we do not propose to transition from one abstraction level to another, but rather reason about low and high-level description jointly, via a lower-resolution intervention W compared to the complex cause X being studied. We consider observations of X, in its full complexity, as well as higher-abstracted descriptions in the form of $\Phi(X, Z)$. Furthermore, we leave room for the choice of construction of such high-level descriptions – they can be guided by domain knowledge hypotheses, as well as from more data-driven approaches of segmentation of the complex object (e.g. convolution windows in the case of image perturbation examples). A somewhat related notion of different abstraction levels between observations available and latent ones which are more directly related to the target of analysis is offered by the disentangled representation literature (Bengio et al., 2014; Locatello et al., 2019). Locatello et al. in particular focused on the feasibility of reaching disentanglement of independent latent explanatory factors from observed data and showed that it is impossible in an unsupervised manner without additional inductive bias baked into modeling[3].

Two-stage Estimation Procedures. So far, we have explored connections between our *problem setup and goals* to those previously appearing in the literature. At this point, we would like to make a final comparison to existing works relating to our *estimation strategy*, which is fundamentally a two-step one. Two-stage procedures for unbiased causal effect estimation have a long tradition in causal inference (Wright, 1928; Theil, 1958; Angrist and Imbens, 1995; Chernozhukov et al., 2018; Hahn et al., 2019; van der Laan and Starmans, 2014). The key idea behind such approaches is the separate computation of models for different conditional expectations (e.g., the propensity score as explored in Section 2.2.8) that are needed for the final conditional expectation representing the causal effect of interest. By breaking down the estimation goal into two models, rather than a single direct estimation, such methods propose to curb bias and the influence of confounding. In our own approach, we similarly learn two separate models in order to obtain less biased estimates. However, our goal is

[3]such as was proposed in (Leeb et al., 2020)

different from previous two-stage works, as was previously explained. We propose to estimate a different causal effect than the usual one, via the *crude* intervention W, and find mediators inside a set of possible high-level descriptions $\Phi(X, Z)$ of the complex object X on which the intervention via W is applied. Crucially, however, we do not deal in the current formulation with unmeasured confounding.

2.4 Trustworthy Machine Learning

As was established through the previous sections, causal modeling holds great promise for policy evaluation and design, as it relates to various fields of scientific inquiry and decision-making. In this section, we will focus on introducing topics in trustworthy machine learning (henceforth, trustworthy ML), some of which can be aided by causal concepts. We will survey works in explainable AI (XAI) and algorithmic fairness, and provide background for Chapters 5, 6, and 7.

2.4.1 Explainable Machine Learning

Explainable ML and AI (henceforth, XAI) emerged recently as an area of much interest in the ML literature. In particular, with the advance of deep learning (DL) methods that show high predictive accuracy, questions about their transparency and interpretability were raised by practitioners, researchers, and policymakers, especially as they relate to social data used to predict outcomes involving humans.

The practical needs emerging from the blackbox nature of ML and DL have led various researchers to weigh in. While some proposed methods to extract information from trained models in various ways, others have criticized their shortcomings, so much so that they called into question these efforts or the use of ML and DL for tasks to which interpretation is central (Rudin, 2019; Lipton, 2018).

In Chapter 5.3.2, we elucidate the goals of XAI via the causal concepts of sufficiency and necessity (Tian and Pearl, 2000). We suggest quantitative measures of those while explaining how these two basic goals can unify existing approaches in the field, that are seemingly disparate, seemingly conflicting at times, and are ripe for misuse (Mothilal et al., 2020a; Ramon et al., 2020; Fernández-Loría et al., 2020; Kumar et al., 2020). Further, we propose a sound and complete algorithm to quantify these measures under different levels of system knowledge and with different use cases in mind.

To prepare ourselves for the consideration of this proposed framework, we will first describe the landscape of the XAI debate to further help orientate Section 5.3.2 within it. While some authors propose to focus on models that are intrinsically interpretable,

such as sparse linear regression and rule lists (Rudin, 2019), we will focus instead on *post-hoc explanations* of more complex models. In fact, we will take a *model-agnostic approach*, explaining the output from any black-box model f, rather than focus on specific architectures or algorithms.

One axis that defines XAI methods is whether they are focused on *local* or *global* explanations. A local explanation looks only to apply to the immediate neighborhood around a specific data point of interest, e.g., one image or one row in a tabular dataset. A global explanation in turn aims to hold for the entire feature space. As will become apparent in Chapter 5.3.2, our focus will be on local explanations, but our method will also allow the exploration of the feature space when that is of interest, and so will include a flavor of global explanations.

Among the various methods proposed so far in the XAI literature, a few prominent families emerged.

2.4.1.1 Feature Attribution

Feature attributions, or the quantification of feature importance for explaining model predictions (Sundararajan and Najmi, 2020; Ribeiro et al., 2016), are some of the most popular explanation approaches for black-box ML models.

Feature attribution methods are techniques used in explainable AI to understand the contribution of individual features to a model's prediction. These methods provide insights into the inner workings of complex models and aid in identifying model biases, which can help improve the model's performance.

We will quickly survey four of the most popular among these approaches: LIME, SHAP (via Shapley values), and integrated gradients.

LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016). LIME is a technique used to explain the predictions of any black-box model. It generates explanations by creating a simpler, interpretable model that approximates the behavior of the black-box model in the local vicinity of the prediction. The interpretable model is trained on perturbed versions of the input data to obtain feature weights that indicate the contribution of each feature to the prediction.

LIME randomly samples instances similar to the input to be explained and fits an interpretable model (e.g. linear regression) on this sample. The learned local model is then used to provide explanations about the original prediction.

Mathematically, LIME computes the feature importance scores by minimizing a distance metric between the interpretable model's predictions and the black-box

model's predictions. Specifically, given an input x to be explained, LIME constructs an interpretable model g that locally approximates the black-box model f, and can provide an explanation with the following:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{2.24}$$

where G is the space of interpretable models, π_x is the probability distribution over similar instances (defined via a use-case specific distance metric), L is a distance metric between f and g, and Ω is a regularization term controlling the complexity of the local model. The feature importance scores are then computed based on the learned coefficients of the interpretable model.

Shapley Values (Shapley, 1953). Shapley values is a method for assigning values to each feature in a model by considering all possible combinations of features and their contributions to the prediction. Shapley values are based on the concept of cooperative game theory, which considers the contribution of each player (feature) in a coalition (model) to the overall outcome (prediction). The Shapley value of a feature is the average marginal contribution of the feature across all possible coalitions.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{(|N| - |S| - 1)!|S|!}{|N|!} (v(S \cup \{i\}) - v(S)) \tag{2.25}$$

where N is the set of all features, v is the prediction function, S is the set of features in a coalition, and $\phi_i(v)$ is the Shapley value of feature i under prediction function v.

SHAP (Shapley Additive Explanations) (Lundberg and Lee, 2017). SHAP is an extension that closely follows the Shapley values method introduced above, applied explicitly to ML models. SHAP offers an interpretation of the output of any model by combining the Shapley values with the concept of local explanations. It assigns importance scores to each feature by considering the contribution of the feature in the prediction for a particular instance.

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \tag{2.26}$$

where $\phi_i(f, x)$ denotes the SHAP value for feature i given a specific input instance x under a prediction model f. The sum is over all possible subsets z' of x', where x' is a simplified version of x, with binary values for each feature. x and x' are associated

via a mapping function, $x = h_x(x')$. The term $|z'|$ represents the number of nonzero entries in z' and M is the total number of features. The vector x_i is zero everywhere except at the ith position where it equals the corresponding entry from the instance x. The function $f_x(z')$ represents the expectation of the model's output when feature i is included in the subset z', while $f_x(z' \setminus i)$ denotes the expectation when feature i is excluded. Therefore, $f_x(z') - f_x(z' \setminus i)$ captures the marginal contribution of feature i to the prediction. The term $\frac{|z'|!(M-|z'|-1)!}{M!}$ serves as the weight of each subset size, ensuring that all possible combinations of features are considered and that larger subsets do not disproportionately influence the SHAP value. By summing over all possible subsets, we are effectively computing the average marginal contribution of feature i across all contexts.

Note that the resulting SHAP values are coefficients defining a linear additive explanation model g, ensuring that $g(z') \approx f(h_x(z'))$ whenever $z' \approx x'$. Lundberg and Lee show that such a g, defined via Shapely values, is the only possible explanation model that would satisfy the following desirable properties (Theorem 1).

- **Local Accuracy:** When approximating the original model for a specific input x, local accuracy requires the explanation model to at least match the output of the original model f for the simplified input x', and decompose as the sum of shapely values across features (including the baseline one, which is the prediction for of the original model when all features are set to 0 in the simplified version of the inputs, i.e., $\phi_0 = f(h_x(0))$). Overall,

$$f(x) = g(x') = \phi_0 + \sum_{i=i}^{M} \phi_i x'_i$$

- **Missingness:** If a feature's value is missing in the simplified instance, the SHAP value for that feature is 0 for that instance, i.e.

$$x_i^{'} = 0 \implies \phi_i = 0$$

- **Consistency:** As a model changes such that it relies more on a certain feature, the SHAP value of that feature should not decrease. In other words, for any two models f and f', if

$$f_x^{'}(z') - f_x^{'}(z' \setminus j) \geq f_x(z') - f_x(z' \setminus j)$$

for all inputs $z' \in \{0,1\}^M$, then

$$\phi_i(f', x) \geq \phi_i(f, x)$$

The SHAP method has been widely adopted due to its theoretical grounding and the axioms suggested above. However, they have also been criticized following their wide adoption, as we will explore when introducing our own XAI framework in Chapter 5.

Integrated Gradients (Sundararajan et al., 2017). Integrated Gradients is a feature attribution method that assigns importance scores to the inputs of a black-box model by integrating the model's gradient with respect to the input along a path from a baseline input (e.g., an input of all zeros) to the input of interest. The method is designed to satisfy several desirable properties, such as sensitivity and implementation invariance.

Formally, the Integrated Gradient of a feature i with respect to a prediction y for an input x is defined as:

$$\text{IntegratedGrads}_i(x) = (x_i - x_i^{baseline}) \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha, \qquad (2.27)$$

where x' is the baseline input, f is the black-box model, and x_i and $x_i^{baseline}$ are the values of feature i in the input and the baseline input, respectively.

The integral can be approximated using a numerical method such as the trapezoidal rule or the rectified trapezoidal rule.

The Integrated Gradient of a feature can be interpreted as the average marginal contribution of that feature to the model's output along the path from the baseline input to the input of interest.

2.4.1.2 Rule Lists

Anchors (Ribeiro et al., 2018a). The Anchors method is a model-agnostic feature attribution technique for Explainable AI. The Anchors method aims to generate high-precision explanations for individual predictions of ML models. The method identifies

a set of anchor rules: concise and human-readable conditions that describe a set of features and their corresponding values that are sufficient to guarantee a specific prediction.

The Anchors method defines an anchor as a binary decision rule (or conjunction thereof) defined over the feature values, which describe how the prediction of the model for a particular instance was achieved. Formally, given a prediction model f, an instance x, and the prediction $f(x) = \hat{y}$, an anchor A is a binary vector of the same dimensionality as x such that whenever A holds, the prediction $f(x) = \hat{y}$ holds with high probability.

The quality of an anchor is then assessed via two metrics: its precision and coverage.

- **Precision:** The precision of an anchor A is the expectation over the conditional perturbation distribution $D(z \mid A)$, which describes how likely it is that the model's prediction for a random instance that fulfills A matches the prediction of the model f for the instance x, i.e., $f(x)$.

  The precision can be defined mathematically as follows:

  $$\text{prec}(A) = \mathbb{E}_{D(z \mid A)}\left[\mathbb{I}(f(z) = f(x)\right] \tag{2.28}$$

  Here, $D(z \mid A)$ denotes the distribution of sampled instances z given that anchor A holds. The indicator function $\mathbb{I}(f(z) = f(x))$ is 1 if the model prediction $f(z)$ equals the instance's prediction $f(x)$, and 0 otherwise. Thus, the expectation $\mathbb{E}_{D(z \mid A)}[\mathbb{I}(f(z) = f(x))]$ gives the probability that a random instance z for which A holds is classified in the same way as $f(x)$.

  Note that under an arbitrary D and a black-box model f, it is intractable to compute this quantity directly. The authors therefore suggest a probabilistic reformulation, asking that their chosen anchors satisfy precision with high probability:

  $$P(\text{prec}(A) \geq \tau) \geq 1 - \delta \tag{2.29}$$

  τ is the pre-specified level at which a user would like precision to hold and δ controls the level of probability at which this statement would satisfactorily be held.

- **Coverage:** The coverage of an anchor A is the expectation over the dataset's distribution D, which describes how likely it is that the conditions specified by A hold for a randomly selected instance x.

The coverage can be defined mathematically as:

$$\text{cov}(A) = \mathbb{E}_{D(z)}[A(z)]$$

In this equation, $D(z)$ is the distribution of instances z in the dataset. The expectation $\mathbb{E}_{D(z)}[A(z)]$ gives the probability that anchor A applies to a random instance in the dataset z.

These two metrics allow the construction of the search problem for optimal anchors:

$$\max_{A \text{s.t.} P(\text{prec}(A) \geq \tau) \geq 1-\delta} \text{cov}(A) \tag{2.30}$$

It uses a two-step process for generating explanations. In the first step, the method generates a set of candidate anchors. In the second step, it utilizes a bandit algorithm to select the best anchor that explains the model's prediction.

Anchors is a local explanation method, meaning that suitable anchors to individual predictions might not provide insight into the global behavior of the model (a fact users may not always appreciate as the urge to generalize to rules that are true for the model as a whole can be strong). Furthermore, predictions that are near a black box decision boundary, or predictions of very rare classes may require very specific "sufficient conditions", and thus their anchors may be complex and provide low coverage. It is also possible multiple anchors may apply to the same instance.

2.4.2 Counterfactual Explanations

This family of methods involves the matching of input instances with proximal ones in feature space which receive counterfactual outcomes from the model (i.e., "flipped" outcomes for classification setting) (Wachter et al., 2018; Russell, 2019; Poyiadzi et al., 2020; Wexler et al., 2020). Note that term counterfactual here is not used in its causal sense (as we explored for example at the top of the hierarchy in Section 2.2.7), but rather in the sense that these methods are looking to base explanations on instances with contrary outcomes to the one seen for the input instance of inquiry.

Counterfactual Explanations (Wachter et al., 2018) Motivated by the "right to explanation" included in the European Union's GDPR regulations, Wachter et al. set out to propose an explanation method that can provide individuals with reasoning about the outputs of black box models. Crucially, the method relies on access to the outputs of the model, but does not require "opening" the black box itself: model

outputs can be explained without having a deeper understanding of the inner workings of the model. Instead, they offer users individual-level explanations: given input features and some original prediction by the model, the method provides the user with an example with minimal changes to its inputs that instead received a desirable output from the model. For example, for a user who got a loan rejected, a counterfactual explanation would point at an alternative instance, as close as possible to the inquiring user, who did get the loan.

To achieve this goal, the authors propose to minimize a loss function involving two terms: (1) the squared distance between the desired outcome and the model's prediction for some alternative input features x' we're searching over, and (2) some distance d between the user's original features x, and the alternative features' vector x'. By including both elements, this objective function can find suitable features x' that are both as close as possible to the inquiring user's original features, and yet see a model output that is as close as possible to the desired output y'. The resulting objective is the following

$$\arg\min_{x'} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x') \tag{2.31}$$

λ is a parameter maintaining a desired balance between the importance given to the distance between the original and alternative input (x, x'), and that between the model output and the desired target $(f(x'), y')$. The authors note that in practice, the maximization of λ is done by iteratively minimizing x', and then increasing λ until a sufficiently close solution is found.

As an advisable distance function, the authors propose the Manhattan distance weighted by the inverse median absolute deviation (MAD) of each feature.

$$d(x, x') = \sum_{i=1}^{p} \frac{|x_i - {x'}_i|}{\text{MAD}_j}$$

where p is the length of the features' vector. [4]

2.4.2.1 Causal Approaches to Explainability

The methods above seem disparate at first, may lead to different outputs and suggested conclusions to users under different use cases, and may be ambiguous in their interpretation. We will examine evidence of all these points in chapter 5. We will propose to unify them all under one framework, inspired by the causal notions of the

[4] MAD can be thought of as an alternative to variance, that is centered at the median rather than the mean: $\text{MAD}_i = \text{median}_{j \in P}(|X_{j,i} - \text{median}_{l \in P}(X_{l,i})|)$, where P is the set of points in the dataset.

probability of necessity and sufficiency, jointly named the probabilities of causation (Pearl, 1999; Tian and Pearl, 2000).

Probability of Necessity (PN). Let X and Y be two binary variables in a causal mode $\mathcal{M}$, let x and y stand for the propositions $X = \text{true}$ and $Y = \text{true}$, respectively, and x' and y' for their complements. The probability of necessity is defined as the expression

$$\begin{aligned} PN &\triangleq P(Y_{x'} = \text{false} \mid X = \text{true}, Y = \text{true}) \\ &\triangleq P(y'_{x'} \mid x, y) \end{aligned} \tag{2.32}$$

And its tightly linked converse is defined as,

Probability of Sufficiency (PS).

$$PS \triangleq P(y_x \mid y', x') \tag{2.33}$$

We will propose to build on these existing definitions by extending and refining them for greater expressiveness in the XAI context, which allows us to discuss factors (e.g., features, combinations of features, feature summaries or abstractions), and their necessity and sufficiency for obtaining certain outcomes of interest from a predictive system.

A more explicitly causal perspective on XAI methods has recently started to emerge (Chattopadhyay et al., 2019; Janzing et al., 2019). The incorporation of causal modeling is crucial when the goal is to understand not just the model itself, but processes in which dependencies between features of the system play a key part. Furthermore, the explicit consideration of the data-generation process can help focus the discussion around which explanations are actually sought in which setting: those that explain a specific model reliant on i.i.d. assumptions, or those that aim to explain a larger process happening in the world, beyond the narrow scope of a specific model (Janzing et al., 2019; Chen et al., 2020).

Another closely related line of work aims not only at finding explanations via influential feature or examples identification but defined useful explanations as ones that are actionable interventions and could help individuals understand how to better their outcomes in a future encounter with the system. This would be particularly relevant in settings where ML models are used to make high-stakes decisions for individuals, such as education admissions, hiring decisions, or loan granting. Karimi et al. (2020a) provides an overview of this emerging field of Algorithmic Recourse and a causal modeling approach to this problem was offered in Karimi et al. (2020b,c).

Algorithmic Recourse (Karimi et al., 2020b). The algorithmic recourse method is an approach for providing model explanations in terms of what changes can be made to the input data to achieve a desired outcome. However, unlike the XAI methods we surveyed in the previous section, it focuses on actionable steps an individual can actually take to achieve a desired outcome. To do so, it goes beyond the optimization described above and proposes to focus not simply on some potentially artificial realizations x' that can get a desired model outcome, but rather on interventions, borrowing from the causal tradition.

The optimization problem is formulated as follows:

$$\min_{A^* \in \arg\min_A} \text{cost}(A; \mathbf{x}^F) \text{ s.t.} \tag{2.34}$$
$$h(\mathbf{x}^{\text{SCF}}) \neq h(\mathbf{x}^F)$$
$$\mathbf{x}^{\text{SCF}} = F_A(F^{-1}(\mathbf{x}^F))$$
$$\mathbf{x}^{\text{SCF}} \in \mathcal{P}, A \in \mathcal{F}$$

In this optimization objective, the goal is to find a set of interventions A (from a set of feasible interventions F) that minimizes the cost function applied to the factual instance $\mathbf{x}^F$, and achieves a different prediction output from the model h. The constraints require that the intervention results in a counterfactual instance $\mathbf{x}^{SCF}$ that is in a pre-defined space P and has a different prediction than the original instance. The counterfactual instance is created by applying an intervention function F_A to the factual instance $\mathbf{x}^F$ and then mapping it back to the input space with F^{-1}. The specific form of the cost function and the set of possible interventions depend on the problem domain.

Unlike previous methods proposed in this space (notably Ustun et al. (2019)), algorithmic recourse uses the notion of intervention within a given structural causal model (SCM) $\mathcal{M}$ and thus relies on its knowledge. The optimization problem above provides Recourse through Minimal Interventions (MINT) for $\mathcal{M}$ and a set of feasible interventions. Solving Equation 2.34 to generate minimal interventions requires the computation of the structural counterfactual, denoted as $\mathbf{x}^{SCF}$, for the individual $\mathbf{x}^F$ in $\mathcal{M}$ under any feasible action, A. To compute the counterfactual, the SCM $\mathcal{M}$ is assumed to be an additive noise model (ANM). The counterfactual $\mathbf{x}^{SCF}$ can then be derived deterministically using the Abduction-Action-Prediction procedure (Pearl, 2000), assuming cross-world assumptions hold.

We will connect Algorithmic Recourse, as a form of counterfactual explanation as described in Section 2.4.2, via our proposed framework. We will also introduce a

complete and sound algorithm based on the necessity and sufficiency of interventions on factors.

2.4.3 Algorithmic Fairness

The notion of algorithmic recourse we just explored in Section 2.4.2 is also tightly linked with algorithmic fairness. While multiple statistical metrics were proposed in the last decade to conceptualize of fairness of predictive algorithms (Dwork et al., 2012; Kleinberg et al., 2017; Chouldechova, 2017; Corbett-Davies and Goel, 2018) and (Barocas et al., 2019, Chp. 2), an emerging section of this literature started promoting the advantages of causal modeling to this problem (Kusner et al., 2017b, 2019).

2.4.3.1 Supervised Prediction Systems

The increasing use of ML methods to guide decisions in our everyday lives has placed the problem of designing fair ML methods at the mainstream of research in the field. In its machine learning variant, work on the impact of automated decision-making on protected groups goes back at least to the early 2010s (Kamishima et al., 2011; Dwork et al., 2012; Zemel et al., 2013). However, it gathered much steam after the next event.

The COMPAS Debate. Following the publication of ProPublica's critique of the criminal recidivism predictive system COMPAS in 2016, a debate on how to quantify and correct for potential harm done to marginalized groups by predictive systems has begun (Larson et al., 2016). Various papers (Zafar et al., 2017b,a; Woodworth et al., 2017; Calmon et al., 2017; Agarwal et al., 2018; Kusner et al., 2017b; Chiappa, 2019; Dwork et al., 2018) were dedicated to the design, enforcement, and critique of fairness notions for decision making systems with societal impact. As a rejoinder to the ProPublica analysis, multiple groups of researchers reached an impossibility result that has enjoyed a place of prominence. The result relied directly on the fairness notions ProPublica measured COMPAS against: *separation* (related to balance for the positive/negative class, equalized odds, equality of opportunity, and conditional procedure accuracy equality), and *sufficiency* (related to calibration within groups, test-fair score and conditional use accuracy equality).[5]

[5]See (Barocas et al., 2019) for a comprehensive survey, and the 'dictionary of criteria' appearing in Chapter 2 therein for a handy summary.

Group Fairness Definitions And Impossibility. The data distribution together with a classifier, generates a probability distribution on the triple $(\hat{Y}, Y, A)$, where $\hat{Y}$ is the predicted label, Y the true label, and A the protected attribute. Both measures of fairness are defined in terms of conditional independence among these three quantities. *Sufficiency* is defined as $Y \perp\!\!\!\perp A \mid \hat{Y}$ and *separation* as $\hat{Y} \perp\!\!\!\perp A \mid Y$. The impossibility result states that, if one has an imperfect predictor and unequal base rates across protected attribute of the studied outcome phenomenon in the population of interest (i.e. $Y \not\perp\!\!\!\perp A$), then it is impossible for both *separation* and *sufficiency* to hold (Kleinberg et al., 2017; Chouldechova, 2017; Corbett-Davies et al., 2017; Berk et al., 2018). This is an often-stated fact in the field of algorithmic fairness, which might lead one to conclude it is impossible to optimize for both notions of fairness when training or consider both metrics when assessing the fairness of systems. A practitioner would thus be forced to choose one or the other, attributing it greater subjective importance in a given use case.[6] Kleinberg et al. also proved an approximate version, showing that a (slightly unusual) measure of accuracy γ has to be close to 1 if suitably quantified versions of separation and sufficiency approximately hold. In Chapter 6 we discuss this in greater detail, and provide an arguably more interpretable result in terms of more standard notions of accuracy based on precision and recall.

We take a different perspective on the impossibility result. Theoretically, we provide a more detailed picture of how to achieve approximate sufficiency and separation, by showing that values of suitably quantified metrics can be held below a certain level while guaranteeing a bound on the accuracy even when the base rates are different ($Y \not\perp\!\!\!\perp A$). In particular, using this theoretical result, one can quantify how unfair *any* algorithm with a certain level of accuracy must be. Concretely, this indicates that an algorithm that returns a predictor that is more unfair than this level can be made more fair *without* sacrificing accuracy. This allows us to compare different algorithms on the basis of how close they are to achievable trade-offs. On the practical side, we show how these quantitative measures can be directly optimized achieving different trade-offs. In particular, we consider $\Delta^{\text{suff}}_{y,\hat{y},a} = |\mathbb{P}[Y = y|\hat{Y} = \hat{y}, A = a] - \mathbb{P}[Y = y \mid \hat{Y} = y]|$ and $\Delta^{\text{sep}}_{\hat{y},y,a} = |\mathbb{P}[\hat{Y} = \hat{y} \mid Y = y, A = a] - \mathbb{P}[\hat{Y} = \hat{y} \mid Y = y]|$. We treat $\Delta^{\text{suff}}_{\max} = \max_{y,\hat{y},a} \Delta^{\text{suff}}_{y,\hat{y},a}$ and $\Delta^{\text{sep}}_{\max} = \max_{\hat{y},y,a} \Delta^{\text{sep}}_{\hat{y},y,a}$ as representing a deviation from perfect sufficiency and separation and minimize an upper bound over these quantities.

[6]Note that other works also consider the group-specific fairness notion requiring $\hat{Y} \perp\!\!\!\perp A$, called independence in (Barocas et al., 2019). That would be the case when trying to enforce strict affirmative action, for example, requiring an equal acceptance rate for each group, disregarding their observed features and base rates. We focus on conditional-independence fairness notions, as they are often more realistic in several cases.

Enforcing Group Fairness. Multiple works considered forms of constrained optimization of predictors to hold one of the fairness notions we consider in this work (Zafar et al., 2017a; Wu et al., 2019a; Donini et al., 2018). However, as far as we know, they consider holding a notion of either the Δ^{suff} or the Δ^{sep} family; none took our approach, of trying to address both at the same time and explicitly to do so as a means to achieve a quality predictor, possibly due to recommendation in works such as Pleiss et al. (2017) or similar intuitions following the impossibility result. Furthermore, previous authors studied the fairness and accuracy trade-off via Pareto Frontiers (Kearns et al., 2018; Kearns and Roth, 2019). However, unlike the Pareto frontiers explored in the aforementioned, we focus instead on extending the impossibility result involving holding both fairness notions of Δ^{suff} and Δ^{sep} in an approximate fashion and propose optimization methods inspired by it. The possible trade-off between each of these notions and accuracy is a secondary point to our work, and we show empirically that we can largely achieve better tradeoffs between Δ^{suff} and Δ^{sep}, without great loss in accuracy. There were also more recent works on fairness relaxations (Lohaus et al., 2020) and group and subgroup fairness (Martinez et al., 2020; Diana et al., 2020). While they all raise important and interesting points about how to hold one chosen family of notions in a properly relaxed version that comes with guarantees, or exploring subgroup fairness, they deal with implications of previous works, usually still referring to the impossibility result. Our focus is in a way more fundamental, going back to the original result, for just two groups, defined by a single sensitive attribute, and in the binary classification setting as a starting point.

2.4.3.2 Fair Policy Optimization

While in the section above we introduce the membership in a sensitive group as A, we switch to S in the context of fair policy optimization, as we will need A to refer to the action node that our regime indicator/optimized policy parametrization will act on.

As explored above, great attention has been given to the problem of developing supervised fair prediction models, resulting in a variety of methods for enforcing relations of the model's outcomes $\hat{Y}$, predictions of the ground truth of Y, with respect to membership in a sensitive group S. Those are tailored to different unfairness settings:

- Independence of the outcomes across groups may be suitable to settings in which dependence is deemed unfair (demographic parity, $\hat{Y} \perp\!\!\!\perp S$, connected to Barocas et al.'s independence, $Y \perp\!\!\!\perp S$), or

- Independence of the errors across groups which may be more relevant in settings in which dependence of the outcomes is deemed fair (equality of opportunity (EoP), closely related to Barocas et al.'s separation, $\hat{Y} \perp\!\!\!\perp S \mid Y$).

In contrast, relatively little attention has been given to the problem of designing optimal policies that have some fairness guarantees (Joseph et al., 2016, 2018; Gillen et al., 2019; Kusner et al., 2019; Nabi et al., 2019; Chohlas-Wood et al., 2021). In such a case, the goal is to design a decision-making system that specifies how to select actions that maximize some downstream outcome of interest from the decision of the system, subject to some fairness constraints. One example of such a problem is the creation of a policy for the allocation of funds or program participation that can achieve a desirable social outcome without penalizing certain demographic groups.

Optimizing Policies with Fairness Constraints. We are instead interested in the decision-theoretic problem of selecting a policy that gives a more favorable outcome, and ask that the distribution of *downstream outcomes* is fair. The solution to this problem cannot be obtained by using fair predictors. In other words, we are not concerned with simply learning predictions $\hat{Y}$ of Y in a fair way, i.e. by imposing constraints on the relations between $\hat{Y}$ and S. Instead, we are concerned with changing the data-generating process itself to give rise to a more favorable outcome Y. Consider, e.g., the problem of fair loan allocation: a fair predictor of loan repayment would simply change the rates at which loans were granted, even if the person taking the loan will not be able to repay it. Thus, using such a predictor as a criterion for giving someone a loan could become counterproductive: in some cases, it could worsen the downstream outcome of the unsuccessful re-payers, or even expose them to exploitation. The appropriate approach would be to ask for a policy that maximizes the probability of repayment, subject to fairness constraints. This issue has been studied under the terms "self-fulfilling prophecy" and "delayed impact of fair ML" in Dwork et al. (2012) and Liu et al. (2018), respectively, where deploying a particular fair predictor $\hat{Y}$ can cause unfavorable changes in the actual eventual outcome of interest.

Another example of this phenomenon may emerge when recruiting underrepresented students to higher-education institutions. While admission might rely on the prediction of success, and we can penalize such predictions such that they offer more spots to such groups, we may still not provide them with the resources and support to fulfill their potential and actually succeed, which is the downstream outcome we are in fact interested in. D'Amour et al. (2020) suggested and provided a tool to explore this issue further in simulated multi-stage online settings.

In Chapter 7, we propose a causal framework for learning optimal policies with fairness constraints that are inspired by the public health literature (Jackson and VanderWeele, 2018; Jackson, 2018, 2020). Unfairness is defined as the presence of differences in the distribution of the downstream outcome Y stratified by levels of chosen sensitive attributes S, where that distribution depends on a policy that a decision maker is responsible for selecting. We reason about the problem by treating sensitive attributes as effect modifiers rather than causes, and by considering the actual effect of our actions, as opposed to planning under an idealized outcome distribution that removes the unfair contributions of the sensitive attributes. By doing so, we diverge from literature in this field, and in particular the causal variants of it (Kusner et al., 2017b, 2019; Nabi et al., 2019). We do so while still leveraging causal modeling for the use of historical data in the estimation of effects under the new policy we optimize.

Like us, Nabi et al. (2019) uses observational data, but in order to learn a policy as if particular path-specific effects between S-A and S-Y were completely deactivated. They do not place any constraints on $p(y_{\sigma_A} \mid s, x)$, where σ_A parameterizes the newly learned policy. Chohlas-Wood et al. (2021) considers a more general utility function than ours as the optimization objective but focuses on enforcing a notion similar to demographic parity on the choice of the policies, rather than considering a fairness notion on the outcomes. It also does not consider individual-level effects, and cannot be extended to continuous S (e.g., age) as easily as our approach.

Still, in the observational data regime, the most similar work has been Kusner et al. (2019), which also does not consider individual-level effects and parameterized policy spaces. It is also different in its motivation: it consists mostly of budget treatment allocations and interference problems and requires manipulation of the sensitive attribute S.

Considering interventional data in the online setting, Joseph et al. (2016, 2018) derive algorithms that have "meritocratic fairness," requiring that, at every round, no arm with a lower expected reward is preferred to one with a higher expected reward. The challenge is that this guarantee should hold with respect to the true expected rewards, which can only be estimated from data. This goal can be seen as an individual-level fairness metric, but it is not at odds with a policy that maximizes utility; instead, its main concern is to offer more conservative exploration, by treating individuals similarly if they are plausibly identical given the current data. This fairness notion does not consider protected classes and is orthogonal to the notions considered in our work. A similar notion, combined with calibration, was explored in Liu et al.

(2017), and in the Reinforcement Learning setting in Jabbari et al. (2017). Recently, fairness in bandit settings has received a lot of additional attention (Huang et al., 2021; Schumann et al., 2022; Patil et al., 2020). Recent work in the fair bandit literature includes Huang et al. (2021) which explores the extension of counterfactual fairness (Kusner et al., 2017b) to a UCB algorithm. Schumann et al. (2022) assumed rewards received for arms are biased due to some societal factor, and propose to learn a bias term to correct for it. Chen et al. (2019) instead conceptualized fairness in terms of a minimum rate at which a task or resource is assigned to a user; Patil et al. (2020) suggested a requirement to pull each arm at least a given fraction of total rounds; and Ron et al. (2021) enforces fairness in a setting where each arm represents a sub-population, and needs to be pulled at least some number of times according to a budget. Other works exploring fairness in the personalized recommendation systems context are Burke (2017); Burke et al. (2018); Ekstrand and Kluver (2020); Huang et al. (2020); Zhu et al. (2018); Celis et al. (2018). In Chapter 7, we explain in more detail why our suggested settings may be more suitable for certain fairness use cases.

3 | Differentiable Causal Backdoor Discovery

3.1 Introduction

Causal modeling allows one to go beyond observational quantities to estimate the *causal effect* of interventions in the real world (Pearl, 2000; Rosenbaum, 2017). Most commonly, we are interested in how a single outcome Y varies as we change the values of a decision or treatment X. However, without knowledge of the true causal graph, estimating the causal effect from observational data can be confounded by unobserved variables. While it is possible to "adjust" for these confounding variables, one must know what observed variables to use in the adjustment. Otherwise, the adjusted estimate can be worse than the unadjusted estimate (Pearl, 2009a).

One way to solve this is to use causal discovery algorithms (Spirtes et al., 2000a; Peters et al., 2017) to identify as much of the causal graph as possible, and then make an adjustment. However, there are a number of problems: 1. In general, causal discovery algorithms involve expensive combinatorial optimization problems; 2. Traditional nonparametric causal discovery methods (Spirtes et al., 2000a) do not discover the full causal graph, but a Markov-equivalent set of graphs that imply the same set of conditional independences. This is an issue for adjustment as edge existence and directionality play key roles in whether a variable should be used in an adjustment or not; 3. Nonparametric methods also require combinations of multiple tests of independence constraints which suffer from rapidly diminishing power as dimensionality grows.

Our contribution is to provide a continuous optimization approach to the problem of learning what variables should be used for adjustment. Although the optimization is non-convex and our practical implementation makes parametric assumptions, our search problem has important benefits compared to typical causal discovery methods (Spirtes et al., 2000a; Peters et al., 2017). Specifically, it avoids all of the problems listed above as follows: 1. Instead of resorting to combinatorial search, our method uses gradient-based optimization that runs extremely fast on modern hardware to discover the adjustment set; 2. Our method does not need to know the full connectivity

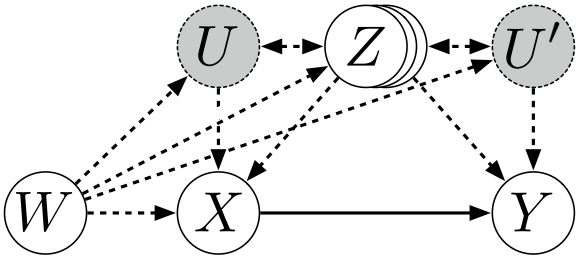


Figure 3.1: An informal visualization of our assumed knowledge of the causal graph. Treatment X must be a parent of outcome Y, and dashed arrows/nodes are optional as long as a subset $Z^\star$ of Z blocks the backdoors between X and Y, and W has no unblocked path into Y given X and $Z^\star$. Additionally, W must be associated with X (and possibly confounded).

of the causal graph. All that is needed is to identify a suitable auxiliary variable W and covariate set Z which precede the treatment X and outcome Y; 3. Our method directly targets functions of the covariate space Z that are useful for covariate adjustment. Thus it does not directly perform high-dimensional multiple-testing and does not assume a small number of parents for the variables of interest.

Our goal is to provide a generally-applicable algorithm to be added to the toolbox of the practitioner while focusing on the "causal supervised learning" problem of targeting a given cause-effect pair (X, Y) as opposed to the full graph learning problem (Zheng et al., 2018a). We will begin by reviewing causal primitives necessary for our work in the next section.

3.2 Background

Using the notation of Pearl (2000), we let the quantity $P(Y = y \mid do(X = x), Z = \mathbf{z})$ describe the variability of Y under an intervention that fixes X at value x, conditioned on observations $\mathbf{z}$ of a set of random variables Z (throughout: non-bold lower-case variables x are scalars, bold lower-case $\mathbf{z}$ are vectors, non-bold capital X are either single random variables or a set of random variables in the case of Z). This is different from the usual conditional probability of Y given $\{X = x, Z = \mathbf{z}\}$, denoted, $P(Y = y \mid X = x, Z = \mathbf{z})$.

We can formalize interventional probabilities using a directed acyclic graph (DAG) in the following way. Given a DAG $\mathcal{G}$ with vertex set $\{V_1, V_2, \ldots, V_d\}$, let $\text{pa}_{\mathcal{G}}(i)$ be the parents of V_i in $\mathcal{G}$. When each vertex corresponds to a random variable, this DAG induces a probabilistic model. In this model the probability (density or mass) function over $\{V_1, \ldots, V_d\}$ factorizes as $\prod_{i=1}^{d} p(v_i \mid \text{pa}_{\mathcal{G}}(i))$ (Lauritzen, 1996). Each

edge into V_i in the graph qualitatively describes a contribution to the model factor $p(v_i \mid \mathrm{pa}_{\mathcal{G}}(i))$. Given the DAG and associated model, the intervention $do(V_A = v_A)$ corresponds to a two-step procedure: (a) remove all model factors $p(v_A \mid \mathrm{pa}_{\mathcal{G}}(A))$ (i.e., all edges entering V_A in the graph), and (b) fix the value of V_A to v_A in any other factor where V_A is a parent node. We say that a model is *causal* if the *do-operator* $do(\cdot)$ is defined for it. If so, then $\mathcal{G}$ is a *causal graph*, and V_i *causes* V_j only if V_i is an ancestor of V_j in $\mathcal{G}$.

One key summary of the interventional distribution $P(Y \mid do(X = x))$ is the *average treatment effect* (ATE),

$$\mathbb{E}[Y \mid do(X = x')] - \mathbb{E}[Y \mid do(X = x)],$$

defined for two treatment levels x and x'. Another effect of interest is the partial derivative $\partial \mathbb{E}[Y \mid do(X = x)]/\partial x$, which we will also refer to as the ATE whenever X is continuous. This will be our main case in the sequel. The classic way to estimate ATE is via a randomized control trial. However, with observational data only, all we can directly estimate is the joint distribution and an adjustment will be necessary based on the causal graph. In particular, if there are common causes of X and Y (*confounders*), then off-the-shelf regression of Y on X can be severely biased.

To fix this, one approach is *covariate adjustment*: find a set $Z^\star \subseteq Z$ of ancestors of X or Y in $\mathcal{G}$ that can "block" such common causes and apply a formula such as the *backdoor adjustment* (Pearl, 2000),

$$p(y \mid do(X = x)) = \int_{\mathbf{z}^\star} p(y \mid x, \mathbf{z}^\star) p(\mathbf{z}^\star) d\mathbf{z}^\star. \tag{3.1}$$

We say that $Z^\star$ is a *valid* covariate set if it satisfies the above. Notice that the marginalization is with respect $p(\mathbf{z}^\star)$ instead of $p(\mathbf{z}^\star \mid x)$, as the link between X and its ancestors in $\mathcal{G}$ is broken by the *do* operator.

Finding valid covariate adjustments. If the full graph is known, there is a graphical criterion by which we can test whether $Z^\star$ is a valid set for covariate adjustment (cf. Pearl, 2000, for details). However, specifying a full causal graph is often difficult, particularly when all we need is to provide a valid covariate set for a given cause-effect pair (X, Y). As formalized by VanderWeele and Shpitser (2011a), partial knowledge of the causal structure may suffice.

Consider the causal setup in Figure 3.1 (solid arrows indicate a causal link, and dashed arrows indicate a causal link may or may not exist). Let $Z \cup \{W\}$ be a known set of observed non-descendants of $\{X, Y\}$, and U, U' possible unobserved parents

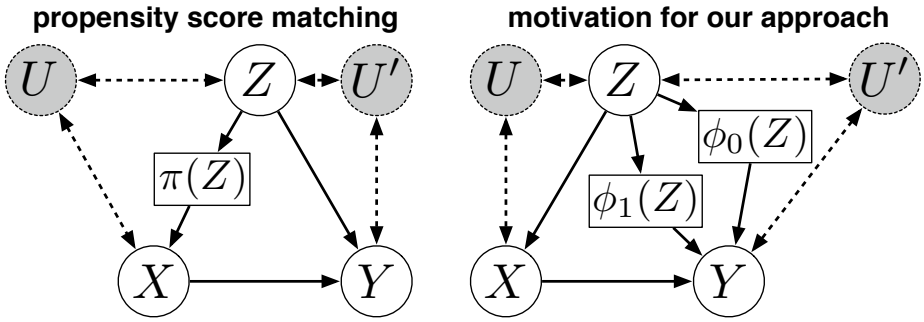


Figure 3.2: On the left, a representation of the propensity score $\pi(z) \equiv p(x \mid z)$ as a vertex that satisfies the backdoor criterion once placed in the graph. On the right, an analogous representation, in terms of two extra vertices, of the outcome function $P(Y = 1 \mid z, x)$ for $x = 0$ (ϕ_0) and $x = 1$ (ϕ_1). Here, the value taken by X works as a selection indicator of a mixture model so that, for binary Y, $p(y \mid x, z) = \phi_x(z)^y(1 - \phi_x(z))^{1-y}$. In fact, $\phi_x(z)$ can be the result of any invertible transformation of $P(Y = 1 \mid x, z)$.

of $Z \cup \{X\}$ and $Z \cup \{Y\}$. Assuming faithfulness (Spirtes et al., 2000a)[1], Entner et al. (2013a) observed that it is possible to recover the causal effect $X \to Y$ between treatment X and outcome Y so long as an observed adjustment set $Z^\star \subseteq Z$ satisfies the following criterion for some observed variable W:

$$\begin{aligned} &W \perp\!\!\!\perp Y \mid Z^\star \cup \{X\}, \\ &W \not\perp\!\!\!\perp Y \mid Z^\star. \end{aligned} \tag{3.2}$$

That is, $Z^\star$ is a valid covariate set "certified" by an auxiliary variable W. In a simplified sense, W plays the role of a pseudo "intervention indicator" into X with all paths from W into Y mediated by X. We choose to focus in this work on the difficult step of finding a valid $Z^\star$. An auxiliary variable W can be found in linear time by an outer loop, or by choosing based on background knowledge (which still requires much weaker conditions than full graph elicitation). In fact, different criteria can be used to combine multiple candidate Ws (Silva and Evans, 2016). Our contribution therefore is in identifying a valid covariate adjustment set, and we assume the existence of a suitable W in the following. The benefit of the above criteria is that it relies on much weaker partial ordering assumptions as opposed to knowing a full graph.

Our work. Finding a valid covariate set satisfying these criteria, in general, requires combinatorial optimization on Z. This is usually done by greedy/random search

[1]In fact, the required faithfulness is just w.r.t. the relationship between the auxiliary variable W and the outcome Y.

(Entner et al., 2013a). In this chapter, we propose instead a fully-differentiable optimization problem for learning a backdoor adjustment. Instead of attempting to find the exact adjustment set our approach finds a set of functions

$$\Phi(\mathbf{z}) \equiv \{\phi_x(\mathbf{z}) \mid x \in \mathcal{X}\},$$

where $\mathcal{X}$ is the sample space of X and $\mathbf{z}$ is in the sample space of $\mathbf{Z}$, such that

$$\begin{aligned} &W \perp\!\!\!\perp Y \mid \{\phi_X(Z^\star), X\}, \\ &W \not\perp\!\!\!\perp Y \mid \Phi(Z^\star). \end{aligned} \tag{3.3}$$

We will show that, under some general conditions, covariates $Z^\star$ satisfying Eq. 3.3 will also satisfy Eq. 3.2. Importantly, to simplify the presentation, we will assume that either Y is binary or the whole causal system is linear Gaussian. This will allow us to define $\phi_x(\mathbf{z})$ to be scalars for any $\mathbf{z}$[2]. The extension to non-binary Y or non-linear Gaussian systems is conceptually simple, but notation gets considerably more involved.

For intuition about why this is true, first consider a graphical representation of the propensity score, $\pi(\mathbf{z}) \equiv \Pr(X = 1 \mid \mathbf{z})$ added to a postulated causal graph $\{Z \to X, Z \to Y, X \to Y\}$ and binary X, This is shown in Figure 3.2: here informally Z is a set of vertices, where a single vertex $\pi(Z)$ (see Hernán and Robins, 2019, Chapter 15, for a more formal discussion) blocks all backdoors between X and Y and hence is a valid adjustment variable. This is particularly helpful as a way of reducing the dimensionality of the problem if we can reasonably estimate $\pi(Z)$. However, discovering this backdoor adjustment by finding a suitable $Z^\star$ will not be possible if, for instance, W and $Z^\star$ are adjacent in the causal graph: we will still need to explicitly condition on $Z^\star$ when verifying the independence between W and Y.

An alternative is to consider the analog to the propensity score with respect to the outcome variable Y, as illustrated by the following example.

Example 1. Assume all variables are binary, with $\log[\Pr(Y=1 \mid X, Z)/(1-\Pr(Y=1 \mid X, Z))] = (1-X)\beta_{yz0}^\top Z + X\beta_{yz1}^\top Z$ and $\log[\Pr(X=1 \mid W, Z)/(1-\Pr(X=1 \mid W, Z))] = (1-W)\beta_{xz0}^\top Z + W\beta_{xz1}^\top Z$, with W being an exogenous variable. Then we can check that $\Phi(Z) \equiv \{\phi_0(Z) \equiv \beta_{yz0}^\top Z, \phi_1(Z) \equiv \beta_{yz1}^\top Z\}$ will satisfy $W \perp\!\!\!\perp Y \mid \{\phi_X(Z), X\}$ and $W \not\perp\!\!\!\perp Y \mid \Phi(Z)$. The former can be shown by noting that we can predict Y purely from X and $\phi_X(Z)$; no further information about W will help. The latter can be verified by observing that W provides further information about X, which we can use

[2]More precisely, we assume we can write $p(y \mid \mathbf{z}', x) \equiv h(y, x, \phi_x(\mathbf{z}'))$ for some $h(\cdot)$ and any subvector $\mathbf{z}'$ of $\mathbf{z}$.

to refine our prediction of Y. A graphical illustration of this idea is shown in Figure 3.2. $\square$

This suggests that if we parameterize $\phi_x(\mathbf{z})$ to be in the same family of the response of Y given X and any other set of observable covariates, we will be able to directly search for this representation without performing high-dimensional tests of conditional independence. However, this raises the immediate concern of what to do if X is continuous, as in this case $\Phi(\mathbf{z})$ is uncountable. We can compress the information in $\Phi(\mathbf{z})$ by making further assumptions about the outcome regression model, as shown in the following example.

Example 2. Assume that $Y = \beta_{yx}X + \beta_{yz}^\top Z + \varepsilon_y$ and $X = \beta_{xw}W + \beta_{xz}^\top Z + \varepsilon_x$ describe the conditional distributions of Y and X, with W being an exogenous variable and $\varepsilon_x, \varepsilon_y$ being independent error terms. Then we can check that $\phi_X(Z) \equiv \beta_{yz}^\top Z$ for all X will satisfy $W \perp\!\!\!\perp Y \mid \{\phi_X(Z), X\}$ and $W \not\perp\!\!\!\perp Y \mid \Phi(Z)$. $\square$

We will prove the existence of a solution of Eq. 3.3 that solves Eq. 3.2 in the following section. This suggests we can obtain a valid adjustment from the optimization of functions $\phi_X(\cdot)$. Instead of searching for exact conditional independence, our approach is to minimize dependence measures motivated by Eq. 3.3. In the parametric case, we will derive a continuous optimization problem, avoiding greedy/random selection (Entner et al., 2013a). While continuous optimization methods exist for discovering the entire causal graph (Mooij et al., 2009; Zheng et al., 2018a), our technique is tailored to discovering a backdoor adjustment.

It is tempting to see our definition of $\phi_X(Z)$ as a similar idea to propensity scores (Hernán and Robins, 2019), as the examples suggest: can we make Y and Z independent given X and $\phi_X(Z)$, so that covariate adjustment can be done directly with $\phi_X(Z^\star)$ as opposed to $Z^\star$? Unfortunately, this is not true: eq. Eq. 3.3 can only identify an equivalence class of $\phi_X(Z)$, not all of which will be valid adjustments on their own. For instance, in the first example, regressing W using X and Z will also satisfy eq. (3.3). Therefore, when we solve for $\phi_X(Z)$, our goal is to discover which variables $Z^\star \subseteq Z$ should make up its domain. In the next section, we give conditions that allow us to identify a suitable adjustment set $Z^\star$ via $\phi_X(\cdot)$, and we describe how to solve for it.

3.3 Method

We now describe the general problem formulation, starting with the idealized scenario where we know the true population observational distribution. Practical implementations of this formulation for linear models with homoscedastic errors are discussed in the sequel.

Let $d(V_i, V_j \mid S)$ be a measure of probabilistic dependence between random variables V_i and V_j given a set of random variables S. Let $d(\cdot, \cdot \mid \cdot)$ have the following properties: (a) it is non-negative, and (b) it equals zero if and only if $V_i \perp\!\!\!\perp V_j \mid S$ [3]. An example of such a measure is the conditional mutual information. In linear models, absolute partial correlation could be used. Let $\phi_X(\mathbf{z})$ have a parametric representation, with θ_X being the respective parameters. Let sparsity$(\{\theta_X\})$ be a penalty term that induces sparsity in this set of parameter vectors θ, e.g. $\sum_x \|\theta_x\|_1$. We define the following optimization problem for $\{\theta_X\}$:

$$\begin{array}{rl} \text{minimize} & d(W, Y \mid X, \phi_X(Z)) \\ \text{subject to} & d(W, Y \mid \Phi(Z)) > \alpha, \\ & \text{sparsity}(\{\theta_X\}) < c. \end{array} \tag{3.4}$$

3.3.1 Theory Behind Learning ϕ_X

In this section, we will present the theoretical justification of our method. The main idea is to show the following: i) if $W \perp\!\!\!\perp Y \mid Z^\star \cup \{X\}$ for some $Z^\star$, then there exists some scalar $\phi_X(Z^\star)$ where $W \perp\!\!\!\perp Y \mid \{\phi_X(Z^\star), X\}$; ii) if $W \perp\!\!\!\perp Y \mid \{\phi_X(Z^\star), X\}$, then $W \perp\!\!\!\perp Y \mid \{X\} \cup Z^\star$ up to some "general" arrangement of the parameters of the model; iii) under *faithfulness* (conditional independences in the data arise from conditional independences in the causal graph), we can search for a $Z^\star$ satisfying ii), and use it to estimate the ATE using the backdoor adjustment with adjustment set $Z^\star$. All results assume the partial ordering described in Figure 1, and that for simplicity of presentation Y is binary or the system is linear-Gaussian so that each $\phi_x(\mathbf{z})$ can be written as scalar.

Point i) was implicitly discussed in the previous section, and it is formalized here for the general case where some elements of $Z^\star$ are not parents of Y:

Lemma 1. If $W \perp\!\!\!\perp Y \mid Z^\star \cup \{X\}$, then there exists some scalar $\phi_X(Z^\star)$ such that $W \perp\!\!\!\perp Y \mid \{\phi_X(Z^\star), X\}$.

[3] If d is a probabilistic dependence measure, (a) and (b) are necessary and sufficient conditions to define a valid optimization problem as described in Eq. 3.4.

Proof. We will discuss the case for binary Y, as the proof for linear-Gaussian models follows the same idea. Let the structural equation for Y be given by $f_y(x, \pi_Z, \pi_U)$, where π_Z and π_U are the observed and unobserved parents of Y in the corresponding causal graph. The conditional distribution of Y is given by

$$\begin{aligned} &p(y \mid x, \mathbf{z}^\star) = \\ &p(f_y(x, \pi_Z, \pi_U) = 1 \mid x, \mathbf{z}^\star)^y \times \\ &(1 - p(f_y(x, \pi_Z, \pi_U) = 1 \mid x, \mathbf{z}^\star))^{1-y}. \end{aligned}$$

By assumption, $f_y(\cdot)$ is functionally independent of W. Now we just have to show that the random variable $f_y(x, \pi_Z, \pi_U)$ is conditionally independent of W given X and $Z^\star$. Since $W \perp\!\!\!\perp Y \mid Z^\star \cup \{X\}$, it cannot be the case that W and $\pi_{\setminus Z^\star, X}$, the parents of Y not in $Z^\star \cup \{X\}$, are conditionally dependent given $Z^\star \cup \{X\}$. We define $\phi_X(\mathbf{z}^\star)$ as $p(f_y(x, \pi_Z, \pi_U) = 1 \mid x, \mathbf{z}^\star)$ for each possible realization of X. Given X, we can fully reconstruct from $\phi_X(\mathbf{z}^\star)$ a conditional distribution of Y that makes information about W irrelevant. □

The result for point ii) is as follows. To simplify the presentation, we assume that Z follows a multivariate discrete distribution, but this is not essential. We also define $\Phi^f_{x\mathbf{z}^\star}$ to be subset of the sample space of $Z^\star$ such that $\phi_x(\mathbf{z}^\star) = f$ for all $\mathbf{z}^\star \in \Phi^f_{x\mathbf{z}^\star}$. We assume the causal model is parametric with a Lebesgue measure on the parameter space. As an abuse of notation, we sometimes use $p(x_i, X_j = f, \dots)$ to mean the pmf $p(x_i, f, \dots)$ of the joint distribution of $X_i, X_j, \dots$, when it is not obvious that "f" is a value taken by X_j.

Theorem 1. If $W \not\perp\!\!\!\perp Y \mid Z^\star \cup \{X\}$, and

$$\begin{aligned} &\sum_{\mathbf{z}^\star \in \Phi^f_{x\mathbf{z}^\star}} p(y \mid w, x, \mathbf{z}^\star) \frac{p(\mathbf{z}^\star \mid w, x)}{\Pr(Z^\star \in \Phi^f_{x\mathbf{z}^\star} \mid w, x)} \neq \\ &\qquad \sum_{\mathbf{z}^\star \in \Phi^f_{x\mathbf{z}^\star}} p(y \mid x, \mathbf{z}^\star) \frac{p(\mathbf{z}^\star \mid x)}{\Pr(Z^\star \in \Phi^f_{x\mathbf{z}^\star} \mid x)}, \end{aligned} \tag{3.5}$$

for some value f in the range of $\phi_x(\cdot)$, then $W \not\perp\!\!\!\perp Y \mid \{\phi_X(Z^*), X\}$.

Proof. Assume, contrary to the hypothesis, that $W \perp\!\!\!\perp Y \mid \{\phi_X(Z^\star), X\}$. Then

$$p(y \mid w, x, \phi_x(Z^\star) = f) = p(y \mid x, \phi_x(Z^\star) = f) \Rightarrow$$

$$\sum_{\mathbf{z}^\star} p(y \mid w, x, \phi_x(Z^\star) = f, \mathbf{z}^*) p(\mathbf{z}^* \mid w, x, \phi_x(Z^\star) = f) =$$

$$\sum_{\mathbf{z}^\star} p(y \mid x, \phi_x(Z^\star) = f, \mathbf{z}^\star) p(\mathbf{z}^\star \mid x, \phi_x(Z^\star) = f) \Rightarrow$$

$$\sum_{\mathbf{z}^\star} p(y \mid w, x, \mathbf{z}^\star) p(\mathbf{z}^\star \mid w, x, \phi_x(Z^\star) = f) =$$

$$\sum_{\mathbf{z}^\star} p(y \mid x, \mathbf{z}^\star) p(\mathbf{z}^* \mid x, \phi_x(Z^\star) = f) \Rightarrow$$

$$\sum_{\mathbf{z}^\star \in \Phi^f_{x\mathbf{z}^\star}} p(y \mid w, x, \mathbf{z}^\star) \frac{p(\mathbf{z}^\star \mid w, x)}{\Pr(Z^\star \in \Phi^f_{x\mathbf{z}^\star} \mid w, x)} =$$

$$\sum_{\mathbf{z}^\star \in \Phi^f_{x\mathbf{z}^\star}} p(y \mid x, \mathbf{z}^\star) \frac{p(\mathbf{z}^\star \mid x)}{\Pr(Z^\star \in \Phi^f_{x\mathbf{z}^\star} \mid x)},$$

which contradicts the hypothesis. □

One observation: equation (3.5) is a relationship that we do not expect to hold in general, even if it can hold for particular parameter arrangements. For instance, if for simplicity $Z \perp\!\!\!\perp W \mid X$, then the last step of the derivation above can be written as

$$\sum_{\mathbf{z}^\star \in \Phi^f_{x\mathbf{z}^\star}} p(y \mid w, x, \mathbf{z}^\star) \frac{p(\mathbf{z}^\star \mid x)}{P(Z^\star \in \Phi^f_{x\mathbf{z}^\star} \mid x)} =$$

$$\sum_{\mathbf{z}^\star \in \Phi^f_{x\mathbf{z}^\star}} p(y \mid x, \mathbf{z}^\star) \frac{p(\mathbf{z}^\star \mid x)}{P(Z^\star \in \Phi^f_{x\mathbf{z}^\star} \mid x)},$$

this is $\mathbb{E}[p(y \mid w, x, Z^\star) - p(y \mid x, Z^\star) \mid X = x, \phi_x(Z^\star) = f] = 0$. This is not an identity if $p(y \mid w, x, \mathbf{z}^\star) \neq p(y \mid x, \mathbf{z}^\star)$. In this sense, we say that "in general" $W \not\perp\!\!\!\perp Y \mid \{X, Z^\star\}$ implies $W \not\perp\!\!\!\perp Y \mid \{X, \phi_X(Z^\star)\}$ so that the contrapositive holds.

The main exception to avoid is the case where $\phi_x(\mathbf{z}^\star)$ is functionally independent of some elements of $Z^\star$. A simple example is the graph $\{W \to Z_1 \leftarrow U \to Y, W \to X, X \to Y\}$: inequality (3.5) *will* be violated for $\phi_x(z_1) =$ constant under *any* parameterization, even though we expect $W \not\perp\!\!\!\perp Y | \{Z_1, X\}$. Hence, it is crucial to return functions $\phi_X(\cdot)$ that are as sparse as possible.

Our method then follows from using the optimization problem in eq. (3.4) to find a function $\phi_x(\mathbf{z}^\star)$ for each x, that is *sparse*. Specifically, it may be a function of a strict subset of Z. Assuming (a) that the conditions of Theorem 1 hold; (b) that our choice of W, variable selection algorithm, and function space for $\phi_X(\cdot)$ can minimize $d(W, Y \mid X, \phi_X(Z))$ all the way to zero; (c) W is a parent of X so that the second condition of Entner et al. (2013a) is satisfied ($W \not\perp\!\!\!\perp Y \mid \{X\} \cup Z^\star$), then by faithfulness we recover a valid adjustment set for the causal effect of X and Y.

Given these results, we will show in the next section how to optimize a fully-differentiable objective for covariate adjustment search that does not require solving a full graph discovery problem. The price to be paid is the assumption about the existence of an auxiliary variable W. This assumption is nevertheless falsifiable by simply testing whether we can indeed minimize our objective function to zero. Note that technically we need to prove similar results for the second term in Eq. 3.3 ($W \not\perp\!\!\!\perp Y \mid \phi_X(Z^\star)$) to show that we do not need the edge $W \to X$. We omit those proofs for simplicity of presentation, as they are nearly identical. We now describe an optimization procedure to find a differentiable backdoor adjustment $\phi_X(Z)$.

3.3.2 Optimization & Implementation

The shape of the function $\phi_X(Z)$ should be decided based on the assumptions about the outcome model $p(y \mid x, \mathbf{z}^\star)$. We will consider Gaussian models. In particular, consider the following,

$$X \sim \mathcal{N}(g_{\text{con}}(Z, W, U), \sigma_X^2)$$
$$Y \sim \mathcal{N}(h_{\text{con}}(X, Z, U), \sigma_Y^2),$$

where $g_{\text{con}}(\cdot,\cdot,\cdot), h_{\text{con}}(\cdot,\cdot,\cdot) \in \mathbb{R}$ are unknown functions. One possible modeling assumption for $g_{\text{con}}(\cdot,\cdot,\cdot), h_{\text{con}}(\cdot,\cdot,\cdot)$ which we consider here is that they are linear functions of their inputs. Specifically, this implies that: $p(y \mid x, \mathbf{z}) = p(y \mid \alpha x + \boldsymbol{\beta}^\top \mathbf{z})$ for some (α, β). Thus, we could set $\phi_X(\mathbf{z}) \equiv \phi(\mathbf{z}) \equiv \boldsymbol{\beta}^\top \mathbf{z}$. In this case we use absolute partial correlation $|\rho(\cdot,\cdot|\cdot)|$ as our probabilistic dependence measure as follows

$$d(W, Y \mid X, \phi(Z)) = |\rho(W, Y \mid X, \phi(Z))|.$$

To learn $\boldsymbol{\beta} \in \mathbb{R}^d$, the parameters of $\phi(Z)$, an initial idea would be to solve for it directly by minimizing the absolute partial correlation. However, recall that the partial correlation is computed from a ratio of terms. In our case, both the numerator and denominator include $\boldsymbol{\beta}$. This makes the partial correlation scale invariant with respect to $\boldsymbol{\beta}$. Thus, we reparameterize $\boldsymbol{\beta}$ as $\boldsymbol{\beta} \equiv \boldsymbol{\gamma}/\|\boldsymbol{\gamma}\|_2$ and propose to optimize the Lagrange equivalent of Eq. 3.4, which is:

$$\min_{\boldsymbol{\gamma}} \left|\rho(W, Y \mid X, \boldsymbol{\beta}^\top Z)\right| - \lambda_1 \left|\rho(W, Y \mid \boldsymbol{\beta}^\top Z)\right| + \lambda_2 \|\boldsymbol{\beta}\|_1,$$
$$\text{s.t. } \boldsymbol{\beta} \equiv \boldsymbol{\gamma}/\|\boldsymbol{\gamma}\|_2.$$

Although the above objective is non-convex, it is differentiable everywhere except $\boldsymbol{\beta} = 0$ and hence amenable to optimization using gradient-based methods[4]. In the next section, we demonstrate our method on simulated graphs and graphs fit on real-world data compared to other practical baselines.

3.4 Experiments

To evaluate our method we begin by devising a high-dimensional simulation benchmark. This benchmark will allow us (a) to test the robustness of our technique to different causal graph parameter settings, and (b) to study the sensitivity of our method to different noise and causal effect parameters. Additionally, we devise a causal graph based on real-world health-worker survey data. We compare our method with practical baselines including Entner et al. (2013a). In all cases, our method matches or outperforms all baselines. Code to replicate experiments and run on new data will be released at `https://github.com/limorigu/Diff-causal-backdoor-disc`.

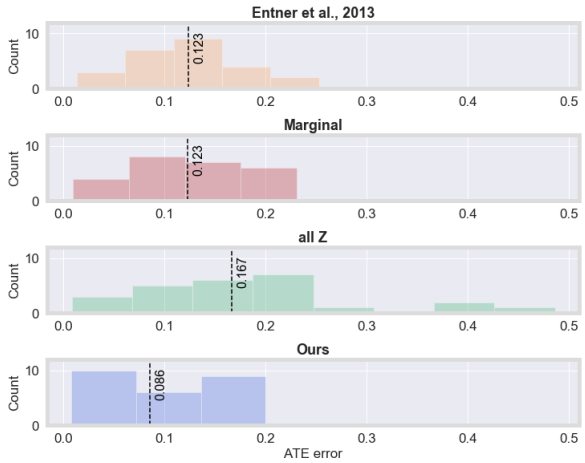
3.4.1 Simulation Benchmark

We design a simulation benchmark to test the robustness and sensitivity of our method. The causal graph is shown in Figure 3.5. The covariates Z are (secretly) grouped into four sets Z_1, Z_2, Z_3, Z_4. While adjusting for Z_3 allows us to correctly estimate the ATE of X on Y, adjusting for Z_1 or Z_2 skews this estimate. We observe all variables except U, U' and assume we have identified W, X, Y, Z but do not know their connectivity except that it satisfies the partial ordering of Figure 3.1. Finally, as described in Section 3.3.2 we will assume that the structural equations are linear Gaussian, which allows us to compare directly to Entner et al. (2013a).

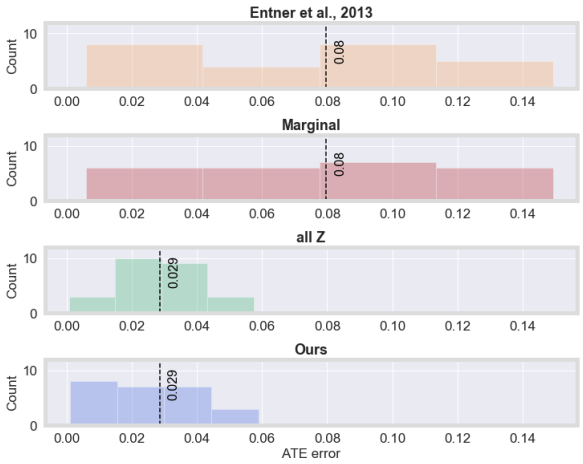
The structural equations for the simulation are:

$$
\begin{aligned}
&U \sim \mathcal{N}(0, \mathbf{I}), U' \sim \mathcal{N}(0, \mathbf{I}), W \sim \mathcal{N}(0, 1), \\
&Z_1 \sim \theta_{Z_1,W} W + \boldsymbol{\Theta}_{Z_1,U} U + \boldsymbol{\Theta}_{Z_1,U'} U' + \mathcal{N}(0, \sigma^2_{Z_1} \mathbf{I}), \\
&Z_2 \sim \theta_{Z_2,W} W + \mathcal{N}(0, \mathbf{I}), Z_3 \sim \mathcal{N}(0, \sigma^2_{Z_3} \mathbf{I}), Z_4 \sim \mathcal{N}(0, \mathbf{I}), \\
&X \sim \boldsymbol{\theta}^\top_{X,U} U + \boldsymbol{\theta}^\top_{X,Z_2} Z_2 + \boldsymbol{\theta}^\top_{X,Z_3} Z_3 + \mathcal{N}(0, \sigma^2_X), \\
&Y \sim \boldsymbol{\theta}^\top_{Y,U'} U' + \boldsymbol{\theta}^\top_{Y,Z_2} Z_2 + \boldsymbol{\theta}^\top_{Y,Z_4} Z_4 + \omega X + \mathcal{N}(0, \sigma^2_Y)
\end{aligned}
$$

[4]This optimization problem could achieve different solutions that all admit the same partial correlations but satisfy different criteria. These criteria could be used to choose among all admissible solutions. See D'Amour and Franks 2021 for an example where the authors characterize a full solution set for a similar optimization problem and motivate a choice among them based on representation overlap.



(a) absolute ATE error for each method on datasets with lower noise on treatment ($\sigma_X^2 = 0.01$).



(b) absolute ATE error for each method on datasets with higher noise on treatment ($\sigma_X^2 = 0.6$).

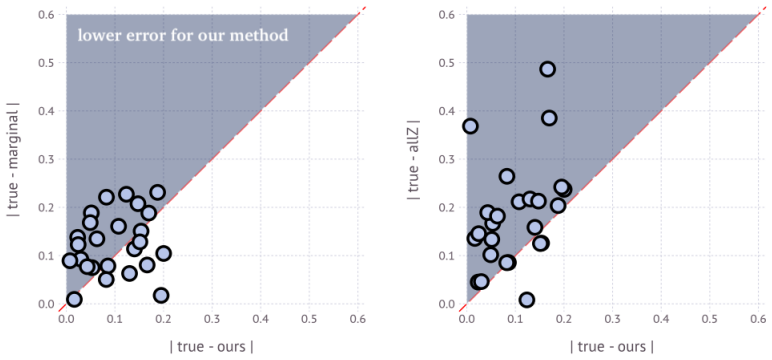
Figure 3.3: Histograms of ATE error of all methods on a simulation with lower treatment effect ($\omega = 0.1$).

where Θ signifies matrices of parameters. Our goal is to learn $\phi(Z)$ in order to correctly estimate the ATE: ω.

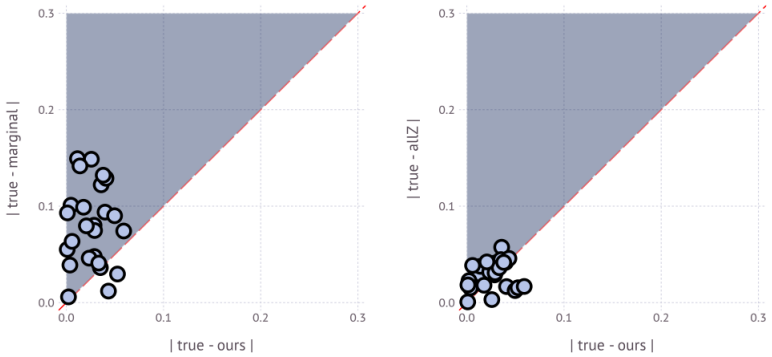
Baselines. We compare our approach with a set of practical baselines that make similar assumptions:

1. (Entner et al., 2013a): We compare against the ATE estimated by the high-dimensional search algorithm of Entner et al. (2013a).

2. **all** Z: This uses all covariates Z in the adjustment set to compute the ATE.



(a) ATE error of baselines vs our method for simulations with lower noise on treatment ($\sigma_X^2 = 0.01$).



(b) ATE error of baselines vs. our method for simulations with higher noise on treatment ($\sigma_X^2 = 0.6$).

Figure 3.4: Scatter plot comparing baselines and our method on a simulation with lower treatment effect ($\omega = 0.1$). Each point is one of the 25 parameter settings. Points in blue regions indicate our method performs better.

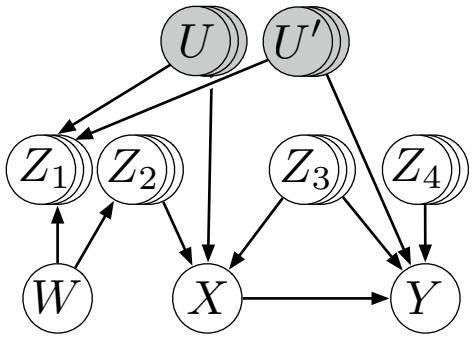
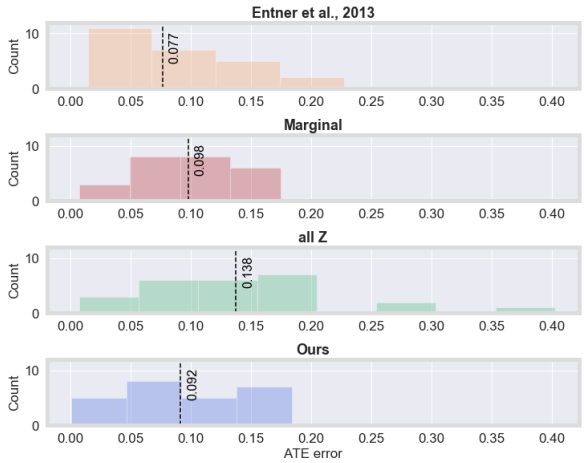
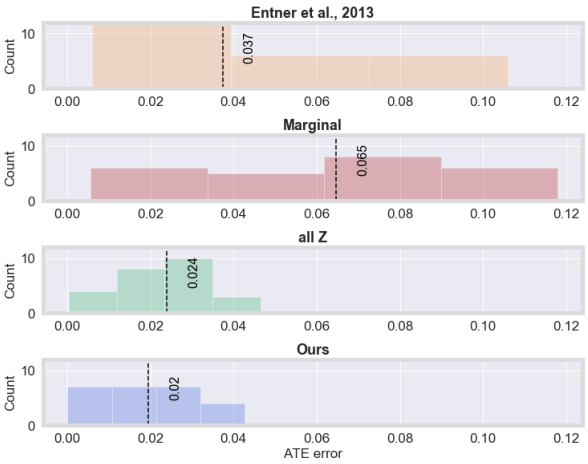


Figure 3.5: The causal graph used in our simulation. There are four (unknown) covariate sets Z_1, Z_2, Z_3, Z_4, two unobserved variable sets U, U', an auxiliary variable W, treatment X, and outcome Y. Notice that the minimal backdoor adjustment set is Z_3 while adjusting for Z_1 or Z_2 can adversely affect the estimation of the average treatment effect (ATE) of X on Y.



(a) absolute ATE error for each method/baseline on datasets with lower noise on treatment ($\sigma_X^2 = 0.01$).



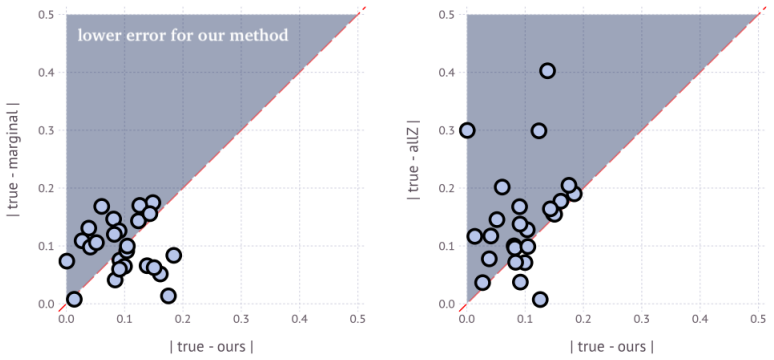
(b) absolute ATE error for each method/baseline on datasets with higher noise on treatment ($\sigma_X^2 = 0.6$).

Figure 3.6: Our method's performance compared to the marginal, all Z and Entner et al.'s method baselines on a simulation with higher treatment effect ($\omega = 0.5$).

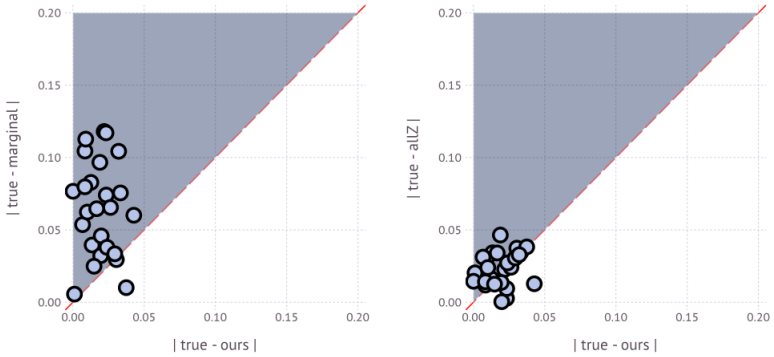
3. **marginal**: This uses no covariates in the adjustment set to compute the ATE.

As described in Section 3.2, there exist many algorithms for covariate selection (Witte and Didelez, 2019). However, none of them work under the weak assumptions of Figure 1, apart from Entner et al. (2013a). Below, we simulate problems in which either baseline 2 or 3 perform well. Our experiments illustrate that our method is competitive even with baseline 1 which is tailored for linear Gaussian models while matching or improving upon the best baseline. Thus, without knowledge of the true causal model, our model is state-of-the-art.

Experimental setup. To evaluate the robustness of all methods we sample 25



(a) Comparison for simulation with lower noise on treatment ($\sigma_X^2 = 0.01$).



(b) Comparison for simulation with higher noise on treatment ($\sigma_X^2 = 0.6$).

Figure 3.7: Performance of our method against the marginal and the all Z baselines on a simulation with a higher effect of treatment ($\omega = 0.5$). Scatter plots represent 25 different datasets.

different parameter settings for the above structural equations (we fix the true ATE ω throughout, more on this in the following paragraph). We sample parameters from a standard Gaussian distribution and then take the absolute value. We then generate signs for parameters by drawing a value from a uniform random variable and flipping the current sign if the value is above a certain threshold. This ensures that sampled parameters are not mean zero and thus will not likely cancel each other out (i.e., violating faithfulness). For each setting, we then sample 20,000 inputs and normalize each variable to have unit variance. We split these inputs 50/50 into train/test. We fix the dimension of each variable $Z_1, Z_2, Z_3, Z_4, U, U'$ to 30.

Hyperparameter tuning. We tune all hyperparameters of our method including λ_1, λ_2, the initialization of $\boldsymbol{\gamma}$, and the learning rate η on the training set. We use cross validation to select λ_1, initial $\boldsymbol{\gamma}$ and η. To select λ_2, we perform a hypothesis test with

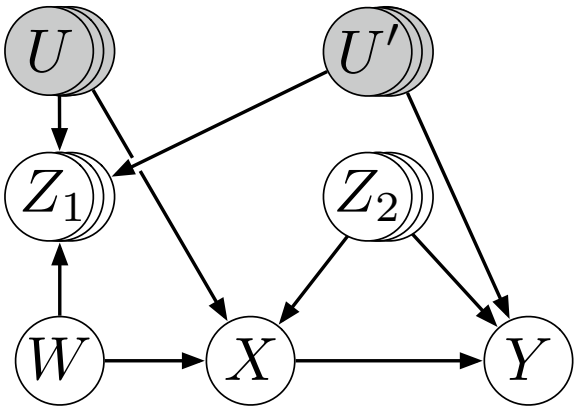


Figure 3.8: DAG representing the NHS dataset we use to test the performance of our model.

null $\rho(W, Y|\boldsymbol{\beta}^\top Z) = 0$. If we reject the null, we increase λ_2 and re-optimize (adjusting significance using Bonferroni correction) until we do not reject the null.

Results. To test the sensitivity of our method we introduced two types of problems: i) problems with high noise on treatment X ($\sigma_X^2 = 0.6$), and ii) problems with lower noise on treatment X ($\sigma_X^2 = 0.01$). Problem i) will cause the marginal baseline to perform poorly, while problem ii) will cause the all Z baseline to perform poorly. Additionally, for each problem, we tested two settings of the true ATE: $\omega = 0.1$ and $\omega = 0.5$. Thus, we have 2 problem types, each with 2 different ATEs, tested over 25 parameter settings, for a total of 100 different simulation scenarios.

Figures 3.3 and 3.4 show test results on the low ATE setting ($\omega = 0.1$) in two different ways. Figure 3.3 shows a histogram of absolute ATE error for each method in both treatment noise settings ($\sigma_X^2 = 0.01$) and ($\sigma_X^2 = 0.6$). The black dashed line indicates the median of the distributions. In both cases, our method matches or outperforms all other methods as measured by the median performance. Figure 3.4 shows the distribution of performance for each individual parameter setting, compared with the marginal and all Z baselines. Points in the blue regions indicate trials where our method outperformed the baseline methods. In the low noise plots (a) our method consistently outperforms the all Z method, while in the high noise plots (b) our method noticeably outperforms the marginal method.

Figures 3.6 and 3.7 present the same type of results for the setting with the higher effect of treatment X ($\omega = 0.5$). In this setting, the high-dimensional method of Entner et al. (2013a) deteriorates. As in the previous settings, our method does better than two baselines, and at least as well as the third. This means that when one encounters an unknown dataset for which we want to estimate the ATE, one doesn't need to

Ours	All Z	Marginal	Entner et al. (2013a)
0.163	0.324	1.747	0.771

Table 3.1: The absolute ATE error for all methods on the NHS health data.

guess whether adjustment using all, none, or some other set of covariates is best. By applying our method one can get accurate ATE estimates regardless of the setting.

3.4.2 NHS Health Data

Alongside the simulation benchmark, we test our method on a causal graph derived from real-world health data. Specifically, we consider data from the 2014 UK National Health Service (NHS) Survey Picker Institute Europe (2015). The aim of the survey was to "gather information that will help to improve the working lives of staff in the NHS". We consider the goal of trying to understand the causal effect of workplace training on personal well-being.

We construct a set of variables by averaging the results of related questions (where most questions are on a five-level Likert scale: from 'strongly disagree' to 'strongly agree'). Specifically, we identified an auxiliary variable W: whether an individual underwent workplace training (Q1 in the survey), a treatment X: one's benefit (or not) from training (Q2), outcome Y: whether an individual's job is good for their well-being (Q14), and covariates Z_1: a set of variables describing personal job satisfaction (Q5-Q9), and Z_2: a set describing the effectiveness of one's organization/managers (Q10-Q12, Q18-Q21). Based on their descriptions we describe the relationships between these variables using the causal structure in Figure 3.8. Additionally, we introduce unobserved variables into this semi-simulated setting, U and U'. We suggest those newly introduced variables could be interpreted, for example, as one's "openness/ability to learn", and "personal affinity for their job", respectively. We use linear Gaussian structural equations for the model and fit the parameters of the model using the real data, 25 variables in total. Once fit, we sample 90,000 data points and partition them into 30,000 train/valid/test splits. We ran each method on sampled data-points (and tuned hyperparameters on the validation set) and evaluated them on the test set.

Results. Table 3.1 shows the ATE error of each method on the NHS health data. Our method outperforms all other methods. It is worth noting that for this dataset, the all Z baseline is much closer to the real ATE than the marginal baseline. Compared to the all Z baseline, our method removes a harmful node in Z_1, which leads to a better ATE estimate.

3.5 Conclusion

In the spirit of Entner et al. (2013a), we showed how causal discovery for covariate adjustment can be tackled directly without the need for full causal graph discovery. In the spirit of Zheng et al. (2018a) and Mooij et al. (2009) we exploited how we can formulate the problem directly as a continuous optimization problem without the need for combinatorial search or indirectly optimizing a likelihood function. To do so, we derived (in)dependence criteria conditional on functions of covariates. These criteria are sufficient for backdoor adjustment. By learning these functions to minimize dependence scores on the observed variables, we have a differentiable way to learn a backdoor adjustment that can control for unobserved confounders. We showed how our method consistently matched or outperformed baselines that make similarly weak assumptions. There are many exciting directions for future work, including formulating semi-parametric versions of the method and considering problems where W and X may be sets of instruments/treatments. In this case, stronger signals may be obtained, making the problem more realistic in practice if the goal is to properly control for favorable outcomes of Y.

4 | Operationalizing Complex Causes: A Pragmatic View of Mediation

4.1 Introduction

Understanding causal mechanisms is a primary goal of scientific inquiry and a crucial prerequisite for planning effective interventions. However, the task of isolating and quantifying treatment effects is complicated by several obstacles. Fundamental questions of identifiability (Shpitser and Pearl, 2008; Correa and Bareinboim, 2020) and transportability (Bareinboim and Pearl, 2016) pose significant challenges to practitioners across a variety of disciplines. The problems are particularly acute in high-dimensional settings where interventions are rarely of the surgical or "atomic" sort envisioned by most authors in this area. For instance, genomic data contains rich information about the pharmacodynamic impact of drug therapies on disease activity. However, careful analysis is required to detect and operationalize these sparse signals, as causal effects are not defined in terms of direct interventions on, say, individual genes, but are instead propagated from a crude treatment (drug administration) on a complex object (the human transcriptome), which affects outcomes (disease activity) through several mediating pathways. Similar complexity arises in other fields, for instance when purported causes are social constructs like "gross domestic product" (Arnold et al., 2020) or large-scale natural phenomena like "El Niño" (Chalupka et al., 2016a).

Despite a substantial and growing literature on causal inference (see Sect. 4.2), existing theory largely fails to accommodate complex systems where the putative causes X of an outcome of interest Y have many internal components $(X_1, X_2, \ldots, X_p)$ not amenable to perfect control. Using the notation of Pearl (2000), there is no clear, non-trivial, physical method for enacting $do(x)$, i.e. setting variable(s) X to a particular value(s) x. Such cases are common in the natural and social sciences, to say nothing of text data and spatial processes captured at a coarse resolution. To continue with the medical example, researchers often design a therapy to target one or several hub genes in full awareness that this may spur unintended interactions with other biological processes. In such a study, researchers want to learn not just whether the

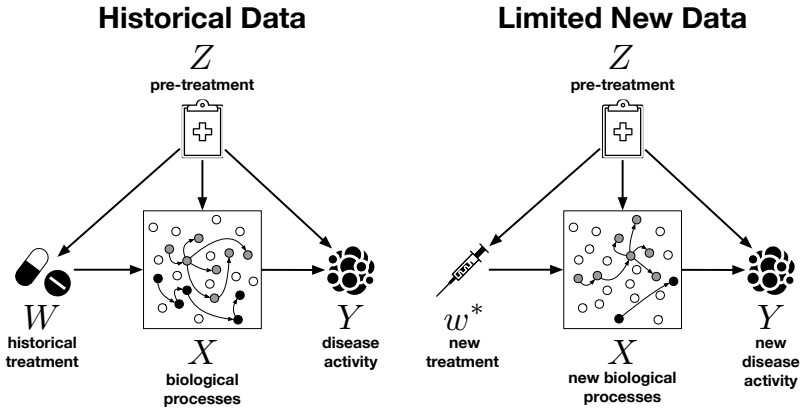


Figure 4.1: The complex cause problem setting. See text for details.

drug is effective but how variability in patient response can be explained by elements of X (perhaps combined with effect modifiers of pre-treatment variables Z).

This gives operational meaning to the notion of X as a *cause* of Y: even if $do(x)$ is undefined, we are interested in framing the effect on Y by some treatment W where the design of W comes from postulated or conjectured mechanisms triggered by X. Under the assumption that X fully *mediates* the actions W (in a technical sense defined in Sect. 4.2), an invariant relationship between X and Y under W becomes a useful building block for predicting the outcomes of new interventions. Furthermore, understanding *which* components of X simultaneously covary with W and Y is of independent interest, as this may suggest new targeted interventions that operate on those elements of X.

In this chapter, we discuss a notion of *pragmatic mediation* that provides a solution to the problem of prediction from new interventions. Our main contributions are threefold: (1) We formalize a general problem in which complex object X causes outcomes Y as a result of crude interventions W, with applications to domains with structured, high-dimensional data. (2) We describe an efficient method for estimating responses to new interventions, tractably marginalizing over the complex object X. (3) We propose a methodology for identifying practical causal mediation paths, which can provide insight into complex systems and suggest new hypotheses for future experiments.

4.2 Problem Setup

Let Y be an outcome of interest, and let X be postulated *causes* of Y, in the sense that hypothetical interventions on X would alter the distribution of Y (Woodward,

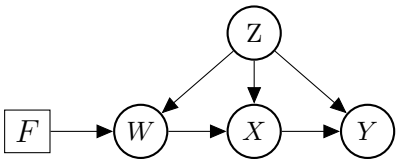


Figure 4.2: A DAG encoding independence assumptions in the set $\{F, W, X, Y, Z\}$, where random variables are circular vertices and intervention variable F is a square vertex. This diagram captures conditional independencies assumed in our setup, but is not Pearlian, as $do(x)$ is undefined. We cover all members of the Markov equivalence class of this graph, including those with unmeasured confounding between X and Z. We do not consider other unobserved confounders, though we discuss this in Sect. 4.6 as a direction for future work.

2003). In the machine learning and artificial intelligence literature, this is typically operationalized in terms of the interventional distribution $p(y \mid do(x))$ (Pearl, 2000). In many domains, however, perfect control is ambiguous or unattainable (VanderWeele and Hernán, 2013). This is often the case when X is a composition of more fundamental variables, as in the examples discussed in Sect. 4.1, as well as in image and text data. We will explore the latter two in our experiments in Sect. 4.4.

Actionable variables and their use. In our setup, actions that change the distribution of X are assumed to exist. We index them by W, allowing this to be a random vector. Pre-treatment variables Z, which are realized before $\{W, X, Y\}$, are also allowed. We say that W are our *actionable* variables, in the sense that in principle we can set them to exact values by an intervention.

By way of contrast, in the instrumental variable scenario, the target is $p(y \mid do(x))$, with W acting as an instrument that is not fundamental to the estimand of interest. If $do(x)$ is not defined and we are primarily driven by policy questions (e.g., choosing an optimal value for W), then W arguably makes the notion of X as a cause redundant. For example, Gelman (2009) suggests interpreting X as little more than a qualifier for our actions W. This is not satisfactory for the many applications where W was chosen because we expect that changing X will also change Y, even if this notion of propagation is unclear when $do(x)$ is undefined. In particular, there are practical scenarios where assumptions about invariances involving X aid the learning and prediction of policy outcomes. That is our motivation here.

Structural assumptions. Our goal is to predict Y under intervention levels $w^\star$ of W that we have not yet seen. Following Dawid (2021), we introduce a *regime indicator* F, which is not a random variable but instead indexes the conditional distributions

$p_F(w \mid z)$, with values ranging over possible interventions on W. Following Pearl 2000, we use the symbol "$do(w)$" $\in \mathcal{F}$ to mean the distribution in which $W = w$ given any $Z = z$. Our goal is then to predict Y given Z under $F = do(w)$, in particular reporting $\mathbb{E}_{F=do(w^\star)}[Y \mid z]$ for some new $w^\star$. In what follows, we will use the Pearlian notation $\mathbb{E}[Y \mid do(w), z]$ to represent the interventional distribution. Note that regime indicators can also accommodate non-atomic interventions – e.g., idle or stochastic regimes (Correa and Bareinboim, 2020) – although we will not make use of this in the sequel.

As is well-understood in the causal modeling literature, assumptions of independence between interventions and random variables can be represented by a directed acyclic graph (DAG). Our structural assumptions are encoded in Fig. 4.2. This is *not* a causal graph in the sense of Pearl (2000), as we are not making any claims about, say, lack of hidden common causes between Z and X or the applicability of do interventions on all variables. Instead, the DAG represents conditional independence claims such as $F \perp\!\!\!\perp Y \mid \{W, Z\}$, a statement interpretable as the lack of unmeasured confounding between W and Y given Z.[1]

Of particular importance to what follows is the implication $\{F, W\} \perp\!\!\!\perp Y \mid \{X, Z\}$. This conditional independence relationship, visually apparent from the d-separation in Fig. 4.2, informs us that there is no direct effect of W on Y. This invariance is the key point that pools together data collected at different values of F.

Problem statement: learning with pragmatic mediation. *Let* $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$. *Moreover, let* $\mathcal{D}_l$ *denote* m' *"labeled" datasets*

$$\mathcal{D}_l \equiv \{(W, X, Y, Z)^{l_1}, \ldots, (W, X, Y, Z)^{l_{m'}}\},$$

where $\mathcal{L} \equiv (l_1, \ldots, l_{m'}) \subset [m]$, *and let* $\mathcal{D}_u$ *denote* m'' *"unlabeled" datasets*

$$\mathcal{D}_u \equiv \{(W, X, Z)^{u_1}, \ldots, (W, X, Z)^{u_{m''}}\},$$

where $\mathcal{U} \equiv (u_1, \ldots, u_{m''}) \subset [m]$. *In particular,* $(W, X, Y, Z)^{l_i}$, $l_i \in \mathcal{L}$, *denotes data collected under regime* f_{l_i}, *the analogous holding for* $u_i \in \mathcal{U}$. *Given* $(\mathcal{D}_l, \mathcal{D}_u)$, *the goal is to return an estimate of*

$$f(w^\star, z) \equiv \mathbb{E}[Y \mid do(w^\star), z]. \tag{4.1}$$

[1] A similar device is used, for instance, in the proof of the back-door adjustment, Thm. 3.3.2 of Pearl (2000); and earlier graphical notions of unconfoundedness, e.g. Fig. 3.19 of Spirtes et al. (1993).

The problem concerns evaluating outcomes under an intervention that sets W to a *specific value* $w^\star$. The *post-treatment* variables X cannot be used as a basis for decision-making, but the invariances encoded by the lack of edges $\{F, W\} \to Y$ and $F \to X$ allow for predictions of policy outcomes even in the case where pairs $(w^\star, y)$ are not in our data. As is typical of causal inference problems, we require some assumptions regarding the support of treatment values in the given data. In particular, we have the following:

Assumptions (identification and support). *For all z in the support of $p(z)$: (i) the distribution $p(x \mid w^\star, z)$ is identifiable from the distributions sampled by $\mathcal{L} \cup \mathcal{U}$; (ii) the support of $p(x \mid w^\star, z)$ is contained in the union of the support of X in each dataset in $\mathcal{L}$.*

Thus, in order to obtain $\mathbb{E}[Y \mid do(w^\star), z]$ from

$$\int \mathbb{E}[Y \mid x, z] p(x \mid w^\star, z)\, dx,$$

we must have some means of generalizing to $p(x \mid w^\star, z)$ from past data, including unlabeled data. Condition (ii) says we can learn the $\mathbb{E}[Y \mid x, z]$ factor across the support of $p(x \mid w^\star, z)$ using the datasets contained in $\mathcal{L}$. This assumption can be relaxed, provided we have some principled way to extrapolate beyond the regions of (X, Z) covered by $\mathcal{L}$.

In our experiments (see Sect. 4.4), we predict causal responses for new interventions with no observed outcomes but some (unlabeled) data on X. Such cases arise when, for instance, Y takes a long time to be observed, or past interventions $w^\star$ took place targeting a different outcome variable. We will not constrain the functional relationship between W and X in any way, meaning that $p(x \mid w^i, z)$ need not contain any information about $p(x \mid w^j, z)$ for $i \neq j$.

This machinery operationalizes what we mean by X being a cause of Y, even if we do not define $do(x)$. At the heart of causal inference is the notion of invariance under intervention, which can be exploited even if the putative causes of interest cannot be directly manipulated. We call X a *pragmatic mediator*, i.e. a set of variables that allows us to decompose a causal model for W in Y by a model (given Z) relating W and X only, and X and Y only, under a space of interventions $\mathcal{F}$. This bears little relation to counterfactual mediation (VanderWeele, 2015) and demonstrates how restricted notions of mediation can be more useful than counterfactual ones in some contexts.

Related work. Although invariance principles have long been cited in formal definitions of causality (Spirtes et al., 1993; Pearl, 2000; Bühlmann, 2020), they have recently found a new life in machine learning approaches that target more focused questions of practical interest.

The Invariant Causal Prediction method of Peters et al. (2016b) – later extended by Heinze-Deml et al. (2018) and Gamella and Heinze-Deml (2020) – exploits the assumption that F does not directly cause outcome Y except for (unknown) causal parents from a candidate pool of observable variables X. There the objective is to discover the causal parents as opposed to learning what happens when we marginalize them. Likewise, Invariant Risk Minimization (Arjovsky et al., 2020) exploits variability in F to better learn the relationship between X and Y in a way that is robust to estimation errors due to spurious, unstable associations. Again, the main focus is on the use of X, here in a (non-causal) prediction problem.

Post-treatment variables are used to improve bandit optimization by factoring the arm space, where variables that are themselves targets of interventions exhibit some (at least partially known) structure (Lattimore et al., 2016; Lee and Bareinboim, 2018b; de Kroon et al., 2020). In contrast, we focus on the class of problems where there is little to be gained by exploiting the inner structure of X, as in many applied scenarios they are composite variables with ambiguous fine causal structure (Arnold et al., 2020).

Domain adaptation, in particular covariate shift, has a long tradition of being analyzed in the context of causal models (e.g., Zhang et al. (2013)). The emphasis of this literature is how to better cope with changes in distribution between, say, training and test regimes. Although techniques such as sample reweighting for improved statistical efficiency are relevant when dealing with multiple regimes, this will not be our focus here. Instead, in Sect. 4.3, we emphasize convenient parametrizations of our causal set up so as to facilitate marginalization over X and parameter learning.

This work is particularly influenced by the literature on ambiguous or undefined interventions (Spirtes and Scheines, 2004; VanderWeele and Hernán, 2013; Lee and Bareinboim, 2019b) as well as causal abstractions and compositional data (Chalupka et al., 2017; Beckers and Halpern, 2019; Arnold et al., 2020). The idea of pragmatic mediators is essentially the grounding of a causal abstraction through imperfect interventions W and the delimitation of its possible values as allowed by a "causal dictionary" $\mathcal{F}$. The ambiguity of X requires explicit assumptions about the space of modifications that we expect to enact on X.

Finally, since completing this work we have learned of a parallel body of work considering the estimation of long-term effect of treatments using short-term outcomes, termed "surrogate index". It was introduced by Athey et al., and further explored in Kallus and Mao 2020 and Battocchi et al. 2021. The aim of this family of methods is to use available short-term outcomes of novel treatment programs in combination with historical long-term outcomes of historical intervention regimes, and in that share some of the implications of our method, as well as a similar problem formulation. However, they do not explore the concept of complex causes, pragmatic mediation and consequences to interpretability or novel intervention design we explore in this work.

4.3 Method

Based on the assumptions introduced above, this section describes (a) an algorithm for learning expected outcomes Y under crude interventions on X, operationalized as $F = do(w)$, conditioned on pre-treatment covariates Z; and (b) a procedure for interpreting the elements of X that play a mediating role in the fitted model.

Causal response estimation. We assume access to a set of features $\phi_i : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$. These features are *candidate mediators*, moderated by covariates Z, which describe the outcome model for Y as

$$Y = \theta_0 + \sum_{i=1}^{d} \theta_i \phi_i(X, Z) + \varepsilon, \tag{4.2}$$

where ε is an independent error term with $\mathbb{E}[\varepsilon] = 0$.

Candidates may come from domain experts (e.g., experimentally validated regulatory pathways) or a data-driven approach (e.g., latent factors learned by an autoencoder). They represent a macro-level summary that clarifies what the existing $\mathcal{F}$ is able to modify in X that simultaneously contributes to Y. For example, if X describes a spatially-distributed object, like neural activations or environmental sensors, features ϕ_i can correspond to smoothing windows with localized information. If X is a text document, ϕ_i may represent aggregate interpretable interactions of relevant entities, topics, and other parts of speech. The linear assumption is substantive but not especially restrictive, given a sufficiently flexible library of basis functions Φ, which, as mentioned above, can be trained directly via neural networks or some other representation learning method.[2]

[2]Note that, though each ϕ_i is a function of X and Z, we occasionally simplify notation by suppressing the dependence, writing ϕ_i for $\phi_i(X, Z)$ and $\Phi = \{\phi_i\}_{i=1}^{d}$.

Given Eq. 4.2, it follows by the assumptions encoded in Fig. 4.2 and by linearity of expectation that

$$\mathbb{E}[Y \mid do(w), z] = \theta_0 + \sum_{i=1}^{d} \theta_i \mathbb{E}[\phi_i(X, Z) \mid w, z]. \tag{4.3}$$

We therefore propose a two-stage procedure to estimate $\mathbb{E}[Y \mid do(w), z]$:

1. Learn $g_i(w, z) \equiv \mathbb{E}[\phi_i(X, Z) \mid w, z]$ for all i via any black-box regression algorithm, and let $\hat{\mathbf{g}}$ denote the d-dimensional vector of resulting expectations.

2. Learn $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}[(Y - \boldsymbol{\theta}^\top \hat{\mathbf{g}})^2]$ via regularized regression (e.g. Lasso, Tibshirani, 1996), to provide sparsity on $\boldsymbol{\theta}$ where supported by the data.

The procedure is detailed in Alg. 1, where we consider the case in which labeled datasets are pooled together into a set with n samples, and we learn a model for $p(x \mid w^\star, z)$ from unlabeled conditions with a single treatment level $w^\star$. This exploits the known structural relationship between W, X, Φ, and Y. In particular, it represents the marginalization of X directly in terms of $\mathbb{E}[\phi_i(X, Z) \mid w, z]$,[3] which avoids the density estimation problem of learning $p(x \mid w, z)$.

There is a relation between this idea and methods for estimating non-linear causal effects in additive-error instrumental variable models based on (potentially infinite) basis expansions (Singh et al., 2019; Muandet et al., 2020). However, given the potential high dimensionality of X and the desire for interpretability, we favor dictionaries that are either hand-constructed or the result of adaptive algorithms. Moreover, although we have the option of fitting θ by regressing Y directly on Φ, we still favor the regression on $\hat{\mathbf{g}}$ instead, as $\phi_i(x, z)$ is a random variable not observable at test time.

Explainable pragmatic mediation. Under the assumptions of our setup, we would like to provide practitioners with qualitative information on the estimated role of the candidate mediators. Informally, we say that $\phi_i(X, Z)$ is a *causal pragmatic mediator* if and only if it covaries with W and Y simultaneously, with adjustments for Z and the other candidate mediators depending on the scenario. More formally, causal pragmatic mediators satisfy two criteria:

(i) $\phi_i(X, Z) \not\perp\!\!\!\perp W \mid Z$,

[3]This can be even further simplified if we opt for product features of the shape $\phi_i(X, Z) \equiv \phi_{ix}(X)\phi_{iz}(Z)$, as in this case, we have $\mathbb{E}[\phi_{ix}(X)\phi_{iz}(Z) \mid w, z] = \mathbb{E}[\phi_{ix}(X) \mid w, z]\phi_{iz}(z)$ (Kaddour et al., 2021).

Algorithm 1 Causal Response Prediction

Require: Historic interventions $\{w_i, \Phi(x_i, z_i), z_i, y_i\}_{i=1}^{n}$, new intervention training set $\{w^\star, \Phi(x_j, z_j), z_j\}_{j=1}^{n^\star_{train}}$, new intervention test set $\{w^{\star\prime}, z'_j\}_{j=1}^{n^\star_{test}}$
Historic Interventions
Learn $g_k(w, z) = \mathbb{E}[\phi_k(X, Z) \mid w, z]$ Stage 1, via any black-box model
Learn $f(w, z) = \mathbb{E}[Y \mid \mathbf{g}(w, z)]$ Stage 2, via an L_1-penalized model
New Intervention
Learn on train split
$g^\star_k(w^\star, z) = \mathbb{E}[\phi_k(X, Z) \mid w^\star, z]$ Stage 1, update for new intervention $w^\star$
Predict on test split
$\hat{y} = f(\mathbf{g}^\star(w^{\star\prime}, z')) = \mathbb{E}[Y \mid do(w^{\star\prime}), z']$ Stage 2, predict using pre-learned f
return $\hat{y}$

Algorithm 2 Pragmatic Mediation Selection

Require: Weights θ, training set $\{w_i, \Phi(x_i, z_i), z_i\}_{i=1}^{n}$, test set $\{w'_i, \Phi'(x'_i, z'_i), z'_i\}_{i=1}^{n'}$, one-sided paired difference test $c(\cdot)$, level α, mediators $\mathcal{M} = \{\}$
for $\phi_i \in \Phi$ **do**
  if $\theta_i \neq 0$ **then**
    Learn $g_i^0(z) = \mathbb{E}[\phi_i(X, Z) \mid z]$ on train split
    Learn $g_i^1(z, w) = \mathbb{E}[\phi_i(X, Z) \mid z, w]$ on train split
    Obtain residual $\varepsilon_i^0 = \phi'_i - g_i^0(z')$ on test split
    Obtain residual $\varepsilon_i^1 = \phi'_i - g_i^1(z', w')$ on test split
    $p = c(|\varepsilon_i^0|, |\varepsilon_i^1|)$
    if $p \leq \alpha$ **then**
      Add mediator $\mathcal{M} = \mathcal{M} \cup \{\phi_i\}$
    end if
  end if
end for
return $\mathcal{M}$

(ii) $\phi_i(X, Z) \not\perp\!\!\!\perp Y \mid \{\Phi_{\setminus i}, Z\}$

where $\Phi_{\setminus i} \equiv \Phi \setminus \phi_i(X, Z)$. We will henceforth refer to (i) and (ii) as $\mathcal{M}$-criteria (for mediation criteria).

Another way of interpreting this is by saying that W has a "nonzero conditional total effect" on ϕ_i for some $Z = z$ (that is, a conditional association without adjusting for $\Phi_{\setminus i}$), while ϕ_i has a "direct effect" on Y (conditional association, also conditioning on $\Phi_{\setminus i}$).

This definition is entirely agnostic to any possible causal structure among the elements of Φ, a structure which is itself indeterminate since $do(x)$ is not defined. Notice that the idea of combining a "total" effect "into" Φ with a "direct" effect "out of" Φ relates to settings where we may want to design new elements of $\mathcal{F}$ that "short-circuit"

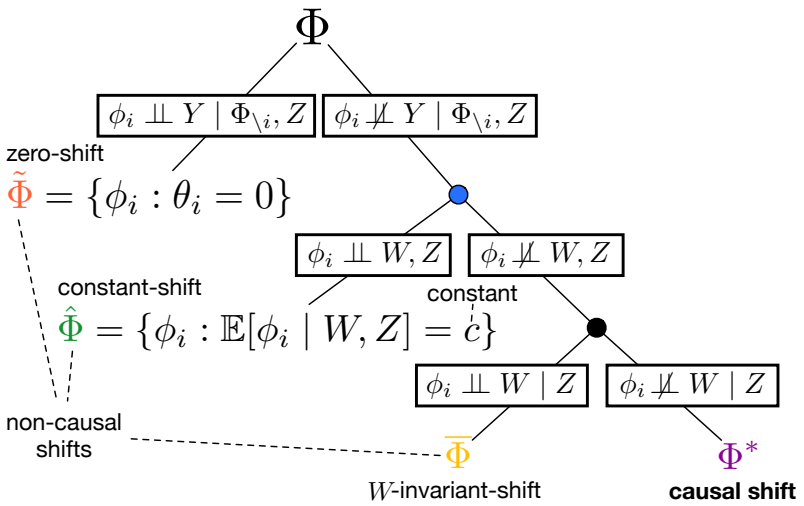


Figure 4.3: Recursive partition of Φ by how elements do or do not shift conditional average treatment effects. See text for details.

the mechanism, by directly targeting ϕ_i if this is at all possible and desirable in a particular domain.[4]

Although the distinction is not crucial for prediction, causal mediators can provide valuable insights about what in X characterizes the effect of W on Y. For instance, if only a subset of regions of the brain respond to stimuli and predict some behavior, then novel interventions can be designed targeting just those regions with detectable causal impact.

The leaf nodes of the tree depicted in Fig. 4.3 correspond to candidate mediators with different functional roles.

1. $\tilde{\Phi} \equiv \{\phi_i : \phi_i \perp\!\!\!\perp Y \mid \Phi_{\setminus i}, Z\}$. These candidates will receive zero weight in the linear formula described by Eq. 4.2 (and, hence, also Eq. 4.3). That is, for each $\phi_i \in \tilde{\Phi}, \theta_i = 0$.

2. $\hat{\Phi} \equiv \{\phi_i : \phi_i \notin \tilde{\Phi} \wedge \phi_i \perp\!\!\!\perp W, Z\}$. In this subset, $\mathbb{E}[\phi_i \mid w, z]$ is constant for all w and z. These terms will be absorbed into the intercept of the linear formula described by Eq. 4.3. That is, $\mathbb{E}[Y \mid do(w), z] = \theta_0 + \sum_{\phi_i \in \hat{\Phi}} \mathbb{E}[\phi_i]$ + "function of w and z".

3. $\overline{\Phi} \equiv \{\phi_i : \phi_i \notin \{\tilde{\Phi} \cup \hat{\Phi}\} \wedge \phi_i \perp\!\!\!\perp W \mid Z\}$. These candidates will receive nonzero weight in Eq. 4.2, but only through the $Z \to \phi_i \to Y$ path. They are invariant

[4]For instance, ignoring Z for simplicity: if there exists a "Pearlian" causal chain $X_1 \to X_2 \to X_3 \to Y$ in the system with no further edges, and we have a rich dictionary Φ, (features of) neither X_1 nor X_2 alone would qualify as causal pragmatic mediators, while (features of) X_3 would, even if all interventions in $\mathcal{F}$ can only directly modify X_1 and X_2.

Table 4.1: General description of experimental setups.

	ImagePert	**Humicroedit**	**DREAM5**
Z	pre-perturbation image	original news headline (GloVe avg. vector)	baseline gene expression
W	location of normal distribution for perturbation	new entity edit (GloVe vector)	transcription factor out-degree
X	post-perturbation image	edited headline (GloVe avg. vector)	post-intervention gene expression
Φ	convolution windows over X	funniness hypotheses (Hossain et al., 2019)	change in kernel eigengene
Y	intensity of pixels, linear combination of Φ	funniness score, via linear combination of Φ	linear combination of Φ

in W and therefore, just like $\tilde{\Phi}$ and $\hat{\Phi}$, do not contribute to conditional average treatment effects $\mathbb{E}[Y \mid do(w), z] - \mathbb{E}[Y \mid do(w'), z]$.

4. $\Phi^* \equiv \{\phi_i : \phi_i \in \Phi \backslash \{\tilde{\Phi} \cup \hat{\Phi} \cup \overline{\Phi}\}\}$. Only this latter subclass satisfies the $\mathcal{M}$-criteria, picking out causal mediators ϕ_i on the $W \to \phi_i \to Y$ path.

This recursive partitioning of Φ immediately suggests a practical method for pragmatic mediation discovery. First, we perform our two-step estimation procedure. Then, for each ϕ_i such that $\theta_i \neq 0$, perform a conditional independence test against the null hypothesis $H_0 : \phi_i \perp\!\!\!\perp W \mid Z$.[5] See Alg. 2 for details.

There exists no uniformly valid conditional independence test for continuous conditioning variables (Shah and Peters, 2020). However, numerous nonparametric methods have been developed with good performance on real and synthetic datasets (Heinze-Deml et al., 2018). In our experiments, we use a simple nested regression procedure, in which we compare the absolute value of out-of-sample residuals for null and alternative models – i.e., $g_i^0(z) = \mathbb{E}[\phi_i \mid z]$ and $g_i^1(z, w) = \mathbb{E}[\phi_i \mid z, w]$, respectively – using a one-sided Wilcoxon rank-sum test.[6] If predictive accuracy significantly improves with the inclusion of W, then we reject H_0. Estimation and testing are performed on separate samples to ensure unbiased inference. The procedure can easily be modified to adjust for multiple testing.

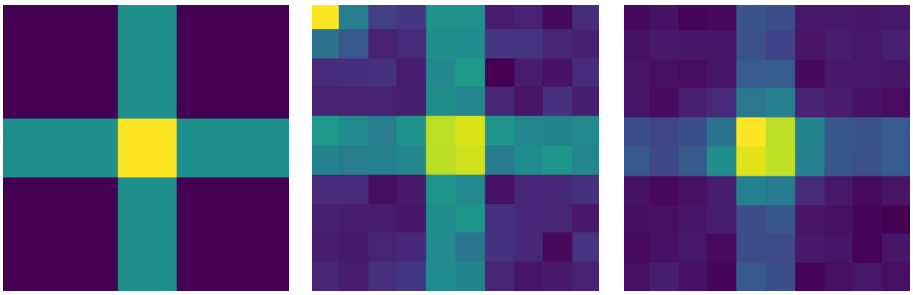
4.4 Experiments

In this section, we demonstrate our method in a variety of domains. We start with a simulated visual simulation task to provide a more concrete intuition for our approach. We then introduce a text data example where users are asked to edit news headlines to make them more humorous. Finally, we describe a genomics experiment where

[5] In randomized trials, where $Z \perp\!\!\!\perp W$ by design, this can be replaced by a marginal association test against $H_0 : \phi_i \perp\!\!\!\perp W$, for those ϕ_i which are non-trivial functions of X.

[6] Other tests could in principle be substituted here, e.g. the binomial test or z-test, depending on what assumptions one is willing to make about residual distributions. See (Lei et al., 2018, Sect. 6).

we simulate the effects of gene knockouts on the *E. coli* transcriptome. The code to reproduce results can be found at `https://github.com/limorigu/ComplexCauses`.



(a) (b) (c)

Figure 4.4: Visual example for image perturbation dataset. (a) Example image (z). (b) Same image, post-perturbation (x, in response to w). (c) Same image under a new perturbation regime, which we leave for the test set (x', in response to w').

4.4.1 Setup

We have two primary goals: (a) causal response prediction, and (b) identification of causal pragmatic mediators. We describe the overall setup for all domains below.

Prediction. We assume access to $m-1$ mutually independent historic training regimes with corresponding labeled datasets $\mathcal{D}_{l_1}, \ldots, \mathcal{D}_{l_{m-1}}$, where each $\mathcal{D}_{l_k} = \{(W_i, X_i, Y_i, Z_i)^{l_k}\}_{i=1}^{|\mathcal{D}_{l_k}|}$. Our goal is to learn $\mathbb{E}[Y \mid do(w^\star), z]$ for a new regime $F = do(w^\star)$ (e.g., a prospective intervention). In this regime, we are given access to limited labeled training data $\mathcal{D}_{l_{w^\star}} = \{(W_i^\star, X_i, Y_i, Z_i)^{l_{w^\star}}\}_{i=1}^{|\mathcal{D}_{l_{w^\star}}|}$ and more unlabeled training data $\mathcal{D}_{u_{w^\star}} = \{(W_i^\star, X_i, Z_i)^{u_{w^\star}}\}_{i=1}^{|\mathcal{D}_{u_{w^\star}}|}$, where $|\mathcal{D}_{u_{w^\star}}| \gg |\mathcal{D}_{l_{w^\star}}|$. This captures settings where measurements for Y are expensive, delayed, or simply unrecorded. All methods are evaluated on a test dataset $\mathcal{T}_{w^\star} = \{(W_i^\star, Y_i, Z_i)^{t_{w^\star}}\}_{i=1}^{|\mathcal{T}_{w^\star}|}$.

Baseline methods that estimate $\mathbb{E}[Y \mid do(w^\star), z]$ can only make use of the labeled dataset $\mathcal{D}_{l_{w^\star}}$, as all regimes are mutually independent. However, by exploiting structural information $\Phi(X, Z)$, we are able to leverage the invariant $p(y \mid x, z)$ distribution from prior regimes. That is, we estimate $\mathbf{g}$ and $\boldsymbol{\theta}$ from $\mathcal{D}_{l_1}, \ldots, \mathcal{D}_{l_{m-1}}$ and predict effects in new regimes using only Z and W, so our method effectively treats $\mathcal{D}_{l_{w^\star}} \cup \mathcal{D}_{u_{w^\star}}$ as a single test set.

We will compare our approach to multiple regression baselines that estimate $\mathbb{E}[Y \mid do(w^\star), z]$ as the proportion of labeled data for the new regime $\mathcal{L}_{w^\star}$ grows from 10%-100%. Specifically, we consider models from four function classes: lasso regression (linear), support vector regression (SVR), random forest (RF), and gradient boosting

(GB). Default hyperparameters are used throughout; see Appendix for details. We also note that other methods that seem to share similarities with our goal, such as co-training (Blum and Mitchell, 1998) and domain adaptation (Chen et al., 2011), would not be relevant baselines for comparison as they differ significantly from our work in two ways. (1) There is a two-stage functional decoupling arising from Eq. 4.3 alongside variable decoupling that we aim to exploit by learning g and then f; co-training does not involve such functional decoupling. (2) We can only leverage the first stage of the decoupling in a new domain. We are not aware of any domain adaptation method that accommodates this specific notion of adaptation.

As an additional check on our performance, we further consider 100 different settings of the target Y, by sampling 100 different parameters (i.e., weights $\boldsymbol{\theta}$) for its structural equations in all three tasks (for the image perturbation example we sample 1500 settings). By demonstrating consistent results across these trials, we illustrate that our method is robust to different configurations of the target variable.

Explanation. Our method is also able to find pragmatic mediators in the complex object X. By studying the high-level descriptions ϕ_i that (a) receive nonzero weight $\theta_i \neq 0$ in the sparse regression, and (b) reject $H_0 : \phi_i \perp\!\!\!\perp W|Z$ at some pre-specified level α, we can identify causal mediators of relevance. We report mediator discovery error rates for all experiments below. Significance levels for all tests were fixed at $\alpha = 0.01$, with p-values adjusted for multiple testing via Holm (1979)'s method.

4.4.2 Image Perturbation Simulation

Setup description. Our first example is simulated and visual, which we hope will provide some intuition for the structure of this problem. We start with five possible pixel patterns (Z) and perform interventions by adding bivariate normal noise with location W. These treatments are influenced with some probability by Z. The resulting post-perturbation image (X) is then summarized via four different convolution windows, $\Phi(X, Z) = \{\phi_1, \phi_2, \phi_3, \phi_4\}$, where each ϕ_i corresponds to a quadrant of the image, and the convolution weights are indexed by the pattern corresponding to the original image Z. Finally, the intensity of the pixels over the whole image leads to an outcome (Y), given by a linear combination of Φ. The generative model used to produce the simulation is described in Fig. 4.5.

$\boldsymbol{p} = [0.2, 0.2, 0.2, 0.2, 0.2]$ defines the multinomial distribution from which we sample shape indicator t. Δ denotes a 5×4 matrix, where each row is a simplex containing different probabilities for selecting W values. $d_0 = 0$ and $d_n = 10$ define

$$
\begin{aligned}
t &\sim \text{Multinomial}(\boldsymbol{p}) \\
Z &= \text{pattern}_t \\
W &\sim \text{Multinomial}(\Delta_t) \\
f_w &= \begin{cases} \text{for } i = 0 \text{ to } 1000: \\ \quad \gamma \sim \mathcal{N}(W, \mathbf{I}) \\ \quad \text{if } (d_0, d_0) < \gamma < (d_n, d_n): \\ \quad\quad f_w[\gamma] = f_w[\gamma] + \eta \end{cases} \\
X &= Z + f_w + \mathcal{N}(0, 0.5) \\
\Phi &= \text{Convolution}_t(X) \\
Y &= \boldsymbol{\theta}^\top \Phi + \mathcal{N}(0, 0.1)
\end{aligned}
$$

Figure 4.5: Description of the generative model used in the experiment of Sect. 4.4.2.

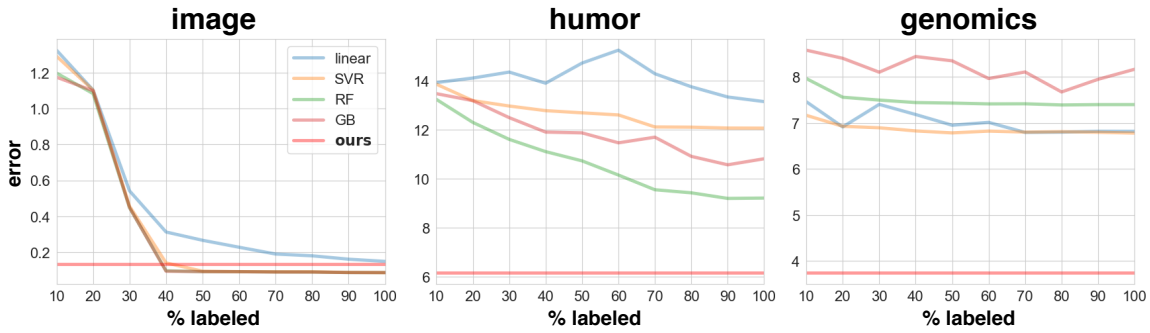


Figure 4.6: The mean squared error (MSE) between the estimated causal effect and the true causal effect as a function of the amount of labeled data that is available in the new regime $do(w^\star)$.

the dimensions of all images. The condition involving them and γ checks whether the sampled location falls within the image size. η is a perturbation parameter (fixed at 0.1 in our experiment) that is added to the sampled location if it passes the check above. This example is designed for demonstrative purposes, and ϕ_1 is constructed to be the pragmatic mediator we intend to find, as it both varies with W and has a nonzero coefficient ($\theta_1 = 0.7$) in the structural equation for Y. See full details in the Appendix. Fig. 4.4 shows an example set of sampled images.

Results. The results of all methods on a new intervention $w^\star$ are presented in Fig. 4.6 (left). Additionally, Fig. 4.7 shows the process of mediator explanation for our method. The true mediator in this simulation, ϕ_1, is indicated in black. Conditional independence tests identify three windows – $\Phi = \{\phi_1, \phi_2, \phi_3\}$, indicated in red – that vary with W after conditioning on Z. We fit a lasso regression to estimate causal effects (see Eq. 4.3), selecting windows $\Phi = \{\phi_1, \phi_4\}$, indicated in blue. Finally,

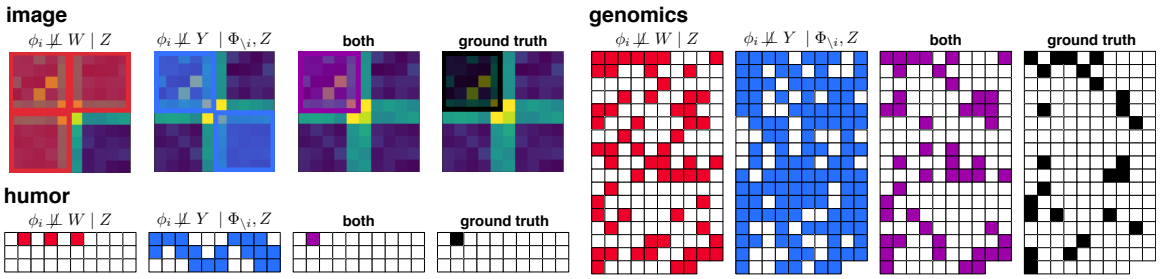


Figure 4.7: The ϕ_i selected by the mediation discovery method. Each set is identified as follows: red by a conditional independence test, blue by sparse regression, and purple those satisfying both tests (**black** are the true mediators). Note for the high-dimensional genomics dataset the ϕ_i are identified by only testing those ϕ_i selected by sparse regression to increase testing power.

the intersection of these two sets, $\Phi^* = \{\phi_1\}$, is our causal mediator, marked in purple. Fig. 4.8 presents performance over 1500 different samples of parameters in the structural equation of prediction target Y. It shows similar trends to the single Y setting, where our method dominates performance by baselines until 30-40% of labels are available. The mean squared error (MSE) in this simplified example is far smaller and required more samples to make std. error scale accordingly. We further note that for the single Y setting, we picked $\theta = \{0.7, 0, 0, -0.5\}$ to clearly demonstrate the idea of pragmatic mediation. For Fig. 4.8, we instead sampled $\boldsymbol{\theta}$ from a distribution (See Appendix for details), which seemed to help the performance of some baselines, while adversely affecting others.

4.4.3 Humorous Edits to News Headlines

Setup description. As a second example, we consider a dataset from a computational humor experiment. Participants were given news headlines and asked to make single entity changes such that the resulting headline would be humorous (Hossain et al., 2019). This work was further extended into a SemEval2020 task, and full datasets were made publicly available.[7]

For our evaluation, we combine all listed datasets and define the following: original headline (Z), new words introduced by edit (W), and a revised headline (X). Following the analysis of Hossain et al. (2019), we carried out the following pre-processing procedures: 1. We generated clusters of edit words (granular W) by performing a k-means clustering on GloVe vector representations (Pennington et al., 2014) of each edit word, with $k = 20$. The aim was to reduce the space of possible interventions

[7]See `https://www.cs.rochester.edu/u/nhossain/humicroedit.html`.

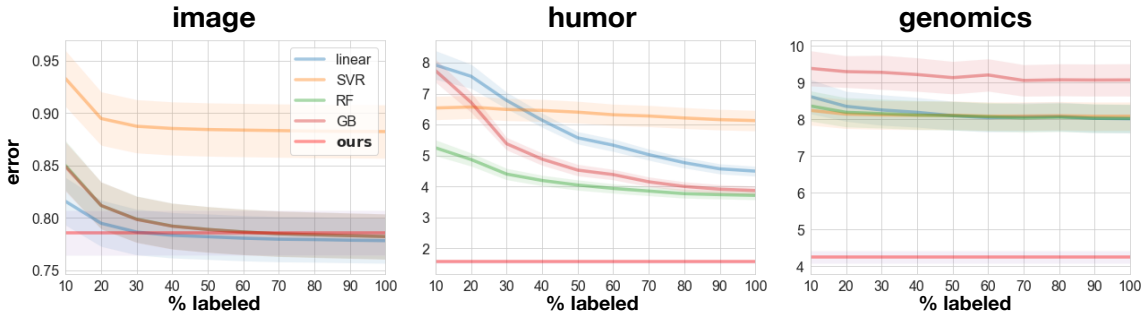


Figure 4.8: The mean squared error (MSE) between the estimated causal effect and the true causal effect as a function of the amount of labeled data that is available in the new regime $do(w^\star)$. Means and std. errors are over 1500 for the Image Perturbation experiments, and 100 for the other two, different configurations of $\boldsymbol{\theta}$, the parameters in the structural equations giving rise to Y.

to topics rather than individual words, for the purpose of defining data subsections as historic and new intervention splits. We used the resulting cluster label to create these splits. 2. We created 30 high-level descriptions ϕ for this setting (full description in the Appendix). One can think of Φ in this scenario as hypotheses to explain the funniness of an edited headline (X). 3. Computational humor is known to be a difficult domain for direct prediction tasks. For the illustrative purpose of this work, we generated funniness scores for the outcome variable Y as a linear combination of Φ with additive noise. A random third of the coefficients are assigned a value of 0, with the rest sampled from a uniform distribution $\mathcal{U}(-1, 1)$.

Results. The results of our estimation method of the outcome Y for a random new intervention $w^\star$ are presented in Fig. 4.6 (middle). As can be seen, we achieve an MSE of 5.33, well below alternative estimation methods of $\mathbb{E}[Y \mid do(w^\star), z]$. Furthermore, our method correctly identified the mediator ϕ_2 in this setting, see Fig. 4.7. Fig. 4.8 provides another angle on the quality of predictions with our method by examining results over 100 trials with different configurations of $\boldsymbol{\theta}$. We can clearly see that our method still outperforms the baseline alternatives, and sees little variation in performance across parameter values, as can be seen by the small std. error bars.

4.4.4 Gene Knockouts

Setup description. As a final experiment, we consider semi-simulated gene knockouts based on data from the Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenge (Marbach et al., 2012). The *E. coli* transcriptome published as part of the DREAM5 challenge comprises a 805×4511 gene expression matrix, with

334 candidate transcription factors.[8] We use GENIE3 (Huynh-Thu et al., 2010), a leading gene regulatory network inference algorithm based on random forests, to fit the 4177 structural equations that govern this system. We treat the resulting model as our ground truth SCM.

We simulate $n = 10^4$ samples of baseline expression data for the transcription factors from a multivariate Gaussian distribution with parameters estimated via maximum likelihood. These values are then propagated by GENIE3 to downstream variables, resulting in a complete set of baseline expression data (Z). We simulate 10 gene knockout experiments, summarized by the out-degree of the corresponding transcription factor (W). Post-intervention expression is once again simulated by GENIE3 (X). We treat each subnetwork of at least 10 genes as a pathway and summarize its expression by taking the first kernel principal component (Schölkopf et al., 1999) of the corresponding submatrix, i.e. the kernel eigengene. The difference between post- and pre-intervention eigengene expression for all 168 modules that meet this dimensionality criterion constitutes our high-level summary (Φ). Modules are subsequently ranked by their Spearman correlation with W. The top and bottom 25 are assigned a nonzero weight in a linear simulation of outcomes Y, with standard normal noise.

Results. Results for a random new gene knockout are presented in Fig. 4.6 (right). The sparsity of this problem poses a particular challenge for baseline regression methods, which could potentially be mitigated with further tuning. In addition to achieving low MSE on the test set, our method additionally recovers 92% of all true mediators, with an overall accuracy rate of 85%. Most of the errors in this example appear to derive from false positives in the lasso regression, which could likely be improved with more cautious tuning of the λ parameter that controls model sparsity. As can be seen in Fig. 4.8, the same trends remain in place when repeating the experiment over 100 different configurations of $\boldsymbol{\theta}$.

4.5 Additional Experimental Details

4.5.1 Method, Evaluation Task and Baselines

We tested our estimation method with three different setups: an image-based simulation, an experimental text dataset, and a naturally-simulated experimental genomics dataset. We considered the same evaluation task for all setups:

[8] See `http://dreamchallenges.org/project/dream-5-network-inference-challenge/`.

1. Using a train set, with various "seen" W values, we train a model for *phi* prediction, and use $\mathbb{E}[\Phi \mid W, Z]$ to fit a lasso regression to predict Y, giving rise to $\mathbb{E}[Y \mid \hat{\Phi}]$ predictions.

2. Next, we see a train split from a test set, corresponding to an "unseen" intervention w', on which we relearn $\mathbb{E}[\Phi \mid w', Z]$

3. Finally, we test Y predictions on a test set of the unseen w' test set, using our relearned $\mathbb{E}[\Phi \mid w', Z]$ model, and our previously trained $\mathbb{E}[Y \mid \hat{\Phi}]$ Lasso regression models. Thus, for a test split of the test set, we predict for Y for previously unseen w', z pairs, i.e. $\mathbb{E}[Y \mid do(w'), Z']$.

For further clarity, we provide a visual description of datasets used in different stages of the method above in Figure 4.9. We compared our estimation accuracy, via Mean Squared Error, as Y labels become available in the unseen w' regime (10-100% of labels). We record our loss against four baselines estimating the same quantity: 1. linear model, as a Lasso regression with cross validation to pick the coefficient λ on the regularization term in the range [0.05,1], 2. SVM predictor with default parameters from (Pedregosa et al., 2011), 3. Random Forest regression with default parameters from (Pedregosa et al., 2011), aside for specifying 5 minimal samples in split, and 4. Gradient Boosting with default parameters from (Pedregosa et al., 2011), aside for maximum depth which was set to 5[9]. We provide code to reproduce our results alongside this document. We also tested a Multi Layered Perceptron baseline, but found it to perform much worse than alternatives without dedicated tuning, and subsequently did not include it in the results.

For the prediction task we include two settings: one for a single configuration of the target Y, and one where we present average performance over 100 different settings of the parameters θ in the structural equation giving rise to Y. Since our test set on which we report result is rather small, for the first setting we average results of the baselines over 10 different shuffles of the dataset on which we train and report result. We do this for the baselines and not for our own method as the baselines have access to 10 different proportions of the target Y, which involved small number of samples and largely varying performance based on the ordering of the dataset. Our method makes no use of labels, and thus is not vulnerable to this variability in performance. However, when we conduct the experiment over 100 different Y configurations, the existence of ample different examples of data better accounts for this variability in

[9]This could be potentially improved with further hyperparameter tuning.

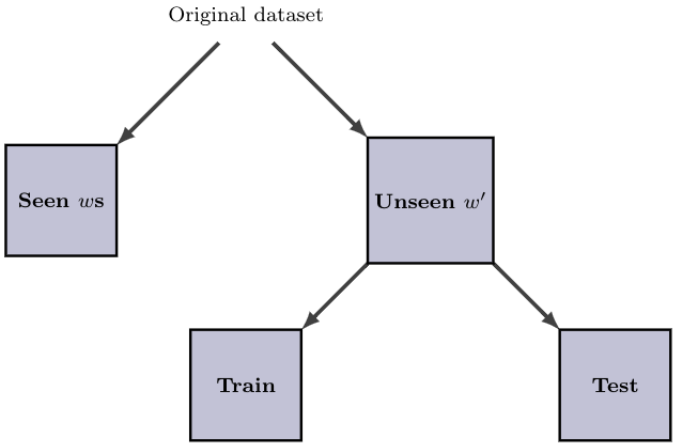


Figure 4.9: Description of dataset splits used in different parts of our experiment, corresponding to items 1-3 in 4.5.1.

performance based on sample ordering. Thus, we simply report results for a single data ordering, and report the mean and standard error over 100 different Y parameters settings instead.

4.5.2 Dataset Construction and Models' Training

4.5.2.1 Image Perturbation

The image perturbation is a simulation dataset that was put together to demonstrate the key ideas of the work. We create a dataset of 10,000 examples, images of size 10X10, equally made up of 5 pixel patterns: cross, square, crossing diagonal, pyramid and diamond. Patterns were allocated for each index by sampling one of 5 pattern indicators from a multinomial with equal probability for each shape. Next, the pattern indicator Z also served to select one of 5 set of probabilities that were used to seed a multinomial from which a perturbation pattern W was picked.

The perturbation was put together via a procedure that used the indicator W as a location to be used in a Multivariate Normal distribution with a two dimensional identity covariance matrix, $I(2)$. For each example in the dataset, we sampled 1000 tuples (x,y) from the Multivariate Normal described, checked whether they fell within the image size (i.e. $\geq (0,0)$ and $< (10,10)$). Each time such tuple fell within the borders of the image, a perturbation of size 0.1 was added to the location (x,y) in the image.

This perturbation regime was added as a mask to the original image reflected by Z, to create the post-perturbation X. For the construction of ϕs, we created five

different 1-d convolution transformations, with randomly initialized weights, which will be indexed by Z and applied to X. There will be 4 resulting ϕs for each image, each corresponding to the 4 quadrants of the image. ϕ_1 most clearly varies with W, with potential little effect on ϕ_2 and ϕ_3, but ϕ_4 should see close to no effect in response to W, based on perturbation locations. Finally, Y was constructed as a linear combination of ϕs, with the weights [0.7, 0, 0, -0.5] applied to them, and added Gaussian noise $N(0, 0.1)$.

Finally, for the test set with an unseen perturbation pattern w', 2,000 examples that were featured in the training set of size 10,000, were used with a different perturbation, with location $w' = 5$. Previous perturbations ranged from 0-3. For specific structural equations corresponding to this description, see Figure 5 in the manuscript.

For the g model, estimating $\mathbb{E}[\Phi \mid w', Z]$, we train a Multi-task MLP with 3 hidden layers, 512 hidden dimensions, and the ReLU activation function. We use 100 epochs and 50 epochs to train the first-stage model g for seen w, and for the unseen w' train split respectively. We use the Adam optimizer, with learning rates of 0.002 and 0.001 used for optimization of each stage respectively. We use a train batch size of 400, and test batch size of 100. We set the seed at 42. We will also include the code to reproduce the results, see code files for any additional hyperparameter setting.

Finally, for the robustness to settings of θ experiment (Figure 8), we had to sample the weights from a distribution such that we can repeat the process 1500 times. We chose a uniform distribution $\theta \sim Unif(-0.3, 0.3)$, and added noise to Y from a $N(0, 0.01)$. We chose those such that similar stats of Y can be achieved, compared to the single Y setting experiment.

4.5.2.2 Humor Micro Edits

The construction of the humor micro edits dataset closely followed the analysis and description in Hossain et al. (2019). The original datasets provided there already contained original headlines (which we used as Z for our purposes), edit words (W for us) and humorous post-edit headlines (X)[10]. Following the analysis in the paper, we chose to represent each of these sentences and words using their pre-trained 6 Billion-token GloVe word-embedding vector representations, trained originally on 2014 English Wikipedia and Gigaword 5[11] (Pennington et al., 2014). We only included examples in which all words were correctly identified in the pre-trained word embedding, following a standard cleaning procedure (see code for exact details).

[10]Access to original dataset at `https://www.cs.rochester.edu/u/nhossain/humicroedit.html`.
[11]Available for download at `https://nlp.stanford.edu/projects/glove/`.

Z, W and X where does based on vector representation of the original dataset. There are three additional steps we have taken to compose the final dataset. First, we clustered the edit words using K-means clustering with K=20 (implemented via pedregosa2011scikit), following the same procedure carried out in Hossain et al. (2019). We used the labels of these clusters to create the training and test split, such that the test set included edit words from one cluster we left out to function as an unseen intervention W'. The unseen intervention was chosen to be one of the larger 5 clusters, to ensure enough training examples for estimation exist. The cluster that was randomly chosen for the results shown in the paper is cluster 11.

Next, we constructed high-level descriptions of the intervention and its implications on X as ϕ. $\Phi = \{\phi\}_{i=1}^{30}$, and each one was inspired by analysis and hypotheses from Hossain et al. (2019):

1. ϕ_1: Length of resulting edited sentence *(does not vary with w)*
2. ϕ_2: Mean cosine distance between GloVe vector of edit word and the rest of words in sentence *(varies with w)*
3. ϕ_3: Location index of replaced word *(does not vary with w)*
4. ϕ_4: Sentiment polarity of edit word, using the pre-trained sentiment processor from (Qi et al., 2020)[12] *(varies with w)*
5. ϕ_5: Sentiment polarity of resulting sentence, using the pre-trained sentiment processor from (Qi et al., 2020) (does not vary with w)
6. ϕ_6: Cosine distance between GloVe vector of edited word and GloVe vector of original word *(varies with w)*
7. ϕ_7-ϕ_{10}: Cosine distance of GloVe vector of edit word from neighboring words (2 preceding, 2 succeeding) *(does not vary with w)*
8. ϕ_{11}-ϕ_{30}: Distance of mean GloVe embedding of final sentence from clusters' centroids *(does not vary with w)*

The set of ϕs, which all correspond to different data types, were all scaled to have 0 mean and unit variance, to make them more comparable. Finally, following Φ as defined above we constructed an outcome variable Y as a linear combination of Φ, with added noise sampled from $N(0, .5)$. We sampled weights for each $\phi, \theta \sim U(-1, 1)$,

[12]Full usage details available at `https://stanfordnlp.github.io/stanza/sentiment.html`.

while keeping a random third of the weights at 0. Additionally, we ensured at least one of the ϕs varying with W is also zeroed out, to have diversity of all possible cases present.

For the g model, estimating $\mathbb{E}[\Phi \mid w', Z]$, we train a Multi-task MLP with 3 hidden layers, 512 hidden dimensions, and the ReLU activation function. We use 700 epochs and 100 epochs to train the first-stage model g for seen w, and for the unseen w' train split respectively. We use the Adam optimizer, with learning rates of 0.002 and 0.001 used for optimization of each stage respectively. We use a train batch size of 400, and test batch size of 100. We set the seed at 42. We also include the code to reproduce the results, see code files for any additional hyperparameter setting.

4.5.2.3 Gene Knockouts

Our GENIE3 model follows the instructions of Huynh-Thu et al. (2010), who scale all genes prior to analysis and fit a series of random forest regressions predicting the expression of each "downstream" gene as a function of the 334 candidate transcription factors (TFs). Each forest contains 1000 trees, with $mtry = \sqrt{334}$. The adjacency matrix is computed using the impurity importance measure originally proposed by Breiman (2001).

We simulate TF data from a multivariate Gaussian distribution with parameters estimated via maximum likelihood. This matrix is then propagated through our GENIE3 model to simulate expression values for downstream genes, with random Gaussian noise $\mathcal{N}(0, \sigma^2)$, where σ is the RMSE of the corresponding random forest on out-of-bag data. This data – TFs and downstream genes – together comprise the matrix Z of simulated baseline *E. Coli* gene expression.

We sort TFs by outdegree and simulate a knockout experiment on the top ten by replacing their values with a scalar 1 unit less than the observed minimum for each (how much less is irrelevant, as random forests are invariant to monotone transformations). We record the outdegreee of these TFs as W and the resulting expression matrix as X.

To compute Φ, we filter out all TFs with outdegree less than 100 and treat each of the remaining 168 TFs as the hub of a module. For downstream genes, module membership is determined by whether the given TF was assigned importance of at least 10 in the GENIE3 adjacency matrix. For each module, we compute the first kernel principal component on a subsample of $n = 1000$ using a radial basis function with default bandwidth given by the median Euclidean distance. These weights are then used to project the remaining data Z and X into the latent space. We define Φ as the difference between pre- and post-intervention expression values for the kernel

eigengene. We proceed to estimate a series of $\mathbb{E}[\phi_j \mid Z, W]$ regressions on a training set comprising 8 of 10 w values using random forests with 500 trees and $mtry = p/3$, where p is the number of genes in a given module.

Each ϕ_j is sorted by its association with W using Spearman correlation. Y is then simulated as a linear function of the top and bottom 25 ϕ's, with nonzero weights drawn from $\mathcal{N}(\pm 4, 1)$, where ± 4 denotes that the amplitude is multiplied by -1 with probability 0.5. A lasso regression $\mathbb{E}[Y \mid \hat{\Phi}]$ is fit to the aforementioned training set, with L_1 penalty λ selected via 10-fold cross-validation. Since, by construction $Z \perp\!\!\!\perp W$ in this experiment, the conditional independence tests can be replaced by a marginal independence test. We use the Spearman correlation to measure the association between W and each ϕ_j using the training set.

4.6 Conclusion

In this chapter, we proposed a general problem setup with applications in various fields of scientific study and policy design. We showed its relevance to different modalities and domain subjects and developed a general estimation framework. This enables the study of causal effects of crude interventions on high-dimensional, complex objects that impact an outcome of interest via some high-level mediator(s). Our approach is useful when one wishes to estimate the effects of new interventions for which little labeled data is available. We further showed how such a method can illuminate the underlying causal structure governing the process by identifying pragmatic mediation pathways between the complex objects and the outcome. Future work could extend this approach to various tasks, including: estimation of causal effects in response to high-dimensional and/or soft interventions; hypothesis design via a search for interventions predicted to achieve the outcome of interest; and prediction and mediation analysis with latent variables or partial knowledge of the true causal graph.

Limitations. We see this work as a first step in the study of complex causes and crude interventions, which are distinct from the atomic, soft, or stochastic interventions that have been previously studied. Though our method is focused on a particular problem setup, we have argued that a wide variety of problems share a similar structure. Additional work could make the method more applicable in settings with unobserved confounders, or where known relationships within objects (e.g., spatial characteristics) could be exploited for greater sensitivity. Other interesting extensions of this work could examine cases where X does not fully mediate W, or when the set of abstract

features Φ is not fully known. For the former, we believe that if the direct effect of W on Y is sufficiently weak, it should still be possible to exploit X as a mediator. For the latter, we envision adding smoothness constraints or parametric assumptions on Φ, or simultaneously learning the abstract features, as in (Xu et al., 2020).

5 | LENS: Local Explanations via Necessity and Sufficiency

5.1 Introduction

Machine learning algorithms are increasingly used in a variety of high-stakes domains, from credit scoring to medical diagnosis. However, many such methods are *opaque*, in that humans cannot understand the reasoning behind particular predictions. Post-hoc, model-agnostic local explanation tools (e.g., feature attributions, rule lists, and counterfactuals) are at the forefront of a fast-growing area of research variously referred to as *interpretable machine learning* or *explainable artificial intelligence* (XAI).

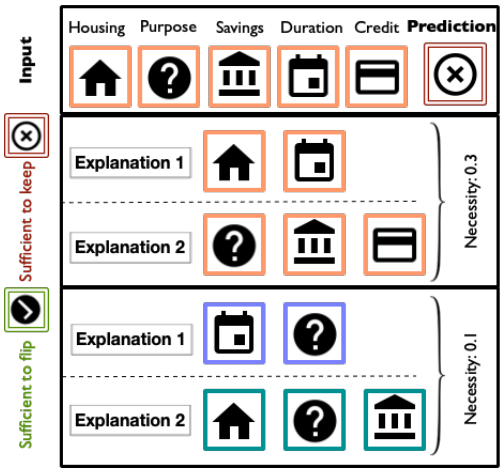


Figure 5.1: We describe minimal sufficient factors (here, sets of features) for a given input (top row), with the aim of preserving or flipping the original prediction. We report a sufficiency score for each set and a cumulative necessity score for all sets, indicating the proportion of paths toward the outcome that are covered by the explanation. Feature colors indicate the source of feature values (input or reference).

Many authors have pointed out the inconsistencies between popular XAI tools, raising questions as to which method is more reliable in particular cases (Mothilal et al., 2020a; Ramon et al., 2020; Fernández-Loría et al., 2020). Theoretical foundations have proven elusive in this area, perhaps due to the perceived subjectivity inherent to notions such as "intelligible" and "relevant" (Watson and Floridi, 2020). Practitioners often seek refuge in the axiomatic guarantees of Shapley values, which have become

the de facto standard in many XAI applications, due in no small part to their attractive theoretical properties (Bhatt et al., 2020). However, ambiguities regarding the underlying assumptions of the method (Kumar et al., 2020) and the recent proliferation of mutually incompatible implementations (Sundararajan and Najmi, 2019; Merrick and Taly, 2020) have complicated this picture. Despite the abundance of alternative XAI tools (Molnar, 2021), a dearth of theory persists. This has led some to conclude that the goals of XAI are underspecified (Lipton, 2018), and even that post-hoc methods do more harm than good (Rudin, 2019).

We argue that this lacuna at the heart of XAI should be filled by a return to fundamentals – specifically, to *necessity* and *sufficiency*. As the building blocks of all successful explanations, these dual concepts deserve a privileged position in the theory and practice of XAI. Following a review of related work (Sect. 5.2), we operationalize this insight with a unified framework (Sect. 5.3) that reveals unexpected affinities between various XAI tools and probabilities of causation (Sect. 5.4). We proceed to implement a novel procedure for computing model explanations that improves upon the state of the art in various quantitative and qualitative comparisons (Sect. 5.5). We conclude with a discussion and directions for future work (Sect. 5.7).

We make three main contributions. (1) We present a formal framework for XAI that unifies several popular approaches, including feature attributions, rule lists, and counterfactuals. (2) We introduce novel measures of necessity and sufficiency that can be computed for any feature subset. The method enables users to incorporate domain knowledge, search various subspaces, and select a utility-maximizing explanation. (3) We present a sound and complete algorithm for identifying explanatory factors and illustrate its performance on a range of tasks.

5.2 Background

Necessity and sufficiency have a long philosophical tradition (Mackie, 1965; Lewis, 1973; Halpern and Pearl, 2005b), spanning logical, probabilistic, and causal variants. In propositional logic, we say that x is a sufficient condition for y iff $x \rightarrow y$, i.e., x implies y; x is a necessary condition for y iff $y \rightarrow x$, i.e., y implies x. So stated, necessity and sufficiency are logically *converse*. However, by the law of contraposition, both definitions admit alternative formulations, whereby sufficiency may be rewritten as $\neg y \rightarrow \neg x$ and necessity as $\neg x \rightarrow \neg y$. By pairing the original definition of sufficiency with the latter definition of necessity (and vice versa), we find that the two concepts are also logically *inverse*. These formulae suggest probabilistic relaxations, measuring

x's sufficiency for y by $P(y|x)$ and x's necessity for y by $P(x|y)$. Because there is no probabilistic law of contraposition, these quantities are generally uninformative w.r.t. $P(\neg x|\neg y)$ and $P(\neg y|\neg x)$, which may be of independent interest. Thus, while necessity is both the converse and inverse of sufficiency in propositional logic, the two formulations come apart in probability calculus. We revisit the distinction between probabilistic conversion and inversion in Rmk. 1 and Sect. 5.4.

These definitions struggle to track our intuitions when we consider causal explanations (Pearl, 2000; Tian and Pearl, 2000). It may make sense to say in logic that if x is a necessary condition for y, then y is a sufficient condition for x; it does not follow that if x is a necessary *cause* of y, then y is a sufficient *cause* of x. We may amend both concepts using *counterfactual probabilities* – e.g., the probability that Alice would still have a headache if she had not taken an aspirin, given that she does not have a headache and did take an aspirin. Let $P(y_x|x', y')$ denote such a quantity, to be read as "the probability that Y would equal y under an intervention that sets X to x, given that we observe $X = x'$ and $Y = y'$." Then, according to Pearl (2000, Ch. 9), the probability that x is a sufficient cause of y is given by $\texttt{suf}(x, y) := P(y_x|x', y')$, and the probability that x is a necessary cause of y is given by $\texttt{nec}(x, y) := P(y'_{x'}|x, y)$.

Analysis becomes more difficult in higher dimensions, where variables may interact to block or unblock causal pathways. VanderWeele and Robins (2008) analyze sufficient causal interactions in the potential outcomes framework, refining notions of synergism without monotonicity constraints. In a subsequent paper, VanderWeele and Richardson (2012) study the irreducibility and singularity of interactions in sufficient-component cause models. Halpern (2016) devotes an entire monograph to the subject, providing various criteria to distinguish between subtly different notions of "actual causality", as well as "but-for" (similar to necessary) and sufficient causes. These authors generally limit their analyses to Boolean systems with convenient structural properties, e.g. conditional ignorability and the stable unit treatment value assumption. Operationalizing their theories in a practical method without such restrictions is one of our primary contributions.

Necessity and sufficiency have begun to receive explicit attention in the XAI literature. Ribeiro et al. (2018a) propose a bandit procedure for identifying a minimal set of Boolean conditions that entails a predictive outcome (more on this in Sect. 5.4). Dhurandhar et al. (2018) propose an autoencoder for learning pertinent negatives and positives, i.e. features whose presence or absence is decisive for a given label, while Zhang et al. (2018) develop a technique for generating symbolic corrections to

alter model outputs. Both methods are optimized for neural networks, unlike the model-agnostic approach we develop here.

Another strand of research in this area is rooted in logic programming. Several authors have sought to reframe XAI as either a SAT (Ignatiev et al., 2019; Narodytska et al., 2019) or a set cover problem (Lakkaraju et al., 2019; Grover et al., 2019), typically deriving approximate solutions on a prespecified subspace to ensure computability in polynomial time. We adopt a different strategy that prioritizes completeness over efficiency, an approach we show to be feasible in moderate dimensions (see Sect. 7.7 for a discussion). Mothilal et al. (2020a) build on Halpern (2016)'s definitions of necessity and sufficiency to critique popular XAI tools, proposing a new feature attribution measure with some purported advantages. Their method relies on the strong assumption that predictors are mutually independent. Galhotra et al. (2021) adapt Pearl (2000)'s probabilities of causation for XAI under a more inclusive range of data-generating processes. They derive analytic bounds on multidimensional extensions of `nec` and `suf`, as well as an algorithm for point identification when graphical structure permits. Oddly, they claim that non-causal applications of necessity and sufficiency are somehow "incorrect and misleading" (p. 2), a normative judgment that is inconsistent with many common uses of these concepts. Rather than insisting on any particular interpretation of necessity and sufficiency, we propose a general framework that admits logical, probabilistic, and causal interpretations as special cases. Whereas previous works evaluate individual predictors, we focus on feature *subsets*, allowing us to detect and quantify interaction effects. Our formal results clarify the relationship between existing XAI methods and probabilities of causation, while our empirical results demonstrate their applicability to a wide array of tasks and datasets.

5.3 Proposed Framework

We propose a unifying framework that highlights the role of necessity and sufficiency in XAI. Its constituent elements are described below.

Target function. Post-hoc explainability methods assume access to a target function $f : \mathcal{X} \mapsto \mathcal{Y}$, i.e. the model whose prediction(s) we seek to explain. For simplicity, we restrict attention to the binary setting, with $Y \in \{0, 1\}$. Multi-class extensions are straightforward, while continuous outcomes may be accommodated via discretization. Though this inevitably involves some information loss, we follow authors in the contrastivist tradition in arguing that, even for continuous outcomes, explanations

always involve a juxtaposition (perhaps implicit) of "fact and foil" (Lipton, 1990). For instance, a loan applicant is probably less interested in knowing why her credit score is precisely y than she is in discovering why it is below some threshold (say, 700). Of course, binary outcomes can approximate continuous values with arbitrary precision over repeated trials.

Context. The context $\mathcal{D}$ is a probability distribution over which we quantify sufficiency and necessity. Contexts may be constructed in various ways but always consist of at least some input (point or space) and reference (point or space). For instance, we may want to compare $\boldsymbol{x}_i$ with all other samples, or else just those perturbed along one or two axes, perhaps based on some conditioning event(s).

In addition to predictors and outcomes, we optionally include information exogenous to f. For instance, if any events were conditioned upon to generate a given reference sample, this information may be recorded among a set of auxiliary variables $\boldsymbol{W}$. Other examples of potential auxiliaries include metadata or engineered features such as those learned via neural embeddings. This augmentation allows us to evaluate the necessity and sufficiency of factors beyond those found in $\boldsymbol{X}$. Contextual data take the form $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{W}) \sim \mathcal{D}$. The distribution may or may not encode dependencies between (elements of) $\boldsymbol{X}$ and (elements of) $\boldsymbol{W}$. We extend the target function to augmented inputs by defining $f(\boldsymbol{z}) := f(\boldsymbol{x})$, which means the outcome under the model f for input $\boldsymbol{x}$ will be taken to be the model outcome of the augmentation $\boldsymbol{z}$, as the only difference is that it additionally includes the auxiliary information tagging $\boldsymbol{w}$, but will not change the prediction model's output for that instance.

Factors. Factors pick out the properties whose necessity and sufficiency we wish to quantify. Formally, a factor $c : \mathcal{Z} \mapsto \{0, 1\}$ indicates whether its argument satisfies some criteria with respect to predictors or auxiliaries. For instance, if $\boldsymbol{x}$ is an input to a credit lending model, and $\boldsymbol{w}$ contains information about the subspace from which data were sampled, then a factor could be $c(\boldsymbol{z}) = \mathbb{1}[\boldsymbol{x}[\mathsf{gender} = \text{"female"}] \wedge \boldsymbol{w}[do(\mathsf{income} > \$50\text{k})]]$, i.e. checking if $\boldsymbol{z}$ is female and drawn from a context in which an intervention fixes income at greater than \$50k. We use the term "factor" as opposed to "condition" or "cause" to suggest an inclusive set of criteria that may apply to predictors $\boldsymbol{x}$ and/or auxiliaries $\boldsymbol{w}$. Such criteria are always observational w.r.t. $\boldsymbol{z}$ but may be interventional or counterfactual w.r.t. $\boldsymbol{x}$[1]. We assume a finite space of factors $\mathcal{C}$.

[1]At first glance, this statement may seem confusing. However, notice $\boldsymbol{x}$ refers to predictors or feature assignments which can be interventional or counterfactual. However, this nature of $\boldsymbol{x}$ is described in the auxiliary information $\boldsymbol{w}$, rendering the criteria observational in $\boldsymbol{z}$ as a whole.

Partial order. When multiple factors pass a given necessity or sufficiency threshold, users will tend to prefer some over others. For instance, factors with fewer conditions are often preferable to those with more, all else being equal; factors that change a variable by one unit as opposed to two are preferable, and so on. Rather than formalize this preference in terms of a distance metric, which unnecessarily constrains the solution space, we treat the partial ordering as primitive and require only that it be complete and transitive. This covers not just distance-based measures but also more idiosyncratic orderings that are unique to individual agents. Ordinal preferences may be represented by cardinal utility functions under reasonable assumptions (see, e.g., von Neumann and Morgenstern 1944).

We are now ready to formally specify our framework.

Definition 1 (Basis)**.** A *basis* for computing necessary and sufficient factors for model predictions is a tuple $\mathcal{B} = \langle f, \mathcal{D}, \mathcal{C}, \preceq \rangle$, where f is a target function, $\mathcal{D}$ is a context, $\mathcal{C}$ is a set of factors, and $\preceq$ is a partial ordering on $\mathcal{C}$.

5.3.1 Explanatory Measures

For some fixed basis $\mathcal{B} = \langle f, \mathcal{D}, \mathcal{C}, \preceq \rangle$, we define the following measures of sufficiency and necessity, with probability taken over $\mathcal{D}$.

Definition 2 (Probability of Sufficiency)**.** The probability that c is a sufficient factor for outcome y is given by:

$$PS(c, y) := P(f(\boldsymbol{z}) = y \mid c(\boldsymbol{z}) = 1).$$

The probability that factor set $C = \{c_1, \ldots, c_k\}$ is sufficient for y is given by:

$$PS(C, y) := P(f(\boldsymbol{z}) = y \mid \sum_{i=1}^{k} c_i(\boldsymbol{z}) \geq 1).$$

Definition 3 (Probability of Necessity)**.** The probability that c is a necessary factor for outcome y is given by:

$$PN(c, y) := P(c(\boldsymbol{z}) = 1 \mid f(\boldsymbol{z}) = y).$$

The probability that factor set $C = \{c_1, \ldots, c_k\}$ is necessary for y is given by:

$$PN(C, y) := P(\sum_{i=1}^{k} c_i(\boldsymbol{z}) \geq 1 \mid f(\boldsymbol{z}) = y).$$

Remark 1. These probabilities can be likened to the "precision" (positive predictive value) and "recall" (true positive rate) of a (hypothetical) classifier that predicts whether $f(\boldsymbol{z}) = y$ based on whether $c(\boldsymbol{z}) = 1$. By examining the confusion matrix of this classifier, one can define other related quantities, e.g. the true negative rate $P(c(\boldsymbol{z}) = 0 | f(\boldsymbol{z}) \neq y)$ and the negative predictive value $P(f(\boldsymbol{z}) \neq y | c(\boldsymbol{z}) = 0)$, which are contrapositive transformations of our proposed measures. We can recover these values exactly via $PS(1-c, 1-y)$ and $PN(1-c, 1-y)$, respectively. When necessity and sufficiency are defined as probabilistic inversions (rather than conversions), such transformations are impossible.

5.3.2 Minimal Sufficient Factors

We introduce Local Explanations via Necessity and Sufficiency (LENS), a procedure for computing explanatory factors with respect to a given basis $\mathcal{B}$ and threshold parameter τ (see Alg. 3). First, we calculate a factor's probability of sufficiency (see probSuff) by drawing n samples from $\mathcal{D}$ and taking the maximum likelihood estimate $\hat{PS}(c, y)$. Next, we sort the space of factors w.r.t. $\preceq$ in search of those that are τ-minimal.

Definition 4 (τ-minimality)**.** We say that c is τ-minimal iff (i) $PS(c, y) \geq \tau$ and (ii) there exists no factor c' such that $PS(c', y) \geq \tau$ and $c' \prec c$.

Since a factor is necessary to the extent that it covers all possible pathways towards a given outcome, our next step is to span the τ-minimal factors and compute their cumulative PN (see probNec). As a minimal factor c stands for all c' such that $c \preceq c'$, in reporting probability of necessity, we expand C to its upward closure.

Thms. 2 and 3 state that this procedure is *optimal* in a sense that depends on whether we assume access to oracle or sample estimates of PS.

Theorem 2. With oracle estimates $PS(c, y)$ for all $c \in \mathcal{C}$, Alg. 3 is sound and complete. That is, for any C returned by Alg. 3 and all $c \in \mathcal{C}$, c is τ-minimal iff $c \in C$.

Proof. Soundness and completeness follow directly from the specification of (P1) $\mathcal{C}$ and (P2) $\preceq$ in the algorithm's input $\mathcal{B}$, along with (P3) access to oracle estimates $PS(c, y)$ for all $c \in \mathcal{C}$. Recall that the partial ordering must be complete and transitive, as noted in Sect. 5.3.

Assume that Alg. 3 generates a false positive, i.e. outputs some c that is not τ-minimal. Then by Def. 4, either the algorithm failed to properly evaluate $PS(c, y)$,

thereby violating (P3); or failed to identify some c' such that (i) $PS(c', y) \geq \tau$ and (ii) $c' \prec c$. (i) is impossible by (P3), and (ii) is impossible by (P2). Thus there can be no false positives.

Assume that Alg. 3 generates a false negative, i.e. fails to output some c that is in fact τ-minimal. By (P1), this c cannot exist outside the finite set $\mathcal{C}$. Therefore there must be some $c \in \mathcal{C}$ for which either the algorithm failed to properly evaluate $PS(c, y)$, thereby violating (P3); or wrongly identified some c' such that (i) $PS(c', y) \geq \tau$ and (ii) $c' \prec c$. Once again, (i) is impossible by (P3), and (ii) is impossible by (P2). Thus there can be no false negatives. □

Theorem 3. With sample estimates $\hat{PS}(c, y)$ for all $c \in \mathcal{C}$, Alg. 3 is uniformly most powerful. That is, Alg. 3 identifies the most τ-minimal factors of any method with fixed type I error α.

Proof. A testing procedure is uniformly most powerful (UMP) if it attains the lowest type II error β of all tests with fixed type I error α. Let Θ_0, Θ_1 denote a partition of the parameter space into null and alternative regions, respectively. The goal in frequentist inference is to test the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta_1$ for some parameter θ. Let $\psi(X)$ be a testing procedure of the form $\mathbb{1}[T(X) \geq c_\alpha]$, where X is a finite sample, $T(X)$ is a test statistic, and c_α is the critical value. This latter parameter defines a rejection region such that test statistics integrate to α under H_0. We say that $\psi(X)$ is UMP iff, for any other test $\psi'(X)$ such that

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\psi'(X)] \leq \alpha,$$

we have

$$(\forall \theta \in \Theta_1)\ \mathbb{E}_\theta[\psi'(X)] \leq \mathbb{E}_\theta[\psi(X)],$$

where $\mathbb{E}_{\theta \in \Theta_1}[\psi(X)]$ denotes the power of the test to detect the true θ, $1 - \beta_\psi(\theta)$. The UMP-optimality of Alg. 3 follows from the UMP-optimality of the binomial test (see (Lehmann and Romano, 2005, Ch. 3)), which is used to decide between $H_0 : PS(c, y) < \tau$ and $H_1 : PS(c, y) \geq \tau$ on the basis of observed proportions $\hat{PS}(c, y)$, estimated from n samples for all $c \in \mathcal{C}$. The proof now takes the same structure as that of Thm. 2, with (P3) replaced by (P3′): access to UMP estimates of $PS(c, y)$. False positives are no longer impossible but bounded at level α; false negatives are no

longer impossible but occur with frequency β. Because no procedure can find more τ-minimal factors for any fixed α, Alg. 3 is UMP. □

Multiple testing adjustments can easily be accommodated, in which case modified optimality criteria apply (Storey, 2007).

Remark 2. We take it that the main quantity of interest in most applications is sufficiency, be it for the original or alternative outcome, and therefore define τ-minimality w.r.t. sufficient (rather than necessary) factors. However, necessity serves an important role in tuning τ, as there is an inherent trade-off between the parameters. More factors are excluded at higher values of τ, thereby inducing lower cumulative PN; more factors are included at lower values of τ, thereby inducing higher cumulative PN. The relationship between τ and cumulative probabilities of necessity is similar to a precision-recall curve quantifying and qualifying errors in classification tasks. In this case, as we lower τ, we allow more factors to be taken into account, thus covering more pathways towards a desired outcome in a cumulative sense. We provide an example of such a precision-recall curve in Fig. 5.2, using an R2I view of the `German` credit dataset. Different levels of cumulative necessity may be warranted for different tasks, depending on how important it is to survey multiple paths towards an outcome. Users can therefore adjust τ to accommodate desired levels of cumulative PN over successive calls to LENS.

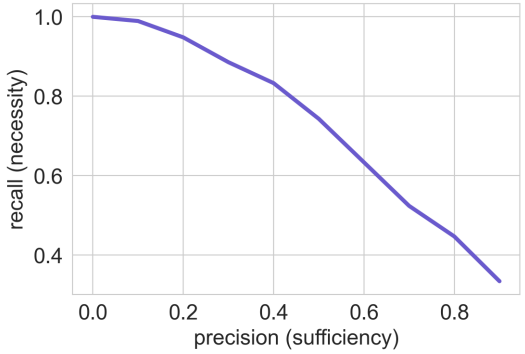


Figure 5.2: An example curve exemplifying the relationship between τ and cumulative probability necessity attained by selected τ-minimal factors.

5.4 Recover Existing Methods

Explanatory measures can be shown to play a central role in many seemingly unrelated XAI tools, albeit under different assumptions about the basis tuple $\mathcal{B}$. In this section, we relate our framework to a number of existing methods.

Algorithm 3 LENS

Input: $\mathcal{B} = \langle f, \mathcal{D}, \mathcal{C}, \preceq \rangle, \tau$
Output: Factor set C, $(\forall c \in C)\ PS(c, y), PN(C, y)$

Sample $\hat{D} = \{\boldsymbol{z}_i\}_{i=1}^{n} \sim \mathcal{D}$

function probSuff(c, y)
    n(c&y) = $\sum_{i=1}^{n} \mathbb{1}[c(\boldsymbol{z}_i) = 1 \wedge f(\boldsymbol{z}_i) = y]$
    n(c) = $\sum_{i=1}^{n} c(\boldsymbol{z}_i)$
    return n(c&y) / n(c)

function probNec(C, y, upward_closure_flag)
    if upward_closure_flag **then**
        $C = \{c \mid c \in \mathcal{C} \wedge \exists\ c' \in C : c' \preceq c\}$
    end if
    n(C&y) = $\sum_{i=1}^{n} \mathbb{1}[\sum_{j=1}^{k} c_j(\boldsymbol{z}_i) \geq 1 \wedge f(\boldsymbol{z}_i) = y]$
    n(y) = $\sum_{i=1}^{n} \mathbb{1}[f(\boldsymbol{z}_i) = y]$
    return n(C&y) / n(y)

function minimalSuffFactors(y, τ, sample_flag, α)
    sorted_factors = topological_sort($\mathcal{C}, \preceq$)
    cands = []
    for c in sorted_factors **do**
        if $\exists(c', \_) \in$ cands $: c' \preceq c$ **then**
            continue
        end if
        ps = probSuff(c, y)
        if sample_flag **then**
            p = binom.test(n(c&y), n(c), τ, alt = >)
            if p $\leq \alpha$ **then**
                cands.append(c, ps)
            end if
        else if ps $\geq \tau$ **then**
            cands.append(c, ps)
        end if
    end for
    cum_pn = probNec($\{c \mid (c, \_) \in$ cands$\}, y$, TRUE)
    return cands, cum_pn

Feature Attributions. Several popular feature attribution algorithms are based on Shapley values (Shapley, 1953), which decompose the predictions of any target

function as a sum of weights over d input features:

$$f(\boldsymbol{x}_i) = \phi_0 + \sum_{j=1}^{d} \phi_j, \tag{5.1}$$

where ϕ_0 represents a baseline expectation and ϕ_j the weight assigned to X_j at point $\boldsymbol{x}_i$. Let $v : 2^d \mapsto \mathbb{R}$ be a value function such that $v(S)$ is the payoff associated with feature subset $S \subseteq [d]$ and $v(\{\varnothing\}) = 0$. Define the complement $R = [d] \backslash S$ such that we may rewrite any $\boldsymbol{x}_i$ as a pair of subvectors, $(\boldsymbol{x}_i^S, \boldsymbol{x}_i^R)$. Payoffs are given by:

$$v(S) = \mathbb{E}[f(\boldsymbol{x}_i^S, \boldsymbol{X}^R)], \tag{5.2}$$

although this introduces some ambiguity regarding the reference distribution for $\boldsymbol{X}^R$ (more on this below). The Shapley value ϕ_j is then j's average marginal contribution to all subsets that exclude it:

$$\phi_j = \sum_{S \subseteq [d] \backslash \{j\}} \frac{|S|!(d - |S| - 1)!}{d!} v(S \cup \{j\}) - v(S). \tag{5.3}$$

It can be shown that this is the unique solution to the attribution problem that satisfies certain desirable properties, including efficiency, linearity, sensitivity, and symmetry.

Reformulating this in our framework, we find that the value function v is a sufficiency measure. To see this, let each $\boldsymbol{z} \sim \mathcal{D}$ be a sample in which a random subset of variables S are held at their original values while remaining features R are drawn from a fixed distribution $\mathcal{D}(\cdot|S)$.[2]

Proposition 1. Let $c_S(\boldsymbol{z}) = 1$ iff $\boldsymbol{x} \subseteq \boldsymbol{z}$ was constructed by holding $\boldsymbol{x}^S$ fixed and sampling $\boldsymbol{X}^R$ according to $\mathcal{D}(\cdot|S)$. Then $v(S) = PS(c_S, y)$.

Proof. As noted in the text, $\mathcal{D}(\boldsymbol{x}|S)$ may be defined in a variety of ways (e.g., via marginal, conditional, or interventional distributions). For any given choice, let $c_S(\boldsymbol{z}) = 1$ iff $\boldsymbol{x}$ is constructed by holding $\boldsymbol{x}_i^S$ fixed and sampling $\boldsymbol{X}^R$ according to $\mathcal{D}(\boldsymbol{x}|S)$. Since we assume binary Y (or binarized, as discussed in Sect. 5.3), we can rewrite Eq. 5.2 as a probability:

$$v(S) = P_{\mathcal{D}(\boldsymbol{x}|S)}(f(\boldsymbol{x}_i) = f(\boldsymbol{x})),$$

[2]The diversity of Shapley value algorithms is largely due to variation in how this distribution is defined. Popular choices include the marginal $P(\boldsymbol{X}^R)$ (Lundberg and Lee, 2017); conditional $P(\boldsymbol{X}^R|\boldsymbol{x}^S)$ (Aas et al., 2021); and interventional $P(\boldsymbol{X}^R|do(\boldsymbol{x}^S))$ (Heskes et al., 2020) distributions.

where $\boldsymbol{x}_i$ denotes the input point. Since conditional sampling is equivalent to conditioning after sampling, this value function is equivalent to $PS(c_S, y)$ by Def. 2. □

Thus, the Shapley value ϕ_j measures X_j's average marginal increase to the sufficiency of a random feature subset. The advantage of our method is that, by focusing on particular subsets instead of weighting them all equally, we disregard irrelevant permutations and hone in on just those that meet a τ-minimality criterion. Kumar et al. (2020) observe that "since there is no standard procedure for converting Shapley values into a statement about a model's behavior, developers rely on their own mental model of what the values represent" (p. 8). By contrast, necessary and sufficient factors are more transparent and informative, offering a direct path to what Shapley values indirectly summarize.

Rule Lists. Rule lists are sequences of if-then statements that describe a hyperrectangle in feature space, creating partitions that can be visualized as decision or regression trees. Rule lists have long been popular in XAI. While early work in this area tended to focus on global methods, more recent efforts have prioritized local explanation tasks.

We focus in particular on the Anchors algorithm (Ribeiro et al., 2018a), which learns a set of Boolean conditions A (the eponymous "anchors") such that $A(\boldsymbol{x}_i) = 1$ and

$$P_{\mathcal{D}(\boldsymbol{x}|A)}(f(\boldsymbol{x}_i) = f(\boldsymbol{x})) \geq \tau. \tag{5.4}$$

The lhs of Eq. 5.4 is termed the *precision*, $\text{prec}(A)$, and probability is taken over a synthetic distribution in which the conditions in A hold while other features are perturbed. Once τ is fixed, the goal is to maximize *coverage*, formally defined as $\mathbb{E}[A(\boldsymbol{x}) = 1]$, i.e. the proportion of datapoints to which the anchor applies.

The formal similarities between Eq. 5.4 and Def. 2 are immediately apparent, and the authors themselves acknowledge that Anchors are intended to provide "sufficient conditions" for model predictions.

Proposition 2. Let $c_A(\boldsymbol{z}) = 1$ iff $A(\boldsymbol{x}) = 1$. Then $\text{prec}(A) = PS(c_A, y)$.

Proof. The proof for this proposition is essentially identical, except in this case our conditioning event is $A(\boldsymbol{x}) = 1$. Let $c_A = 1$ iff $A(\boldsymbol{x}) = 1$. Precision $\text{prec}(A)$, given by the lhs of Eq. 5.4, is defined over a conditional distribution $\mathcal{D}(\boldsymbol{x}|A)$. Since

Table 5.1: Overview of experimental settings by basis configuration.

Experiment	Datasets	f	$\mathcal{D}$	$\mathcal{C}$	$\preceq$
Attribution comparison	`German`, `SpamAssassins`	`Extra-Trees`	R2I, I2R	Intervention targets	-
Anchors comparison: Brittle predictions	`IMDB`	`LSTM`	R2I, I2R	Intervention targets	$\preceq_{subset}$
Anchors comparison: PS and Prec	`German`	`Extra-Trees`	R2I	Intervention targets	$\preceq_{subset}$
Counterfactuals: Adverserial	`SpamAssassins`	`MLP`	R2I	Intervention targets	$\preceq_{subset}$
Counterfactuals: Recourse, DiCE comparison	`Adult`	`MLP`	I2R	Full interventions	$\preceq_{cost}$
Counterfactuals: Recourse, causal vs. non-causal	`German`	`Extra-Trees`	I2R_{causal}	Full interventions	$\preceq_{cost}$

conditional sampling is equivalent to conditioning after sampling, this probability reduces to $PS(c_A, y)$. □

While Anchors outputs just a single explanation, our method generates a ranked list of candidates, thereby offering a more comprehensive view of model behavior. Moreover, our necessity measure adds a mode of explanatory information entirely lacking in Anchors.

Counterfactuals. Counterfactual explanations identify one or several nearest neighbors with different outcomes, e.g. all datapoints $\boldsymbol{x}$ within an ε-ball of $\boldsymbol{x}_i$ such that labels $f(\boldsymbol{x})$ and $f(\boldsymbol{x}_i)$ differ (for classification) or $f(\boldsymbol{x}) > f(\boldsymbol{x}_i) + \delta$ (for regression).[3] The optimization problem is:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \text{CF}(\boldsymbol{x}_i)}{\arg\min}\ cost(\boldsymbol{x}_i, \boldsymbol{x}), \tag{5.5}$$

where $\text{CF}(\boldsymbol{x}_i)$ denotes a counterfactual space such that $f(\boldsymbol{x}_i) \neq f(\boldsymbol{x})$ and *cost* is a user-supplied cost function, typically equated with some distance measure. (Wachter et al., 2018) recommend using generative adversarial networks to solve Eq. 5.5, while others have proposed alternatives designed to ensure that counterfactuals are coherent and actionable (Ustun et al., 2019; Karimi et al., 2020a). As with Shapley values, the variation in these proposals is reducible to the choice of context $\mathcal{D}$.

For counterfactuals, we rewrite the objective as a search for minimal perturbations sufficient to flip an outcome.

Proposition 3. Let *cost* be a function representing $\preceq$, and let c be some factor spanning reference values. Then the counterfactual recourse objective is:

$$c^* = \underset{c \in \mathcal{C}}{\arg\min}\ cost(c) \quad \text{s.t. } PS(c, 1-y) \geq \tau, \tag{5.6}$$

[3]Confusingly, the term "counterfactual" in XAI refers to any point with an alternative outcome, which is distinct from the causal sense of the term (see Sect. 5.2). We use the word in both senses here, but strive to make our intended meaning explicit in each case.

where τ denotes a decision threshold. Counterfactual outputs will then be any $\boldsymbol{z} \sim \mathcal{D}$ such that $c^*(\boldsymbol{z}) = 1$.

Proof. There are two closely related ways of expressing the counterfactual objective: as a search for optimal *points*, or optimal *actions*. We start with the latter interpretation, reframing actions as factors. We are only interested in solutions that flip the original outcome, and so we constrain the search to factors that meet an I2R sufficiency threshold, $PS(c, 1-y) \geq \tau$. Then the optimal action is attained by whatever factor (i) meets the sufficiency criterion and (ii) minimizes cost. Call this factor c^*. The optimal point is then any $\boldsymbol{z}$ such that $c^*(\boldsymbol{z}) = 1$. □

Probabilities of Causation. Our framework can describe Pearl (2000)'s aforementioned probabilities of causation, however in this case $\mathcal{D}$ must be constructed with care.

Proposition 4. Consider the bivariate Boolean setting, as in Sect. 5.2. We have two counterfactual distributions: an input space $\mathcal{I}$, in which we observe x, y but intervene to set $X = x'$; and a reference space $\mathcal{R}$, in which we observe x', y' but intervene to set $X = x$. Let $\mathcal{D}$ denote a uniform mixture over both spaces and let auxiliary variable W tag each sample with a label indicating whether it comes from the original ($W = 1$) or contrastive ($W = 0$) counterfactual space. Define $c(\boldsymbol{z}) = w$. Then we have $\texttt{suf}(x, y) = PS(c, y)$ and $\texttt{nec}(x, y) = PS(1 - c, y')$.

Proof. Recall from Sect. 5.2 that Pearl (2000, Ch. 9) defines $\texttt{suf}(x, y) := P(y_x | x', y')$ and $\texttt{nec}(x, y) := P(y'_{x'} | x, y)$. We may rewrite the former as $P_{\mathcal{R}}(y)$, where the reference space $\mathcal{R}$ denotes a counterfactual distribution conditioned on $x', y', do(x)$. Similarly, we may rewrite the latter as $P_{\mathcal{I}}(y')$, where the input space $\mathcal{I}$ denotes a counterfactual distribution conditioned on $x, y, do(x')$. Our context $\mathcal{D}$ is a uniform mixture over both spaces.

The key point here is that the auxiliary variable W indicates whether samples are drawn from $\mathcal{I}$ or $\mathcal{R}$. Thus conditioning on different values of W allows us to toggle between probabilities over the two spaces. Therefore, for $c(\boldsymbol{z}) = w$, we have $\texttt{suf}(x, y) = PS(c, y)$ and $\texttt{nec}(x, y) = PS(1 - c, y')$.

In other words, we regard Pearl's notion of necessity as *sufficiency of the negated factor for the alternative outcome*. By contrast, Pearl (2000) has no analog for our probability of necessity. This is true of any measure that defines sufficiency and necessity via inverse, rather than converse probabilities. While conditioning on the

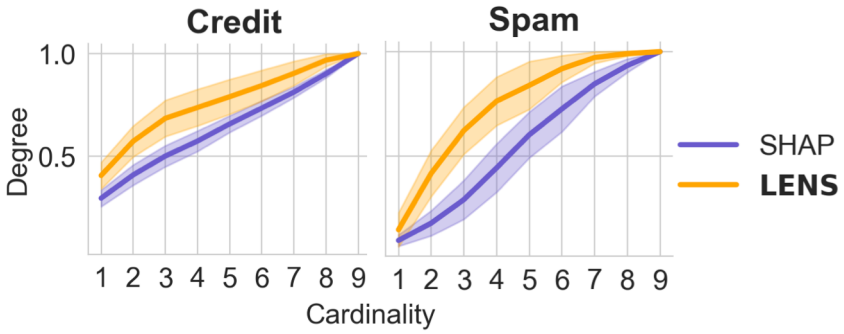


Figure 5.3: Comparison of top k features ranked by SHAP against the best-performing LENS subset of size k in terms of $PS(c, y)$. `German` results are over 50 inputs; `SpamAssassins` results are over 25 inputs.

same variable(s) for both measures may have some intuitive appeal, it comes at a cost to expressive power. Whereas our framework can recover all four explanatory measures, corresponding to the classical definitions and their contrapositive forms, definitions that merely negate instead of transpose the antecedent and consequent are limited to just two. □

Remark 3. We have assumed that factors and outcomes are Boolean throughout. Our results can be extended to continuous versions of either or both variables, so long as $c(\boldsymbol{Z}) \perp\!\!\!\perp Y \mid \boldsymbol{Z}$. This conditional independence holds whenever $\boldsymbol{W} \perp\!\!\!\perp Y \mid \boldsymbol{X}$, which is true by construction since $f(\boldsymbol{z}) := f(\boldsymbol{x})$. However, we defend the Boolean assumption on the grounds that it is well motivated by contrastivist epistemologies (Kahneman and Miller, 1986; Blaauw, 2013) and not especially restrictive, given that partitions of arbitrary complexity may be defined over $\boldsymbol{Z}$ and Y.

5.5 Experiments

In this section, we demonstrate the use of LENS on a variety of tasks and compare results with popular XAI tools, using the basis configurations detailed in Table 5.1. A comprehensive discussion of experimental design, including datasets and pre-processing pipelines, is left to Section 5.6. Code for reproducing all results is available at `https://github.com/limorigu/LENS`.

Contexts. We consider a range of contexts $\mathcal{D}$ in our experiments. For the input-to-reference (I2R) setting, we replace input values with reference values for feature subsets S; for the reference-to-input (R2I) setting, we replace reference values with

Table 5.2: Example prediction given by an LSTM model trained on the `IMDB` dataset. We compare τ-minimal factors identified by LENS (as individual words), based on $PS(c, y)$ and $PS(1-c, 1-y)$, and compare to output by Anchors.

Inputs		Anchors		LENS	
Text	Original model prediction	Suggested anchors	Precision	Sufficient R2I factors	Sufficient I2R factors
'read book forget movie'	wrongly predicted positive	[read, movie]	0.94	[read, forget, movie]	read, forget, movie
'you better choose paul verhoeven even watched'	correctly predicted negative	[choose, better, even, you, paul, verhoeven]	0.95	choose, even	better, choose, paul, even

input values. We use R2I for examining the sufficiency/necessity of the original model prediction, and I2R for examining the sufficiency/necessity of a contrastive model prediction. We sample from the empirical data in all experiments, except in Sect. 5.5.3, where we assume access to a structural causal model (SCM).

Partial Orderings. We consider two types of partial orderings in our experiments. The first, $\preceq_{subset}$, evaluates subset relationships. For instance, if $c(\boldsymbol{z}) = \mathbb{1}[\boldsymbol{x}[\mathsf{gender} = \text{“female”}]]$ and $c'(\boldsymbol{z}) = \mathbb{1}[\boldsymbol{x}[\mathsf{gender} = \text{“female”} \wedge \mathsf{age} \geq 40]]$, then we say that $c \preceq_{subset} c'$. The second, $c \preceq_{cost} c' := c \preceq_{subset} c' \ \wedge \ cost(c) \leq cost(c')$, adds the additional constraint that c has cost no greater than c'. The cost function could be arbitrary. Here, we consider distance measures over either the entire state space or just the intervention targets corresponding to c.

5.5.1 Feature Attributions

Feature attributions are often used to identify the top-k most important features for a given model outcome (Barocas et al., 2020). However, we argue that these feature sets may not be explanatory with respect to a given prediction. To show this, we compute R2I and I2R sufficiency – i.e., $PS(c, y)$ and $PS(1-c, 1-y)$, respectively – for the top-k most influential features ($k \in [1, 9]$) as identified by SHAP (Lundberg and Lee, 2017) and LENS. Fig. 5.3 shows results from the R2I setting for `German` credit (Dua and Graff, 2017) and `SpamAssassin` datasets (SpamAssassin, 2006). Our method attains higher PS for all cardinalities. We repeat the experiment over 50 inputs, plotting means and 95% confidence intervals for all k. Results indicate that our ranking procedure delivers more informative explanations than SHAP at any fixed degree of sparsity. Results from the I2R setting are in Section 5.6.

5.5.2 Rule Lists

Sentiment Sensitivity Analysis. Next, we use LENS to study model weaknesses by considering minimal factors with high R2I and I2R sufficiency in text models. Our

Table 5.3: (Top) A selection of emails from `SpamAssassins`, correctly identified as spam by an MLP. The goal is to find minimal perturbations that result in non-spam predictions. (Bottom) Minimal subsets of feature-value assignments that achieve non-spam predictions with respect to the emails above.

From	To	Subject	First Sentence	Last Sentence
resumevalet info resumevalet com	yyyy cv spamassassin taint org	adv put resume back work	dear candidate	professionals online network inc
jacqui devito goodroughy ananzi co za	picone linux midrange com	enlargement breakthrough zibdrzpay	recent survey conducted	increase size enter detailsto come open
rose xu email com	yyyyac idt net	adv harvest lots target email address quickly	want	advertisement persons 18yrs old

Gaming options	Feature subsets for value changes	
1	From	To
	crispin cown crispin wirex com	example com mailing... list secprog securityfocus... moderator
2	From	First Sentence
	crispin cowan crispin wirex com	scott mackenzie wrote
3	From	First Sentence
	tim one comcast net tim peters	tim

Table 5.4: Recourse example comparing causal and non-causal (i.e., feature independent) $\mathcal{D}$. We sample a single input example with a negative prediction and 100 references with the opposite outcome. For I2R$_{causal}$ we propagate the effects of interventions through a user-provided SCM.

input									I2R		I2R$_{causal}$	
Age	Sex	Job	Housing	Savings	Checking	Credit	Duration	Purpose	τ-minimal factors ($\tau = 0$)	Cost	τ-minimal factors ($\tau = 0$)	Cost
23	Male	Skilled	Free	Little	Little	1845	45	Radio/TV	Job: Highly skilled Checking: NA Duration: 30 Age: 65, Housing: Own Age: 34, Savings: N/A	1 1 1.25 4.23 1.84	Age: 24 Sex: Female Job: Highly skilled Housing: Rent Savings: N/A	0.07 1 1 1 1

goal is to answer questions of the form, "What are words with/without which our model would output the original/opposite prediction for an input sentence?" For this experiment, we train an LSTM network on the `IMDB` dataset for sentiment analysis (Maas et al., 2011). If the model mislabels a sample, we investigate further; if it does not, we inspect the most explanatory factors to learn more about model behavior. For the purpose of this example, we only inspect sentences of length 10 or shorter. We provide two examples and compare with Anchors in Table 5.2.

Consider our first example: READ BOOK FORGET MOVIE is a sentence we would expect to receive a negative prediction, but our model classifies it as positive. Since we are investigating a positive prediction, our reference space is conditioned on a negative label. For this model, the standard 'empty' token UNK receives a positive prediction. Thus we opt for an alternative neutral token, PLATE. Performing interventions on all possible combinations of words with our token, we find the conjunction of READ, FORGET, and MOVIE is a sufficient factor for a positive prediction (R2I). We also find that changing any of READ, FORGET, or MOVIE to PLATE would result in a negative

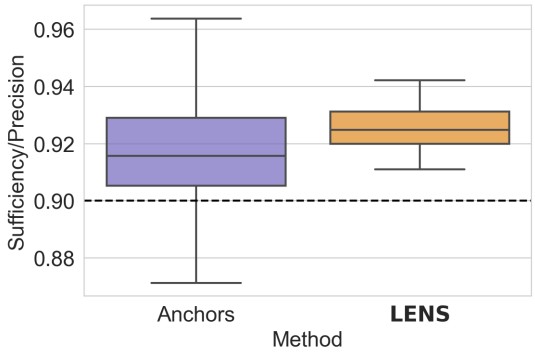


Figure 5.4: We compare $PS(c, y)$ against precision scores attained by the output of LENS and Anchors for examples from `German`. We repeat the experiment for 100 inputs, and each time consider the single example generated by Anchors against the mean $PS(c, y)$ among LENS's candidates. Dotted line indicates $\tau = 0.9$.

prediction (I2R). Anchors, on the other hand, perturbs the data stochastically (see Section 5.6), suggesting the conjunction READ AND BOOK. Next, we investigate the sentence: YOU BETTER CHOOSE PAUL VERHOEVEN EVEN WATCHED. Since the label here is negative, this time we use the 'empty' token UNK. We find that this prediction is brittle – a change of almost any word would be sufficient to flip the outcome. Anchors, on the other hand, reports a conjunction including most words in the sentence. Taking the R2I view, we still find a more concise explanation: CHOOSE or EVEN would be enough to attain a negative prediction. These brief examples illustrate how LENS may be used to find brittle predictions across samples, search for similarities between errors, or test for model reliance on sensitive attributes (e.g., gender pronouns).

Anchors Comparison. Anchors also includes a tabular variant, against which we compare LENS's performance in terms of R2I sufficiency. We present the results of this comparison in Fig. 5.4, and include additional comparisons in Section 5.6. We sample 100 inputs from the `German` dataset, and query both methods with $\tau = 0.9$ using the classifier from Sect. 5.5.1. Anchors satisfies a PAC bound controlled by parameter δ. At the default value $\delta = 0.1$, Anchors fails to meet the τ threshold on 14% of samples; LENS meets it on 100% of samples. This result accords with Thm. 2 and vividly demonstrates the benefits of our optimality guarantee. Note that we also go beyond Anchors in providing multiple explanations instead of just a single output, as well as a cumulative probability measure with no analog in their algorithm.

5.5.3 Counterfactuals

Adversarial Examples: Spam Emails. R2I sufficiency answers questions of the form, "What would be sufficient for the model to predict y?". This is particularly valuable in cases with unfavorable outcomes y'. Inspired by adversarial interpretability approaches (Ribeiro et al., 2018b; Lakkaraju and Bastani, 2020), we train an MLP classifier on the `SpamAssassins` dataset and search for minimal factors sufficient to relabel a sample of spam emails as non-spam. Our examples follow some patterns common to spam emails: received from unusual email addresses, include suspicious keywords such as ENLARGEMENT or ADVERTISEMENT in the subject line, etc. We identify minimal changes that will flip labels to non-spam with high probability. Options include altering the incoming email address to more common domains, and changing the subject or first sentences (see Table 5.3). These results can improve understanding of both a model's behavior and a dataset's properties.

Diverse Counterfactuals. Our explanatory measures can also be used to secure algorithmic recourse. For this experiment, we benchmark against DiCE (Mothilal et al., 2020b), which aims to provide diverse recourse options for any underlying prediction model. We illustrate the differences between our respective approaches on the `Adult` dataset (Kochavi and Becker, 1996), using an MLP and following the procedure from the original DiCE paper.

According to DiCE, a diverse set of counterfactuals is one that differs in *values* assigned to features, and can thus produce a counterfactual set that includes different interventions on the same variables (e.g., CF1: $\mathsf{age} = 91, \mathsf{occupation} =$ "retired"; CF2: $\mathsf{age} = 44, \mathsf{occupation} =$ "teacher"). Instead, we look at the diversity of counterfactuals in terms of intervention *targets*, i.e. features changed (in this case, from input to reference values) and their effects. We present minimal cost interventions that would lead to recourse for each feature set but we summarize the set of paths to recourse via subsets of features changed. Thus, DiCE provides answers of the form "Because you are not 91 and retired" or "Because you are not 44 and a teacher"; we answer "Because of your age and occupation", and present the lowest cost intervention on these features sufficient to flip the prediction.

With this intuition in mind, we compare outputs given by DiCE and LENS for various inputs. For simplicity, we let all features vary independently. We consider two metrics for comparison: (a) the mean cost of proposed factors, and (b) the number of minimally valid candidates proposed, where a factor c from a method M is *minimally valid* iff for all c' proposed by M', $\neg(c' \prec_{cost} c)$ (i.e., M' does not

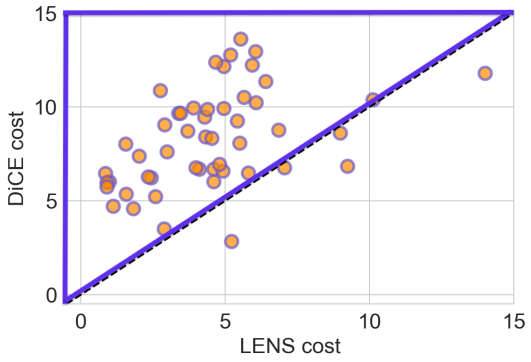


Figure 5.5: A comparison of the mean cost of outputs by LENS and DiCE for 50 inputs sampled from the `Adult` dataset.

report a factor preferable to c). We report results based on 50 randomly sampled inputs from the `Adult` dataset, where references are fixed by conditioning on the opposite prediction. The cost comparison results are shown in Fig. 5.5, where we find that LENS identifies lower cost factors for the vast majority of inputs. Furthermore, DiCE finds no minimally valid candidates that LENS did not already account for. Thus LENS emphasizes *minimality* and *diversity* of intervention targets, while still identifying low-cost intervention values.

Causal vs. Non-causal Recourse. When a user relies on XAI methods to plan interventions on real-world systems, causal relationships between predictors cannot be ignored. In the following example, we consider the DAG in Fig. 5.6, intended to represent dependencies in the `German` credit dataset. For illustrative purposes, we assume access to the structural equations of this data-generating process. (There are various ways to extend our approach using only partial causal knowledge as input (Karimi et al., 2020b; Heskes et al., 2020).) We construct D by sampling from the SCM under a series of different possible interventions. Table 5.4 describes an example of how using our framework with augmented causal knowledge can lead to different recourse options. Computing explanations under the assumption of feature independence results in factors that span a large part of the DAG depicted in Fig. 5.6. However, encoding structural relationships in D, we find that LENS assigns high explanatory value to nodes that appear early in the topological ordering. This is because intervening on a single root factor may result in various downstream changes once the effects are fully propagated.

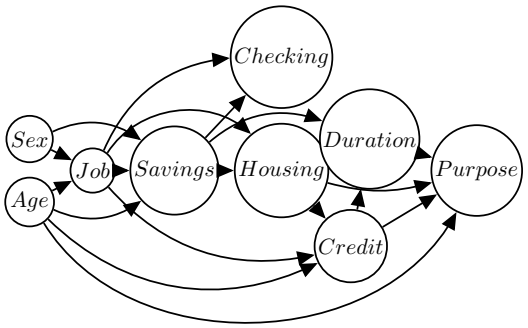


Figure 5.6: Example DAG for `German` dataset.

5.6 Additional Discussion of Experimental Results

5.6.1 Data Pre-Processing and Model Training

German Credit Risk. We first download the dataset from Kaggle,[4] which is a slight modification of the UCI version (Dua and Graff, 2017). We follow the pre-processing steps from a Kaggle tutorial.[5] In particular, we map the categorical string variables in the dataset (`Savings`, `Checking`, `Sex`, `Housing`, `Purpose` and the outcome `Risk`) to numeric encodings, and mean-impute values missing values for `Savings` and `Checking`. We then train an Extra-Tree classifier (Geurts et al., 2006) using pedregosa2011scikit, with random state 0 and max depth 15. All other hyperparameters are left to their default values. The model achieves a 71% accuracy.

German Credit Risk - Causal. We assume a partial ordering over the features in the dataset, as described in Fig. 5.6. We use this DAG to fit a structural causal model (SCM) based on the original data. In particular, we fit linear regressions for every continuous variable and a random forest classifier for every categorical variable. When sampling from $\mathcal{D}$, we let variables remain at their original values unless either (a) they are directly intervened on, or (b) one of their ancestors was intervened on. In the latter case, changes are propagated via the structural equations. We add stochasticity via Gaussian noise for continuous outcomes, with variance given by each model's residual mean squared error. For categorical variables, we perform multinomial sampling over predicted class probabilities. We use the same f model as for the non-causal `German` credit risk description above.

[4]See `https://www.kaggle.com/kabure/german-credit-data-with-risk?select=german_credit_data.csv`.

[5]See `https://www.kaggle.com/vigneshj6/german-credit-data-analysis-python`.

SpamAssassins. The original spam assassins dataset comes in the form of raw, multi-sentence emails captured on the Apache SpamAssassins project, 2003-2015.[6] We segmented the emails to the following "features": `From` is the sender; `To` is the recipient; `Subject` is the email's subject line; `Urls` records any URLs found in the body; `Emails` denotes any email addresses found in the body; `First Sentence`, `Second Sentence`, `Penult Sentence`, and `Last Sentence` refer to the first, second, penultimate, and final sentences of the email, respectively. We use the original outcome label from the dataset (indicated by which folder the different emails were saved to). Once we obtain a dataset in the form above, we continue to pre-process by lower-casing all characters, only keeping words or digits, clearing most punctuation (except for '-' and '_'), and removing stopwords based on nltk's provided list (Bird et al., 2009). Finally, we convert all clean strings to their mean 50-dim GloVe vector representation (Pennington et al., 2014). We train a standard MLP classifier using pedregosa2011scikit, with random state 1, max iteration 300, and all other hyperparameters set to their default values.[7] This model attains an accuracy of 98.3%.

IMDB. We follow the pre-processing and modeling steps taken in a standard tutorial on LSTM training for sentiment prediction with the IMDB dataset.[8] The CSV is included in the repository named above, and can be additionally downloaded from Kaggle or ai.standford.[9] In particular, these include removal of HTML-tags, non-alphabetical characters, and stopwords based on the the list provided in the ntlk package, as well as changing all alphabetical characters to lower-case. We then train a standard LSTM model, with 32 as the embedding dimension and 64 as the dimensionality of the output space of the LSTM layer, and an additional dense layer with output size 1. We use the sigmoid activation function, binary cross-entropy loss, and optimize with Adam (Kingma and Ba, 2015). All other hyperparameters are set to their default values as specified by Keras.[10] The model achieves an accuracy of 87.03%.

Adult Income. We obtain the adult income dataset via DiCE's implementation[11] and followed Haojun Zhu's pre-processing steps.[12] For our recourse comparison, we

[6]See `https://spamassassin.apache.org/old/credits.html`.

[7]See `https://pedregosa2011scikit.org/stable/modules/generated/sklearn.\neural_network.MLPClassifier.html`.

[8]See `https://github.com/hansmichaels/sentiment-analysis-IMDB-Review-using-LSTM/blob/master/sentiment_analysis.py.ipynb`.

[9]See `https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews` or `http://ai.stanford.edu/~amaas/data/sentiment/`.

[10]See `https://keras.io`.

[11]See `https://github.com/interpretml/DiCE`.

[12]See `https://rpubs.com/H_Zhu/235617`.

Table 5.5: Recourse options for a single input given by DiCE and our method. We report targets of interventions as suggested options, but they could correspond to different values of interventions. Our method tends to propose more minimal and diverse intervention targets. Note that all of DiCE's outputs are already subsets of LENS's two top suggestions, and due to τ-minimality LENS is forced to pick the next factors to be non-supersets of the two top rows. This explains the higher cost of LENS's bottom three rows.

input								DiCE output		**LENS output**	
Age	Wrkcls	Edu.	Marital	Occp.	Race	Sex	Hrs/week	Targets of intervention	Cost	Targets of intervention	Cost
42	Govt.	HS-grad	Single	Service	White	Male	40	Age, Edu., Marital, Hrs/week	8.13	Edu.	1
								Age, Edu., Marital, Occp., Sex, Hrs/week	5.866	Martial	1
								Age, Wrkcls, Educ., Marital, Hrs/week	5.36	Occp., Hrs/week	19.3
								Age, Edu., Occp., Hrs/week	3.2	Wrkcls, Occp., Hrs/week	12.6
								Edu., Hrs/week	11.6	Age, Wrkcls, Occp., Hrs/week	12.2

use a pretrained MLP model provided by the authors of DiCE, which is a single layer, non-linear model trained with TensorFlow and stored in their repository as 'adult.h5'.

5.6.2 Tasks

Comparison with attributions. For completeness, we also include here comparison of cumulative attribution scores per cardinality with probabilities of sufficiency for the I2R view (see Fig. 5.7).

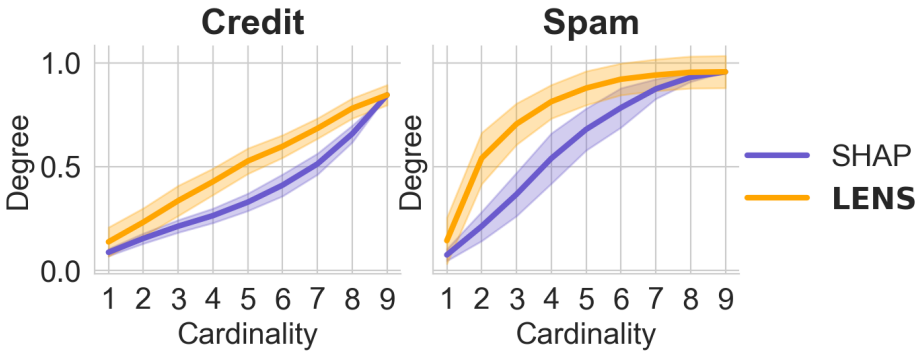


Figure 5.7: Comparison of degrees of sufficiency in I2R setting, for top k features based on SHAP scores, against the best performing subset of cardinality k identified by our method. Results for `German` are averaged over 50 inputs; results for `SpamAssassins` are averaged over 25 inputs.

Sentiment Sensitivity Analysis. We identify sentences in the original `IMDB` dataset that are up to 10 words long. Out of those, for the first example we only look at wrongly predicted sentences to identify a suitable example. For the other example, we simply consider a random example from the 10-word maximum length examples. We noted that Anchors uses stochastic word-level perturbations for this setting. This

leads them to identify explanations of higher cardinality for some sentences, which include elements that are not strictly necessary. In other words, their outputs are not minimal, as required for descriptions of "actual causes" (Halpern and Pearl, 2005a; Halpern, 2016).

Comparison with Anchors. To complete the picture of our comparison with Anchors on the `German` Credit Risk dataset, we provide here additional results. In Section 5.5, we included a comparison of Anchors's single output precision against the mean degree of sufficiency attained by our multiple suggestions per input. We sample 100 different inputs from the `German` Credit dataset and repeat this same comparison. Here we additionally consider the minimum and maximum $PS(c, y)$ attained by LENS against Anchors. Note that even when considering minimum PS suggestions by LENS, i.e. our worst output, the method shows more consistent performance. We qualify this discussion by noting that Anchors may generate results comparable to our own by setting the δ hyperparameter to a lower value. However, Ribeiro et al. (2018a) do not discuss this parameter in detail in either their original article or subsequent notebook guides. They use default settings in their own experiments, and we expect most practitioners will do the same.

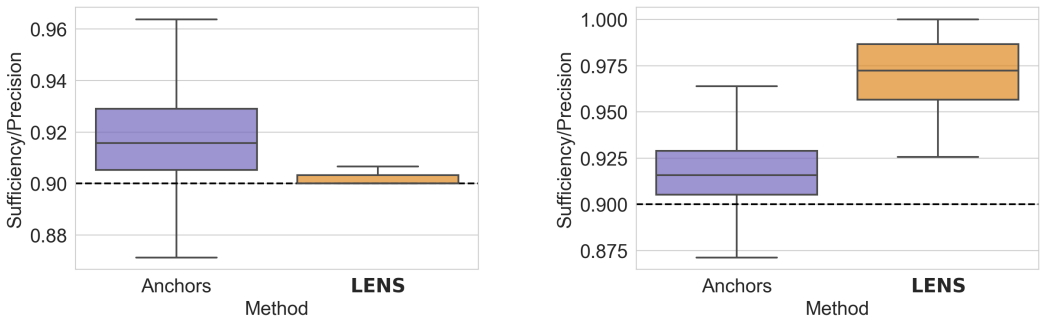


Figure 5.8: We compare degree of sufficiency against precision scores attained by the output of LENS and Anchors for examples from `German`. We repeat the experiment for 100 sampled inputs, and each time consider the single output by Anchors against the min (left) and max (right) $PS(c, y)$ among LENS's multiple candidates. Dotted line indicates $\tau = 0.9$, the threshold we chose for this experiment.

Recourse: DiCE Comparison First, we provide a single illustrative example of the lack of diversity in intervention targets we identify in DiCE's output. Let us consider one example, shown in Table 5.5. While DiCE outputs are diverse in terms of values and target combinations, they tend to have great overlap in intervention targets. For instance, `Age` and `Education` appear in almost all of them. Our method

would focus on minimal paths to recourse that would involve different combinations of features.

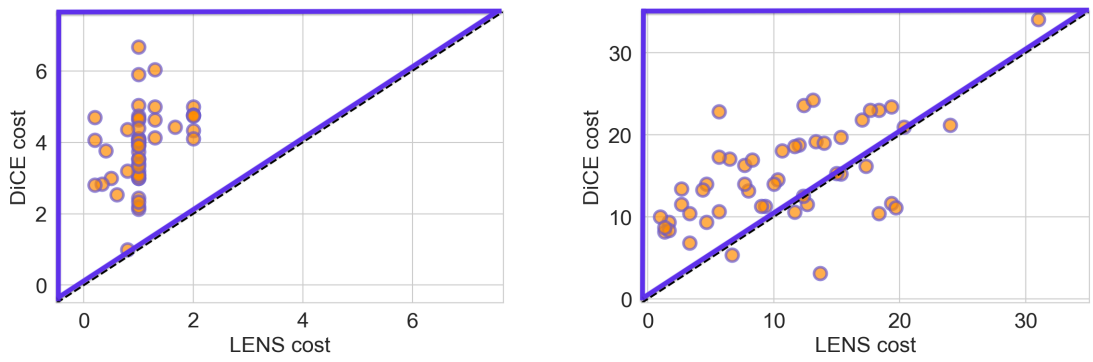


Figure 5.9: We show results over 50 input points sampled from the original dataset, and all possible references of the opposite class, across two metrics: the min cost (left) of counterfactuals suggested by our method vs. DiCE, and the max cost (right) of counterfactuals.

Next, we also provide additional results from our cost comparison with DiCE's output in Fig. 5.8. While in Section 5.5 we include a comparison of our mean cost output against DiCE's, here we additionally include a comparison of min and max cost of the methods' respective outputs. We see that even when considering minimum and maximum cost, our method tends to suggest lower cost recourse options. In particular, note that all of DiCE's outputs are already subsets of LENS's two top suggestions. The higher costs incurred by LENS for the next two lines are a reflection of this fact: due to τ-minimality, LENS is forced to find other interventions that are no longer supersets of options already listed above.

5.7 Conclusion

Our results, both theoretical and empirical, rely on access to the relevant context $\mathcal{D}$ and the complete enumeration of all feature subsets. Neither may be feasible in practice. When elements of $\boldsymbol{Z}$ are estimated, as is the case with the generative methods sometimes used in XAI, modeling errors could lead to suboptimal explanations. For high-dimensional settings such as image classification, LENS cannot be naïvely applied without substantial data pre-processing. The first issue is extremely general. No method is immune to model misspecification, and attempts to recreate a data-generating process must always be handled with care. Empirical sampling, which we rely on above, is a reasonable choice when data are fairly abundant and representative.

However, generative models may be necessary to correct for known biases or sample from low-density regions of the feature space. This comes with a host of challenges that no XAI algorithm alone can easily resolve. The second issue – that a complete enumeration of all variable subsets is often impractical – we consider to be a feature, not a bug. Complex explanations that cite many contributing factors pose *cognitive* as well as computational challenges. In an influential review of XAI, Miller (2019) finds near-unanimous consensus among philosophers and social scientists that, "all things being equal, simpler explanations – those that cite fewer causes... are better explanations" (p. 25). Even if we could list all τ-minimal factors for some very large cardinality of features d, it is not clear that such explanations would be helpful to humans, who famously struggle to hold more than seven objects in short-term memory at any given time (Miller, 1955). That is why many popular XAI tools include some sparsity constraint to encourage simpler outputs. Rather than throw out some or most of our low-level features, we prefer to consider a higher level of abstraction, where explanations are more meaningful to end users. For instance, in our `SpamAssassins` experiments, we started with a pure text example, which can be represented via high-dimensional vectors (e.g., word embeddings). However, we represent the data with just a few intelligible components: `From` and `To` email addresses, `Subject`, etc. In other words, we create a more abstract object and consider each segment as a potential intervention target, i.e. a candidate factor. This effectively compresses a high-dimensional dataset into a 10-dimensional abstraction. Similar strategies could be used in many cases, either through domain knowledge or data-driven clustering and dimensionality reduction techniques. In general, if data cannot be represented by a reasonably low-dimensional, intelligible abstraction, then post-hoc XAI methods are unlikely to be of much help.

We have presented a unified framework for XAI that foregrounds necessity and sufficiency, which we argue are the fundamental building blocks of all successful explanations. We defined simple measures of both and showed how they undergird various XAI methods. Our formulation, which relies on converse rather than inverse probabilities, is uniquely flexible and expressive. It covers all four basic explanatory measures – i.e., the classical definitions and their contrapositive transformations – and unambiguously accommodates logical, probabilistic, and/or causal interpretations, depending on how one constructs the basis tuple $\mathcal{B}$. We illustrated illuminating connections between our measures and existing proposals in XAI, as well as Pearl (2000)'s probabilities of causation. We introduced a sound and complete algorithm for identifying minimally sufficient factors and demonstrated our method on a range of

tasks and datasets. Our approach prioritizes completeness over efficiency, suitable for settings of moderate dimensionality. Future research will explore more scalable approximations, model-specific variants optimized for, e.g., convolutional neural networks, and developing a graphical user interface.

6 | Beyond Impossibility: Revisiting Fairness Trade-offs Between Sufficiency and Separation

6.1 Introduction

The advance of automatic decision-making into social domains has made training and auditing Machine Learning (ML) systems that treat sensitive groups fairly a crucial research area. Fair and Responsible ML has been a growing research area in recent years (Dwork et al., 2012; Barocas et al., 2019), and academic research has fostered important debates around the use of ML for predicting sensitive information, such as for example recidivism in criminal justice systems. In 2016, ProPublica made an important contribution to these debates by analyzing a criminal justice system's score allocation, called COMPAS, and showing that it had twice as high False-Positive Rate (FPR) (44.85 vs. 27.99 on the non-violent COMPAS scores) for black defendants compared to white defendants, while False-Negative Rate (FNR) was twice as high for white defendants compared to black defendants (47.2 vs. 23.45 on the non-violent COMPAS scores) (Larson et al., 2016). Northpointe, the developer of COMPAS, launched a response showing that a different fairness measure, called calibration, was in fact satisfied by the system (Flores et al., 2016). This immediately opened a basic question: what is the *right* fairness criterion for ML systems?

The academic community jumped into the discussion and contributed a much-needed theoretical understanding of these initial results. Multiple researchers showed that the two fairness criteria used in the COMPAS debate, calibration and error rates, were not coincidentally emerging in opposition to each other – when the labels are not equally spread across sensitive groups (as is often the case in criminal justice and other applications), one cannot have both perfect calibration and equal error rates at the same time (unless the predictor is perfect), as proven in a seminal impossibility theorem (Chouldechova, 2017; Kleinberg et al., 2017; Corbett-Davies et al., 2017). The data distribution together with an ML classifier generates a probability distribution on the triple $(\hat{Y}, Y, A)$, where $\hat{Y}$ is the predicted label, Y the true label, and A the protected attribute. The aforementioned fairness measures correspond to two types of conditional

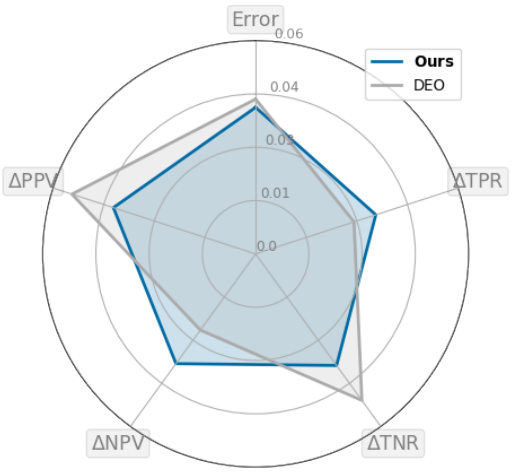


Figure 6.1: We propose, conceptually, that fairness criteria should aim to balance *separation* and *sufficiency* measures. We propose a criterion and show it often leads to better tradeoffs. In the plot above, we show how the method we call **Ours** fares against a Difference of Equality of Opportunity (DEO) constraint Hardt et al. (2016), which only involves one *separation* measure, used in a multi-objective method on the dataset `Color-MNIST` in terms of error (top axis), *separation* violations (TPR, TNR absolute group differences) (right), and *sufficiency* violations (PPV, NPV absolute group differences) (left).

independence among these three quantities: *Separation*, $\hat{Y} \perp\!\!\!\perp A \mid Y$ (related to balance for the positive/negative class, equalized odds, equality of opportunity, conditional procedure accuracy equality, etc.), and *Sufficiency* $Y \perp\!\!\!\perp A \mid \hat{Y}$ (related to calibration within groups, test-fair score, conditional use accuracy equality, etc.).[1]

The impossibility result ruling out equal error rates and perfect calibration simultaneously in practice has had an important and beneficial impact on the development of the fairness literature. In a certain sense, this result puts the COMPAS debate to rest: Northpointe and ProPublica are interested in different incompatible fairness criteria and the choice of Northpointe to focus on calibration can only be questioned outside the realm of mathematical reasoning. Another consequence of the theorem has been that researchers have advocated picking a single arbitrary fairness definition and designing machine learning models that satisfy this fairness definition.[2] Multiple papers thus either 1) recommend practitioners to stick to one group of fairness measures and disregard the others (Pleiss et al., 2017); or 2) suggest that certain results are worse when trying to uphold more than one type of fairness constraint (Padh et al.,

[1]See (Barocas et al., 2019) for a comprehensive survey, and the 'dictionary of criteria' in Chapter 2 for a handy summary.

[2]For example, this one measure can be chosen to be the absolute difference across sensitive groups in one of the possible eight values implied by a standard confusion matrix for a binary classification problem.

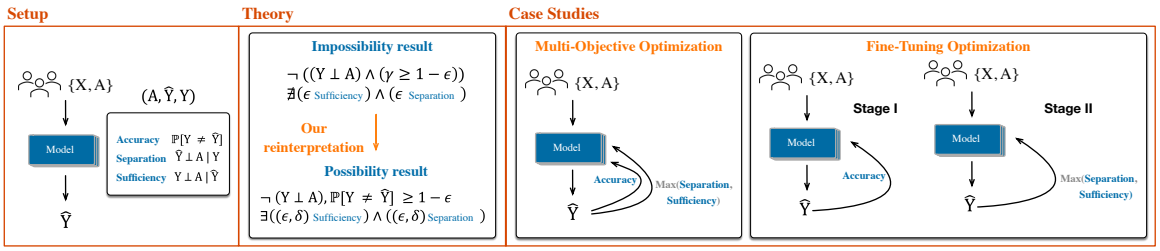


Figure 6.2: A general description of our contributions.

2021), or 3) promote alternative approaches to fairness, partially to bypass this issue (e.g. Counterfactual Fairness (Kusner et al., 2017b)).

Nevertheless, these resolutions are somewhat unsatisfactory: counterfactual fairness, for one, is a very principled way to look at the problem but requires knowledge of the data generation process, and thus may require strong assumptions that make practical implementation more difficult; the other approaches, focusing on one measure of fairness without requiring any guarantees on the others may, for example, allow claims regarding a model's fairness with respect to one (possibly "cherry-picked") criterion while being particularly unfair with respect to several others. Phrased more positively, we ask: given a desired quality level for one fairness criterion (hence seen as a constraint), how well can we optimize the other criteria? We demonstrate that a better balance can be achieved by simultaneously optimizing multiple criteria (cf. Figure 6.1). Concretely, this problem can also be illustrated with the COMPAS debate: acknowledging that the COMPAS system achieves a certain (approximate) calibration as claimed by Northpointe, one may ask how well the system performs in terms of the equal-error-rates measure compared to other systems achieving a similar level of calibration. In other words, we could be satisfied with the COMPAS system if, given the level of calibration achieved, it is the "best" also with respect to equal error rates among all models that have similar accuracy and calibration.

6.1.1 Our Contributions

In this chapter, we introduce an approximate view of the calibration and equal-error-rates criteria that is particularly relevant in practical settings. This leads us to analyze how these quantities can be traded-off and how well can one quantity be approximated under the constraint that the other remains under some desired threshold. We show how to achieve a satisfying balance between approximate versions of these fairness criteria and accuracy as a whole. We focus on the measures $\Delta^{\text{suff}}_{y,\hat{y},a} = |\mathbb{P}[Y = y \mid \hat{Y} = \hat{y}, A = a] - \mathbb{P}[Y = y \mid \hat{Y} = \hat{y}]|$ and $\Delta^{\text{sep}}_{\hat{y},y,a} = |\mathbb{P}[\hat{Y} = \hat{y} \mid Y = y, A = a] - \mathbb{P}[\hat{Y} = \hat{y} \mid Y = y]|$.

We treat $\Delta^{\text{suff}}_{\max} = \max_{y,\hat{y},a} \Delta^{\text{suff}}_{y,\hat{y},a}$ and $\Delta^{\text{sep}}_{\max} = \max_{\hat{y},y,a} \Delta^{\text{sep}}_{\hat{y},y,a}$ as representing deviation from perfect *sufficiency* and *separation* and minimize these quantities.

In Section 6.2 we provide further details on related work. Section 6.3 delves into the impossibility result, and offers a refinement of the theorem in terms of the above criteria: *separation*, *sufficiency* and accuracy of the predictor. Our version of the theorem can be interpreted as an approximate possibility result: if a classifier can hold both *sufficiency* and *separation* approximately (i.e.: under some prescribed threshold), it must be fairly accurate.[3] Stated alternatively, if one has a fairly accurate classifier, it is possible to hold both *sufficiency* and *separation* violations at small values. We therefore argue that we should indeed consider what can be achieved in terms of the above measures beyond the strict impossibility. The implications of our theorem for the COMPAS debate are the following: Given the accuracy and the level of *separation* that are achieved by Northpointe's predictor, our theory prescribes that a certain minimal level of *sufficiency* can be achieved. One may thus ask whether Northpointe's predictor achieves this minimal level of *sufficiency*.

Equipped with this theoretical insight, we design new loss functions for training classifiers that have both good accuracy and minimal fairness requirements for *sufficiency* and *separation* criteria (Sections 6.4 and 6.5). We propose the use of our new loss functions via an in-processing multi-objective framework Padh et al. (2021), as well as using one of them as a post-processing technique, fine-tuning an existing model with a fairness-dedicated objective. We show that our new loss function $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ yields classifiers achieving better trade-offs in terms of accuracy, *separation* and *sufficiency* compared to existing alternatives across the 4 datasets we studied: `COMPAS` (Larson et al., 2016), `Adult Income` (Kohavi, 1996), `NELS` (Ingels, 1990) and `Color-MNIST` (Lecun et al., 1998).[4]

6.2 Background

Constrained Optimization Under Fairness Notions. Multiple works considered forms of constrained optimization of predictors to achieve one of the fairness notions we consider in this work (Zafar et al., 2017a; Wu et al., 2019a; Donini et al., 2018). However, they consider achieving either the Δ^{suff} or the Δ^{sep} family. Previous authors moreover studied the **the fairness and accuracy trade-off via Pareto Frontiers** (Kearns

[3] Assuming base rates are not almost equally distributed across groups (a very unlikely possibility under most fairness applications).

[4] *Color-MNIST* is an adaptation of the original MNIST dataset to this setting, see full description in 6.5.1.

et al., 2018; Kearns and Roth, 2019). Instead, we focus on predictors holding both fairness notions of Δ^{suff} and Δ^{sep} in an approximate fashion and propose optimization methods inspired by this idea. There were also more recent works on **Fairness relaxations** (Lohaus et al., 2020), **Group and subgroup Fairness** (Martinez et al., 2020; Diana et al., 2020; Kearns et al., 2018; Yang et al., 2020) or **fair transformations of scores** (Jiang et al., 2020). While they all raise important and interesting points about how to hold one chosen family of notions in a properly relaxed version that comes with guarantees, or exploring subgroup fairness, our focus is in a way more fundamental, going back to the setting of the original impossibility result. One other set of methods that could enable the balancing of multiple fairness objectives is **Multi-objective approaches to fairness** (Celis et al., 2019; Padh et al., 2021; Ruchte and Grabocka, 2021). While the above works could handle the challenge of balancing *separation* and *sufficiency*, they do not directly address this question, and instead focus their experiments on the combination of other objectives (e.g., Demographic Parity (DP) and Equality of Opportunity (EOP) Hardt et al. (2016)). We believe this might be due to the direct or indirect influence of the **impossibility result**: Kleinberg et al. (2017) and Chouldechova (2017) are each cited over 1,000 times so far. Furthermore, the legacy it has left on the literature goes beyond the original result itself, including Pleiss et al. (2017) which directly recommends picking one of the violation groups to focus on and giving up on the other.

How is the classic impossibility result cited and used in the fairness literature? Pleiss et al. (2017) showed the strongest resistance to an attempt to balance *separation* and *sufficiency* violations, stating multiple times that a practitioner may want to focus on one. "Calibration and error rate constraints are in most cases mutually incompatible goals", Pleiss et al. advised, and proceeded to conclude, "[i]n practical settings, it may be advisable to choose only one of these goals rather than attempting to achieve some relaxed notion of both". Other works used similar language to justify different fairness approaches, including Kusner et al. (2017b) who recommended Counterfactual Fairness as a way to bypass the impossibility result on multiple occasions; Padh et al. (2021) recently discussed performance of their mult-objective approach to fairness in terms of the impossibility result, explaining that "For the COMPAS dataset, [the suggested algorithm] performs well for DEO but not for DDP... it is not possible to satisfy DP and error rate based metrics simultaneously if the base rate of classification is different for different groups... This explains the poor performance of [the suggested algorithm] algorithm on the Dutch dataset as well as the fact that it

performs well only on DEO and not on DDP for the COMPAS dataset". We advocate here both a closer look at the fairness violation resulting from any fairness constraint (see our Figure 6.4 compared to Figure 2 in the original), as well as an additional criterion aiming at achieving a balance between *sufficiency* and *separation* measures, instead of abandoning the hope to find such balance altogether.

We do not know of papers that try to balance both *sufficiency* and *separation* directly. There are **other works that attempt to look beyond the impossibility results** via other means. Lazar Reich and Vijaykumar (2021) show how to provide scores (model output pre-threshold) satisfying calibration, and label predictions (post-threshold) satisfying equal error rates. Blum and Stangl (2019) suggest that when starting with a biased dataset (according to definitions they provide), fairness-motivated constraints can lead to better results even if one cares only about accuracy. Finally, other works looked at **the fairness and accuracy trade-off** (Menon and Williamson, 2018; Chen et al., 2018; Dutta et al., 2020; Wick et al., 2019).

6.3 Theoretical Contribution

6.3.1 Separation and Sufficiency

In the following, let A, Y and $\hat{Y}$ denote the random variable representing the sensitive attribute, the ground truth label, and the outcome of a prediction model. We will denote by a, y and $\hat{y}$ the values taken by these random variables; for simplicity, we will assume that each of these can only take the values $\{0, 1\}$.

Separation refers to the conditional independence of $\hat{Y}$ and A given Y, i.e. $\hat{Y} \perp\!\!\!\perp A|Y$. This, in particular, implies that for any $a, y, \hat{y}$, $\mathbb{P}[\hat{Y} = \hat{y}|Y = y, A = a] = \mathbb{P}[\hat{Y} = \hat{y}|Y = y]$. For an assignment of values y, $\hat{y}$ and a, we denote the *separation* gap by $\Delta^{\text{sep}}_{\hat{y},y,a} = |\mathbb{P}[\hat{Y} = \hat{y}|Y = y, A = a] - \mathbb{P}[\hat{Y} = \hat{y}|Y = y]|$, and we denote by $\Delta^{\text{sep}}_{\max} = \max_{\hat{y},y,a} \Delta^{\text{sep}}_{\hat{y},y,a}$. In particular, *separation* holds if and only if $\Delta^{\text{sep}}_{\max} = 0$. A second criterion called *sufficiency* refers to the conditional independence of Y and A given $\hat{Y}$, i.e. $Y \perp\!\!\!\perp A|\hat{Y}$. As in the case of *separation*, we can define $\Delta^{\text{suff}}_{y,\hat{y},a} = |\mathbb{P}[Y = y|\hat{Y} = \hat{y}, A = a] - \mathbb{P}[Y = y|\hat{Y} = \hat{y}]|$, and denote by $\Delta^{\text{suff}}_{\max} = \max_{y,\hat{y},a} \Delta^{\text{suff}}_{y,\hat{y},a}$. *Sufficiency* holds if and only if $\Delta^{\text{suff}}_{\max} = 0$.

To measure the performance of the classifiers, we can measure the *precision* and *recall* of the positive and negative classes. For the positive class, these are denoted by PPV (positive predictive value) and TPR (true positive rate), whereas for the negative class, these are denoted by NPV (negative predictive value) and TNR (true negative rate).

The complements of these notions are also meaningful. In particular, we have (see Tharwat (2020)):

- $\mathsf{PPV} = \mathbb{P}[Y = 1|\hat{Y} = 1]$, $\mathsf{FDR} = 1 - \mathsf{PPV} = \mathbb{P}[Y = 0|\hat{Y} = 1]$ (false discovery rate)
- $\mathsf{TPR} = \mathbb{P}[\hat{Y} = 1|Y = 1]$, $\mathsf{FNR} = 1 - \mathsf{TPR} = \mathbb{P}[\hat{Y} = 0|Y = 1]$ (false negative rate)
- $\mathsf{NPV} = \mathbb{P}[Y = 0|\hat{Y} = 0]$, $\mathsf{FOR} = \mathbb{P}[Y = 1|\hat{Y} = 0]$ (false omission rate)
- $\mathsf{TNR} = \mathbb{P}[\hat{Y} = 0|Y = 0]$, $\mathsf{FPR} = \mathbb{P}[\hat{Y} = 1|Y = 0]$ (false positive rate)

We also consider these measures conditioned on the sensitive attribute, e.g. $\mathsf{PPV}_a = \mathbb{P}[Y = 1|\hat{Y} = 1, A = a]$; the other measures are defined similarly.

We say that a set of values $\{a_1, \ldots, a_n\}$ are (ε, δ)-approximately equal if for every i, j, we have that $a_i \leq a_j e^{\varepsilon} + \delta$. We say that $(\hat{Y}, Y, A)$ satisfies (ε, δ)-*separation*, if the four sets $\left\{\mathbb{P}\left[\hat{Y} = \hat{y}|Y = y, A = a\right] | a \in \{0, 1\}\right\}$ of two values for each possible setting of $y, \hat{y} \in \{0, 1\}$ are (ε, δ)-approximately equal. Likewise, we say that $(\hat{Y}, Y, A)$ satisfies (ε, δ)-*sufficiency*, if the four sets $\left\{\mathbb{P}\left[Y = y|\hat{Y} = \hat{y}, A = a\right] | a \in \{0, 1\}\right\}$ of two values for each possible setting of $y, \hat{y} \in \{0, 1\}$ are (ε, δ)-approximately equal. We denote by $\rho = \mathbb{P}[Y = 1]/\mathbb{P}[Y = 0]$ and by $\rho_a = \mathbb{P}[Y = 1|A = a]/\mathbb{P}[Y = 0|A = a]$. Note that division by 0 will not occur due to the assumptions, as stated in Thm. 1. We refer to ρ (resp. ρ_a) as the *base odds* (resp. group conditional base odds).

6.3.2 Theoretical Result: Refining the Impossibility Result

As mentioned in the introduction, several authors have shown impossibility results for the *separation* and *sufficiency* criteria. Kleinberg et al. further proved an approximate version of the impossibility result, but at the heart of it was showing that the following quantity $\gamma = \mathsf{TPR} \cdot \mathsf{PPV} + \mathsf{FNR} \cdot \mathsf{FOR}$ is close to 1 whenever the base rates differ substantially and approximate versions of both calibration and *sufficiency* hold. Moreover, γ is 1 if and only if the learned classifier is perfect ($Y = \hat{Y}$) or perfectly flipped ($Y = 1 - \hat{Y}$). However, the interpretation is more difficult in the approximate case.

How does Kleinberg et al.'s γ differ from our notion of accuracy? A concrete example of the difference between our result and Kleinberg et al. (2017) is the following. Consider a classifier achieving the following performances: number of true negatives = 1, number of false positives = 1, number of false negatives = N, number of true

positives = 1, then γ is very close to 1 (essentially larger than $(\frac{N}{N+1})^2$), while PPV = 0.5, leading to a large gap between γ and the actual performance of the classifier. On the other hand, one can come up with an example where the inverse phenomenon happens: γ is close to 0.5 while PPV is close to 1. This is arguably equally bad: It is precisely a phenomenon that we want to prevent in e.g., a criminal justice setting (such as described by COMPAS). Motivated by practice, we would like to show that the accuracy is close to 1 whenever the base rates differ substantially and approximate versions of both calibration and *sufficiency* hold, meaning that a predictor with some level of accuracy could satisfy simultaneously some minimal approximate calibration and *sufficiency*. We thus present the following theorem.

Theorem 1. Assume that for some fixed constant $c \in (0, 1/2)$, the random variable $(\hat{Y}, Y, A)$ satisfies for all $a, y, \hat{y} \in \{0, 1\}$, $\mathbb{P}[Y = y|A = a], \mathbb{P}[\hat{Y} = \hat{y}|A = a] \in (c, 1 - c)$. Let $\varepsilon > 0$ be sufficiently small, in particular $\varepsilon < c$ and $0 < \delta = o(\varepsilon)$. If $(\hat{Y}, Y, A)$ satisfies both (ε, δ)-*sufficiency* and *separation*, then at least one of the following holds:

1. For each a and $\mu \in \{\mathsf{PPV}, \mathsf{NPV}, \mathsf{TPR}, \mathsf{TNR}\}$, we have $\mu_a \geq 1 - O(\varepsilon + \delta/\varepsilon)$,
2. For each a and $\mu \in \{\mathsf{PPV}, \mathsf{NPV}, \mathsf{TPR}, \mathsf{TNR}\}$, we have $\mu_a \leq O(\varepsilon + \delta/\varepsilon)$ (flipped classifiers),
3. The set $\{\rho_0, \rho_1\}$ is $(O(\varepsilon + \delta/\varepsilon), 0)$-approximately equal, where $\rho_a = \mathbb{P}[Y = 1|A = a]/\mathbb{P}[Y = 0|A = a]$.

Proof. Throughout the proof, we assume that $c_1, c_2, \ldots$ are sufficiently small constants. We will prove the statements for TPR and PPV; the statements regarding $\mathsf{NPV}, \mathsf{TNR}$ hold similarly. Without loss of generality assume that $\mathsf{FPR}_0 \leq \mathsf{FPR}_1$. We have,

$$\mathsf{FPR}_0 = \frac{\mathbb{P}[\hat{Y} = 1, Y = 0|A = 0]}{\mathbb{P}[Y = 0|A = 0]}.$$

We break the proof into several cases.

Case 1: $\mathsf{FPR}_0 \leq c_1\varepsilon$. In this case, $\mathsf{FPR}_1 \leq c_1\varepsilon e^{\varepsilon} + \delta$. For $a \in \{0, 1\}$, we have that

$$\mathsf{PPV}_a = \frac{\mathbb{P}[\hat{Y} = 1, Y = 1|A = a]}{\mathbb{P}[\hat{Y} = 1|A = a]} = 1 - \frac{\mathbb{P}[\hat{Y} = 1, Y = 0|A = a]}{\mathbb{P}[\hat{Y} = 1|A = a]} = 1 - \frac{\mathbb{P}[Y = 0|A = a]}{\mathbb{P}[\hat{Y} = 1|A = a]}\mathsf{FPR}_a.$$

From there, using the assumption that $\mathbb{P}[Y = 0|A = a], \mathbb{P}[\hat{Y} = 1|A = a] \in (c, 1 - c)$, and the facts that $e^{\varepsilon} = \Theta(1)$ and $\delta = O(\delta/\varepsilon)$, it follows that $\mathsf{PPV}_a \geq 1 - O(\varepsilon + \delta/\varepsilon)$.

Since $\mathsf{FPR}_0 \leq c_1\varepsilon$, we have that $\mathbb{P}[\hat{Y}=1, Y=0|A=0] \leq (1-c)\cdot c_1\varepsilon$. Then, provided c_1 is small enough, e.g. $c_1 \leq c/2$, we have $\mathbb{P}[\hat{Y}=0, Y=0|A] \geq c/2$. we have, by Bayes' rule,

$$\begin{aligned}
1-\mathsf{TPR}_a = \mathsf{FNR}_a &= \frac{\mathbb{P}[\hat{Y}=0, Y=1|A=a]}{\mathbb{P}[Y=1|A=a]} \\
&= \frac{\mathbb{P}[Y=1|\hat{Y}=0, A=a]\cdot\mathbb{P}[\hat{Y}=0|A=a]}{\rho_a\cdot\mathbb{P}[Y=0|A=a]} \\
&= \frac{\mathbb{P}[Y=1|\hat{Y}=0, A=a]}{\rho_a\cdot\mathbb{P}[Y=0|A=a]}\cdot\frac{\mathbb{P}[\hat{Y}=0|Y=0, A=a]\cdot\mathbb{P}[Y=0|A=a]}{\mathbb{P}[Y=0|\hat{Y}=0, A=a]} \\
&= \mathbb{P}[\hat{Y}=0|Y=0, A=a]\cdot\frac{1}{\rho_a}\cdot\frac{\mathbb{P}[Y=1|\hat{Y}=0, A=a]}{\mathbb{P}[Y=0|\hat{Y}=0, A=a]} \\
&= \mathsf{TNR}_a\cdot\frac{1}{\rho_a}\cdot\frac{\mathsf{FOR}_a}{\mathsf{NPV}_a}.
\end{aligned}$$

The above can be rewritten slightly to get,

$$\rho_a = \frac{\mathsf{TNR}_a}{\mathsf{FNR}_a}\cdot\frac{\mathsf{FOR}_a}{\mathsf{NPV}_a}.$$

If $\mathsf{FNR}_a \leq c_2\varepsilon$, for either $a=0$ or $a=1$, then it clearly follows that $\mathsf{TPR}_a \geq 1-O(\varepsilon+\delta/\varepsilon)$ for $a\in\{0,1\}$. Otherwise, it must be that case that $\mathbb{P}[\hat{Y}=0, Y=1|A=a] \geq c_2c\varepsilon$. Together with $\mathbb{P}[\hat{Y}=0, Y=0|A=a] \geq c/2$, this means that each of TNR_a, FNR_a, FOR_a, and NPV_a are at least $c_3\varepsilon$ for some constant c_3. Since $(\hat{Y}, Y, A)$ satisfies both (ε,δ)-*sufficiency* and *separation* we get for every $q\in\{\mathsf{TNR}, \mathsf{FOR}, \mathsf{FNR}, \mathsf{NPV}\}$ and $a'\neq a, a'\in\{0,1\}$ we have that $\{q_a, q_{a'}\}$ are (ε,δ)-approximately equal. Hence, by Claim 1,

$$e^{-(\varepsilon+(\delta/(c\varepsilon)))} \leq q_a/q_{a'} \leq e^{(\varepsilon+(\delta/(c\varepsilon)))}.$$

Therefore, using the claim below it follows that $\{\rho_0, \rho_1\}$ are $(O(\varepsilon+\delta/\varepsilon), 0)$-approximately equal.

Claim 1. If $\{v_1, v_2\}$ are (ε,δ)-approximately equal and $\min\{v_1, v_2\} \geq c\varepsilon$, we have $e^{-(\varepsilon+(\delta/(c\varepsilon)))} \leq v_1/v_2 \leq e^{(\varepsilon+(\delta/(c\varepsilon)))}$.

Claim proof. Without loss of generality, let $v_1 \leq v_2$. Then we have,

$$v_1 \leq v_2 \leq v_1e^{\varepsilon}+\delta \leq v_1e^{\varepsilon}+v_1\cdot\frac{\delta}{c\varepsilon} \leq v_1\cdot e^{\varepsilon+\frac{\delta}{c\varepsilon}},$$

where we used that $1+x/e^{\varepsilon} \leq 1+x \leq e^x$.

Case 2a: Assume that $\mathsf{FPR}_0 \geq c_1\varepsilon$; further assume that $\mathbb{P}[\hat{Y}=1, Y=1|A=0] \leq c_4\varepsilon$. The latter assumption implies that $\mathsf{PPV}_0 \leq c_5\varepsilon$ and $\mathsf{TPR}_0 \leq c_6\varepsilon$. Using

the fact that $\{\mathsf{PPV}_0, \mathsf{PPV}_1\}$ are (ε, δ)-approximately equal and $\{\mathsf{TPR}_0, \mathsf{TPR}_1\}$ are (ε, δ)-approximately equal, we get the required result.

Case 2b: Assume that $\mathsf{FPR}_0 \geq c_1\varepsilon$; further assume that $\mathbb{P}[\hat{Y} = 1, Y = 1|A = 0] \geq c_4\varepsilon$. Then, as in Case 1, we can write,

$$\begin{aligned}\mathsf{FPR}_a &= \frac{\mathbb{P}[\hat{Y} = 1, Y = 0|A = a]}{\mathbb{P}[Y = 0|A = a]} \\ &= \mathbb{P}[\hat{Y} = 1|Y = 1, A = a] \cdot \rho_a \cdot \frac{\mathbb{P}[Y = 0|\hat{Y} = 1, A = a]}{\mathbb{P}[Y = 1|\hat{Y} = 1, A = a]} \\ &= \mathsf{TPR}_a \cdot \rho_a \cdot \frac{\mathsf{FDR}_a}{\mathsf{PPV}_a}.\end{aligned}$$

From the above, we get,

$$\rho_a = \frac{\mathsf{FPR}_a}{\mathsf{TPR}_a} \cdot \frac{\mathsf{PPV}_a}{\mathsf{FDR}_a}.$$

Under the conditions, each of $\mathsf{TPR}_a, \mathsf{FPR}_a, \mathsf{PPV}_a, \mathsf{FDR}_a$ is at least some $c_7\varepsilon$, which using the claim above implies that $\{\rho_0, \rho_1\}$ are $(O(\varepsilon + \delta/\varepsilon), 0)$-approximately equal. □

We note that the second property is unavoidable. A classifier with accuracy 0, i.e. $\hat{Y} = 1 - Y$ giving a flipped classifier satisfies *separation* and *sufficiency*. The following straightforward observation also allows us to re-frame the result in a more standard notion of accuracy, i.e. $\mathbb{P}[\hat{Y} \neq Y]$.

Observation 1. Assume that for all a, $\mathsf{TPR}_a \geq 1 - \varepsilon$ and $\mathsf{TNR}_a \geq 1 - \varepsilon$, then the accuracy $\mathrm{acc} = \mathbb{P}[\hat{Y} = Y] \geq 1 - \varepsilon$. Likewise, if for all a, $\mathsf{TPR}_a \leq \varepsilon$ and $\mathsf{TNR}_a \leq \varepsilon$, then the accuracy $\mathrm{acc} \leq \varepsilon$.

Proof. We have

$$\begin{aligned}\mathrm{acc} = \mathbb{P}[\hat{Y} = Y] &= \sum_a \Big(\mathbb{P}[\hat{Y} = 0, Y = 0, A = a] + \mathbb{P}[\hat{Y} = 1, Y = 1, A = a]\Big) \\ &= \sum_a \left(\mathsf{TNR}_a \cdot \mathbb{P}[Y = 0, A = a] + \mathsf{TPR}_a \cdot \mathbb{P}[Y = 1, A = a]\right) \geq 1 - \varepsilon.\end{aligned}$$

The proof of the second part follows almost identically. □

The proof of Theorem 1 above includes careful analysis, but is fundamentally based on simple relationships between the measures. Assuming that the quantities under consideration are well defined, i.e. no division by 0 occurs, one can easily establish relationships between the measures defined above using Bayes' rule. For e.g., as shown in Chouldechova (2017), we get,

$$\mathsf{FPR}_a = \rho_a \cdot \frac{\mathsf{FDR}_a}{\mathsf{PPV}_a} \cdot \mathsf{TPR}_a.$$

If *separation* and *sufficiency* were to hold exactly, all of these quantities are the same regardless of the group a. The only way this can happen is if the base odds, ρ_a are the same or $\mathsf{FPR}_a = 0$ for all a. Similar observations about other quantities show that we must have a perfect classifier. These equations as written down do not hold if some quantities, e.g. PPV_a are 0. A careful analysis shows that that actually happens when we have a perfectly flipped classifier, i.e. $\hat{Y} = 1 - Y$.

6.4 Methods

Following our theoretical result, a natural next step would be to consider an optimization procedure for classifiers that takes advantage of this rediscovered relationship between Δ violations and accuracy, as viewed through $\mathsf{PPV}, \mathsf{NPV}, \mathsf{TPR}$ and TNR. Thus, we suggest to simply minimize the upper bound of Δ^{suff} and Δ^{sep}, as a possible regularizer added to classification accuracy, an objective in a multi-objective framework, or a finetuning objective:

$$\mathcal{L}(\theta, Y, A) = \max(\Delta^{\text{suff}}, \Delta^{\text{sep}}) \tag{6.1}$$

$\hat{Y}$ for the purpose of $\Delta^{notion}_{c,a}$ is defined as $S = \theta^T X > \tau$, and $\tau = 0.5$ is our default value for classification thresholding.

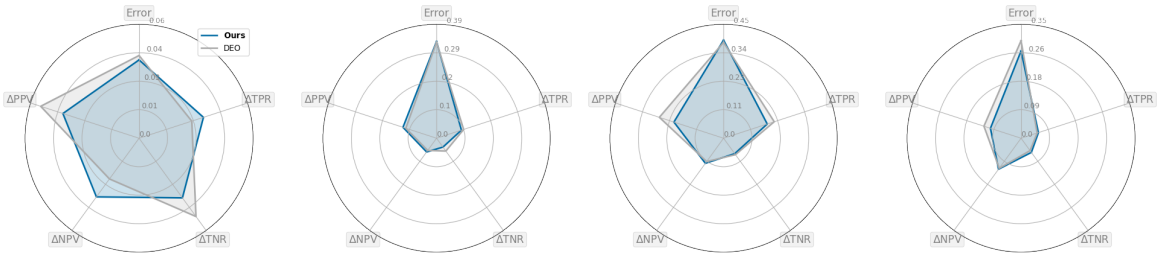
Given that we study the binary group case, in the following we implemented the objective $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ as the max between $\Delta^{\text{suff}}_{y,\hat{y}} = |\mathbb{P}[Y = y|\hat{Y} = \hat{y}, A = 1] - \mathbb{P}[Y = y|\hat{Y} = \hat{y}, A = 0]|$ and $\Delta^{\text{sep}}_{y,\hat{y}} = |\mathbb{P}[\hat{Y} = \hat{y}|Y = y, A = 1] - \mathbb{P}[\hat{Y} = \hat{y}|Y = y, A = 0]|$. We name this modification $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$. Alternating between the two formulations of the objectives leads only to slight differences, as can be seen in the results included in Section 6.7.

Using $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ as part of a multi-objective approach. The balancing act we aim to achieve is compatible with many works published in recent years supporting fairness constraints, such as Multi-Objective techniques. Although multiple such methods were suggested recently Celis et al. (2019); Ruchte and Grabocka (2021), we chose to focus on the work by Padh et al.. MAMO, the algorithm developed in Padh et al. (2021) is model-agnostic, proposes a novel hyperbolic tangent-based relaxation to fairness criteria, and is accompanied by an easy-to-use implementation that is modular and allows for adaptation and extensions. We test the performance of our $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ as an objective to be optimized with MAMO.

Finetuning approach. We also consider $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ as an addition to the common binary cross-entropy (BCE) optimization, making it a finetuning objective. It

Dataset	Number of samples	Training set size	Validation/Test set size	P(Y=0), P(Y=1)	P(A=0), P(A=1)	P(Y=1 | A=0), P(Y=1 | A=1)	γ_0, γ_1
`Color-MNIST`	60,000	40,000	20,000	0.49, 0.51	0.51, 0.49	0.69, 0.29	2.25, 0.41
`COMPAS`	6,172	4,115	2,057	0.53, 0.47	0.6, 0.4	0.53, 0.39	1.1, 0.64
`NELS`	4,743	3,162	1,581	0.533, 0.467	0.099, 0.901	0.311, 0.484	0.45, 0.94
`Adult Income`	48,842	32,562	16,280	0.76, 0.239	0.67, 0.33	0.3, 0.11	0.12, 0.44

Table 6.1: Description of datasets.



(a) `Color-MNIST` (b) `COMPAS` (c) `NELS` (d) `Adult Income`

Figure 6.3: Error and fairness violations in terms of PPV, NPV, TNR and TPR absolute difference across groups for DEO Hardt et al. (2016) vs. $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ (indicated as **Ours** in the legend above). Values plotted are the mean of each metric for 40 runs of all datasets, except for `Color-MNIST`, which is averaged over 25 runs. We look for a centered shape and smaller area where applicable.

can easily follow common model training procedures, saving in costs and implementation difficulties, effectively being a post-processing step of trained models. One point of inspiration for this approach is provided by existing work that shows unconstrained accuracy-focused optimization of predictors tends to minimize Δ^{suff} as a natural by-product (Liu et al., 2019). Therefore, we propose to finetune a predictor trained with BCE, which is expected to already minimize Δ^{suff}, by forcing the upper bound over $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ to be lower. Crucially, it can act as a safeguard against too sharp of an increase in one Δ violation in response to applying a fairness constraint, e.g. just minimizing Δ^{sep} may lead to a high increase in Δ^{suff} and we would like to upper bound such increases.

6.5 Experiments

All experiments were completed on a MacBook Pro 2019 using the Pytorch library for Python, and could be easily performed on any modern CPU configuration, aside for the `Color-MNIST` experiments which involved training a Convolutional Neural Network on a GPU. In all the following plots, Δs defined as, e.g. TPR $= |\mathbb{P}[\hat{Y} = 1|Y = 1, A = 1] - \mathbb{P}[\hat{Y} = 1|Y = 1, A = 0]|$).

6.5.1 Datasets

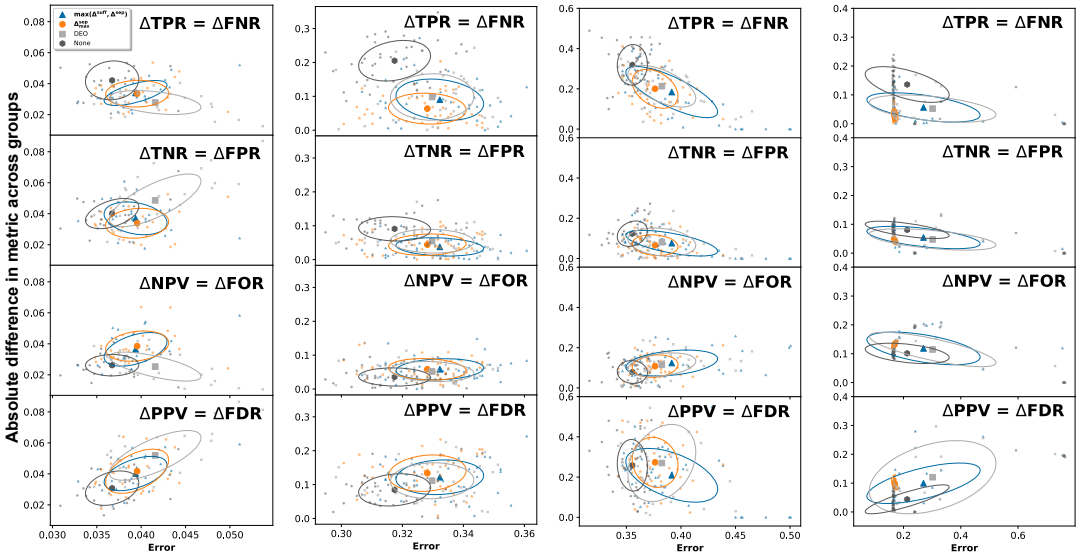
See descriptive statistics in Table 6.1. All are available under creative commons license CC0. We elaborate on `Color-MNIST` as it is the only dataset that we curate for fair ML experiments. We defer the exact details of the rest to Section 6.6 and make datasets and processing code available.

Color-MNIST. The first dataset we considered is a tweaked version of the original MNIST dataset Lecun et al. (1998). In order to clearly demonstrate the dynamics the theory points to, this setting enables us to study a quality predictor that can achieve high accuracy, and thus also small fairness violations. At the same time, it also involves a larger dataset of the more challenging image modality, compared to the often tabular data studied in fairness contexts, and is thus closer to certain real-world applications. As a convolutional neural net is known to achieve near-prefect accuracy on the original dataset, we know we can achieve such statistics on the tweaked dataset.

We create a version of the original dataset where images are colored based on their label. First, we binarize the label, to make all images labeled as $0-4$ the negative class, and all others, $5-9$ the positive class. Next, all images are assigned a color such that, initially, the images in the positive class are assigned a red color, and all images in the negative class are assigned a green color. However, with probability 0.3, we flip the color assigned to each image. We use the assigned colors as the sensitive attribute, and by the procedure above we end up with a dataset that includes different base rates for each label-color combination, such that each sensitive group is more associated with a different label (see Table 6.1). Similar procedures were described in Arjovsky et al. (2020); Wang et al. (2020) for different purposes.

6.5.2 Results Multi-Objective

As one can see in Figures 6.3 and 6.4, $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ can help achieve a better balance across all Δ violations compared to DEO as well as other possible constraints or an unconstrained approach. We directly compare in this case to the setting and results presented in Padh et al. (2021) (See Figure 2 in their manuscript. We change their error bars to ellipses, visualizing how those errors are correlated). We adapt their code and keep most of their original hyperparameter choice for `COMPAS` and `Adult Income`. We average results over 40 runs and set the regularizer coefficient λ to 0.1 and 0.3 respectively. For `Color-MNIST` we use 10 training epochs and set the hyperparameter λ's value to 2. In all settings, we achieve lower violations in at least 2 out of the 4 possible violations, without trading much accuracy or much



(a) `Color-MNIST` (b) `COMPAS` (c) `NELS` (d) `Adult Income`

Figure 6.4: Multi-Objective results, using DEO, $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ or $\Delta^{\text{sep}}_{\max}$ as an additional objective to Binary-Cross Entropy, implemented via Padh et al.'s framework, MAMO. Location of shapes in the middle of ellipses is the mean of 40 runs (25 runs for `Color-MNIST`). The first two rows correspond to *separation* measures; the bottom 2 to *sufficiency* measures. 'None' refers to the unconstrained case (training with Binary Cross-Entropy loss). Ellipses correspond to 1-std. of the distribution of results over runs. y-axis: absolute difference of metric across groups; x-axis: error (1-accuracy).

increase in the other measures. We sacrifice relatively little in accuracy compared to the unconstrained case; we also report better or equivalent accuracy compared to DEO. We further experimented with a variety of alternatives but found $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ to lead to better trade-offs. We present results for alternative criteria in Section 6.7, and include below the 3 most promising constraints: $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ $\Delta^{\text{sep}}_{\max}$ and DEO.

Color-MNIST. DEO vs. $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$. As expected, DEO achieves lower TPR/FNR[5] and NPV/FOR values (by about 0.01 points), but at the cost of higher PPV/FDR values (*sufficiency*) (0.012 points difference, and 0.02 compared to the unconstrained model's violation) as well as TNR/FPR values (*separation*) (0.0116 difference from $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$). In other words, DEO achieves high fluctuations of two pairs of violations. On the other hand, $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ and $\Delta^{\text{sep}}_{\max}$ keep all 4 violations within a similar range, between 0.033 and 0.04, thus balancing the violations

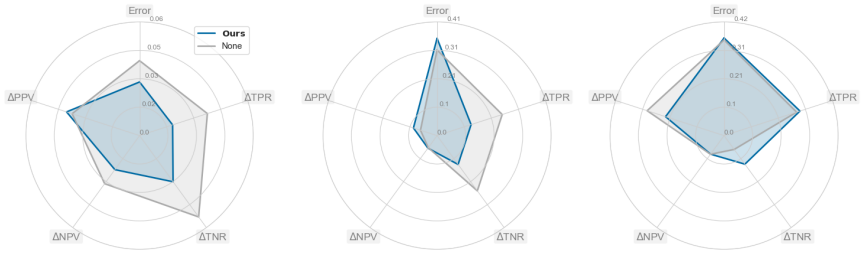
[5]The DEO constraint directly involves the absolute difference between TPR/FNR for both groups as indicated by the sensitive attribute. In the following, whenever we refer to the constraint itself we will call it DEO, but when referring to the measured violation in TPR/FNR we will simply refer to TPR/FNR.

across all measures and keeping accuracy rather close to the unconstrained model training (0.037 vs. 0.039). $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ vs. $\Delta^{\text{sep}}_{\max}$. $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ and $\Delta^{\text{sep}}_{\max}$ stay close together in this dataset (up to 0.003 difference). However, notice how $\Delta^{\text{sep}}_{\max}$ leads to slightly lower *separation* values in this case, while $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ achieves slightly lower *sufficiency* values. Thus we see $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ achieving a better balance across the 4 violation pairs. The variance across runs is much higher for the DEO objective. $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ and $\Delta^{\text{sep}}_{\max}$ are within the same range.

COMPAS. DEO vs. $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$. Notice that TPR/FNR is the biggest original violation for the unconstrained model (~ 0.2). Thus, DEO and $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ achieve similar results for that criterion (0.098 vs. 0.09). However, $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ additionally improves group-difference in TNR/FPR (0.054 vs. 0.037), without much sacrifice of increase of group-differences in PPV/FDR (0.11 vs. 0.12). Considering Figure 6.4 more closely, we see $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ offers lower violation for all *separation* measures (top two rows), while not trading off much of the *sufficiency* violations. $\Delta^{\text{sep}}_{\max}$ vs. $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ Compared to $\Delta^{\text{sep}}_{\max}$, we see similar *separation* values to $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ (TPR/FNR, TNR/FPR). Yet, $\Delta^{\text{sep}}_{\max}$ achieves smaller TPR/FNR values compared to $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ (0.06 vs. 0.089), but does so at the prices of slightly higher TNR/FPR violation (0.044 vs. 0.037), and more importantly higher violations in the *sufficiency* violations PPV/FDR (0.134 vs. 0.122). The variance across runs is overall similar for the 3 constraint methods. Thus, we see $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ as achieving a better balance overall, although $\Delta^{\text{sep}}_{\max}$ could be a reasonable alternative for specific use cases.

NELS. DEO vs. $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$. Both approaches attain similar results for TNR/FPR (0.081 vs. 0.076) and NPV/FOR (0.115 vs. 0.124), however, there is a slight advantage for $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ for TPR/FNR and especially for PPV/FDR (0.27 vs. 0.209). $\Delta^{\text{sep}}_{\max}$ vs. $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$. The two objectives show similar results for TNR/FPR and NPV/FOR with slightly smaller values for $\Delta^{\text{sep}}_{\max}$. Nevertheless, there is an even greater advantage for $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ for TPR/FNR and especially for PPV/FDR (0.273 vs. 0.209). Variance across runs is quite high for $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ for this setting, yet seems to be dominated by 3 runs out of 40 (which dominate the scale of the plot as a whole), representing runs that are both high error and particularly high PPV/FDR violations.

Adult Income. DEO and $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$. *Separation* values in this case remain quite close. As for *sufficiency*, DEO achieves slightly lower NPV/FOR (0.114 vs. 0.118); notice, however, that $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ achieves far lower PPV/FDR violation (0.12 vs. 0.099), and that's with an even lower variance in performance.



(a) `Color-MNIST` (b) `COMPAS` (c) `NELS`

Figure 6.5: Finetune experiment with $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ objective. Error and fairness violations in terms of PPV, NPV, TNR, and TPR absolute difference across groups for BCE vs. $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ (indicated as **Ours** in the legend above). We look for a greater 'balance' of shape around the center, and smaller area where applicable.

$\Delta^{\text{sep}}_{\max}$ and $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$. It is rather clear $\Delta^{\text{sep}}_{\max}$ leads to lower *separation* values, while $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ achieves one far lower *sufficiency* violation. The accuracy achieved by $\Delta^{\text{sep}}_{\max}$ is higher here, and so $\Delta^{\text{sep}}_{\max}$ might be a better choice for this dataset overall. Variance across runs. For $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ and DEO the variance is quite large for the `Adult Income` dataset, with a slightly higher variance for DEO. $\Delta^{\text{sep}}_{\max}$ achieves remarkably low variance, and also a high accuracy that is very competitive with the unconstrained case. Thus, it could be an advantage for this criterion, depending on the specifics of the use case. However, two things are worth noting: the variance of $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ and DEO seem to be driven by a few runs that achieve particularly high error. The `Adult Income` dataset is highly imbalanced in terms of label distribution (the ratio of negative to positive labels is 3.16, see Table 6.1), which might explain why it can be vulnerable to different random splits of the dataset into train, validation, and test.

6.5.3 Results Finetuning

As a second possibility of putting our objective to the test, we consider using $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ as part of a finetuning approach, following standard Binary Cross Entropy training. We test the performance of such an objective on multiple datasets as we did above (see main results achieved on the test set in Figure 6.5). Notice how $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ leads to improved mean violations of a greater scale than a small increase in error, or even improvement for the `Color-MNIST` dataset. These results are overall in line with the performance of the Multi-Objective approach of the previous section. Given the performance of $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ above, we focus on it as the objective used in this finetuning setting and compare its results to

the baseline of training a predictor via BCE throughout training. Similarly, given the high variance and imbalanced nature of the `Adult Income` dataset, we did not include it in the following. **COMPAS.** We applied the method discussed in Section 6.4 to the `COMPAS` dataset to train a predictor. In the previous experiment, we mostly adopted the hyper-parameter settings used in Padh et al. (2021) for ease of comparison. In this setting, we performed our own hyperparameter tuning, as described in the Appendix. We see similar training dynamics in terms of accuracy and $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ on both training and validation sets (the indicated validation results are the mean over the 3 folds and the 5 random initializations, see Figure 6.9 in the Appendix). We can see some trade-off between high accuracy and low $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$, yet `finetune` remains well above 60% accuracy while achieving a decrease of about 40% in $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ on the train/validation splits. Inspecting $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ and accuracy results over the test set in Figure 6.5, we see that our proposed method is competitive with BCE. It presents the better-balanced violations achieved by $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$, captured by the more centered and smaller area it covers in the radar plot compared to the unconstrained BCE, while the top edge representing the error remains quite close.

6.6 Additional Experimental Details

6.6.1 Datasets

COMPAS. A natural interesting choice for a dataset to investigate this approach is the one that ignited the debate to begin with: the `COMPAS` dataset as studied by ProPublica (Larson et al., 2016)[6]. We apply minimal pre-processing, and finally obtain a dataset of 9 features, including JAIL IN, JAIL OUT, AGE, PRIOR COUNTS, DAYS B SCREENING ARREST, CHARGE DEGREE, RACE, AGE CATEGORY and SEX. We use RACE as the protected attribute A, and TWO YEAR RECIDIVISM as the target Y. For the purpose of this analysis, we only consider samples where A was either AFRICAN-AMERICAN or CAUCASIAN. The dataset is overall balanced in terms of group sizes, but with unequal base rates, as can be seen in Table 6.1.

NELS. As an additional dataset we consider the `NELS` dataset, or the US Department of Education's National Education Longitudinal Study of 1988 (Ingels, 1990) as well as its followups. The dataset was used in (Kleinberg et al., 2018), and we follow the

[6]We use the COMPAS SCORES TWO YEARS dataset from `https://github.com/propublica/compas-analysis`.

pre-processing steps taken by the authors in the accompanying code[7]. It consists of 427 features, including the protected attribute RACE, as well as HIGH-SCHOOL GRADES, COURSE TAKING PATTERNS, EXTRACURRICULARS, and STANDARDIZED TESTS IN MATH, READING, SCIENCE, SOCIAL STUDIES. The target for prediction is GPA. The only change we make to Kleinberg et al.'s procedure is setting the threshold for binarizing the target GPA slightly higher: while Kleinberg et al. consider setting the threshold at "At least mostly B's received" we set it at "At least A's and B's received"[8]. The resulting dataset is overall balanced in terms of the prediction classes, but unbalanced in base rate and highly unbalanced group sizes. See Table 6.1.

Adult Income. We adopt the `Adult Income` dataset Kohavi (1996) for the purpose of a fairness application from Padh et al. (2021), and thus follow their pre-processing steps. The resulting dataset consists of 48,842 samples, and 49 features, with a combination of categorical and continuous variables. The target for prediction is a binarized INCOME variable, and the sensitive attribute in this case is SEX[9]. We note the `Adult Income` dataset shows an imbalance in class labels, as it includes many more negative examples than positive examples. While Padh et al. (2021) do not adjust their learning strategies or their reporting to this setting, we note that a constant predictor would achieve an accuracy of 76% by simply predicting the negative label for all examples. Thus, for the finetuning approach we consider, we employed a correction for imbalance during training (see Pytorch's BCE documentation). Finally, we would like to acknowledge our familiarity with recent work suggesting to stop using this dataset for fairness studies Ding et al. (2021), but we would like to emphasize we include it here to allow for direct comparison with Padh et al. (2021).

6.6.2 Data Pre-Processing

For the `COMPAS` and `NELS` datasets, we mainly followed the pre-processing steps taken by the original authors and analyzers (Larson et al., 2016; Kleinberg et al., 2018). We will further include the exact steps taken, as well as resulting csv files with data as used by us.

[7]The original code is at `https://www.openicpsr.org/openicpsr/project/114435/version/V1/view`.

[8]A and B are used in this context as grades in American educational institutions; another way to view the difference is setting the threshold for $Y = 1$ at > 3.25 instead of > 2.75 in Kleinberg et al..

[9]Padh et al. also allow for using RACE as an additional sensitive attribute, but we focus on the single group setting in this work.

Finally, we adopt the treatment of the `Adult Income` dataset directly from Padh et al. (2021).

6.6.3 Data Splits and Cross Validation

For all datsets, two-thirds of the dataset were used as a training set. As for validation and test splits, those were created and selected via kfold cross validation on the remaining third of each dataset. Notice that there is a slight difference in pipelines between the multi-objective experiments, where we tried to stay as close as possible to the procedures in Padh et al. (2021) for comparison purposes, while for the finetuning objective experiments we developed our own pipeline. Thus, for details on the exact procedure for validation and averaging of results over runs for the multi-objective case, see Padh et al. (2021).

For the finetuning experiments, we used 3-folds for cross-validation for `COMPAS` and `MNIST` (as the model was more complex and required a little longer training time), while for NELS we used 5-folds. Those were operationalized via the pedregosa2011scikit (Pedregosa et al., 2011) library, using StratifiedKFold in particular. Stratification was done over the sensitive attribute, to make sure the same proportion of the minority group is achieved in each fold.

6.7 Additional Experimental Results

6.7.1 Multi-Objective Results

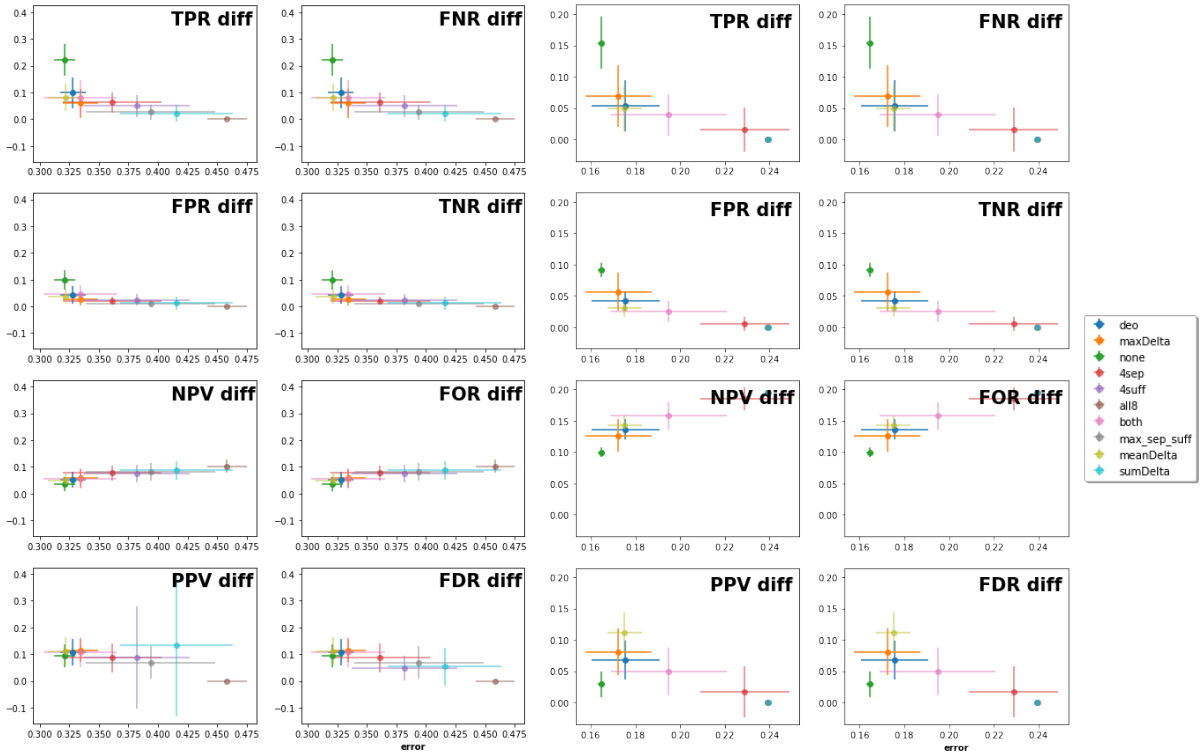
In addition to the results reported in Section 6.5, we have also considered other forms of objectives to represent different forms of constraints. Among these were the sum of all Δ-violations, the mean of all Δ-violations, all 4 Δ^{sep} (4sep), all 4 Δ^{suff} (4suff) or both (all8), both $\Delta^{\text{sep}}_{\text{max}}$ and $\Delta^{\text{suff}}_{\text{max}}$ (called max in the plot below), as well as both DEO and DP (exactly like in Padh et al. (2021), and called both in the following figure's legend).

In Figure 6.6 we provide an example of such objectives applied to the `COMPAS` and `Adult` datasets. The pattern of performance was somewhat similar across our trials.

6.7.2 Finetuning Results

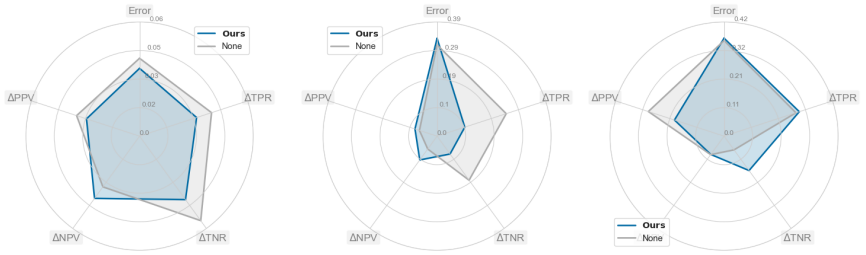
6.7.2.1 Hyperparmeter Choice for Fine-Tuning Experiments

A range of hyperparameters were considered for obtaining the results in the manuscript. We will attach the code to reproduce the results alongside this work, to make results



(a) `COMPAS` (b) `Adult`

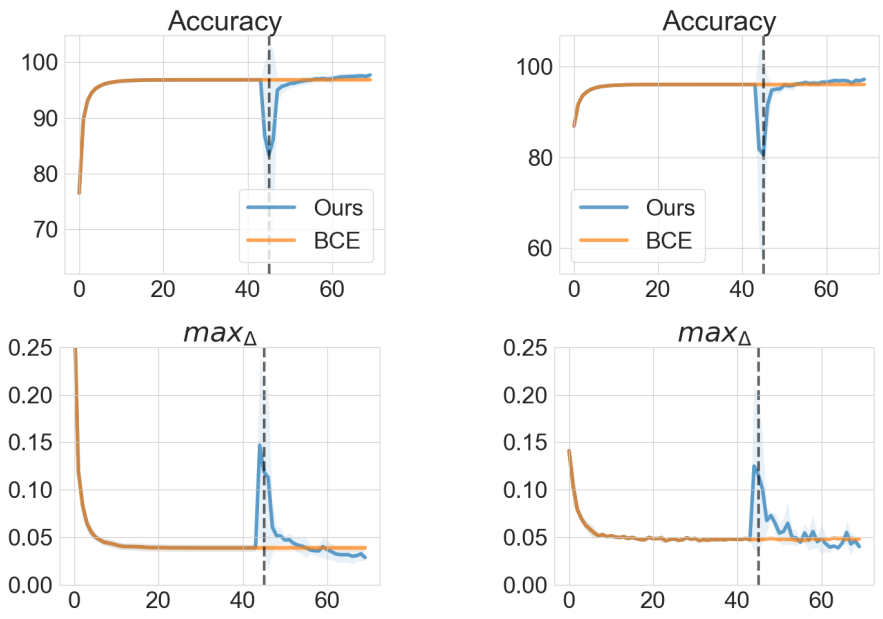
Figure 6.6: An example of performance of alternative objectives on the COMPAS dataset, averaged over 25 on the test set. The main point of this figure is to demonstrate we did consider alternative objectives, as the multi-objective framework indeed allowed, but chose to focus on the most promising forms we ended up including in Section 6.5.



(a) `Color-MNIST` (b) `COMPAS` (c) `NELS`

Figure 6.7: Finetune experiment with $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ objective (instead of $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ used in practice in Section 6.5). Error and fairness violations in terms of PPV, NPV, TNR and TPR absolute difference across groups for BCE vs. $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ (indicated as **Ours** in the legend above).

easy to reproduce, and hopefully make all choices we made explicit. We discuss the gist of them here.



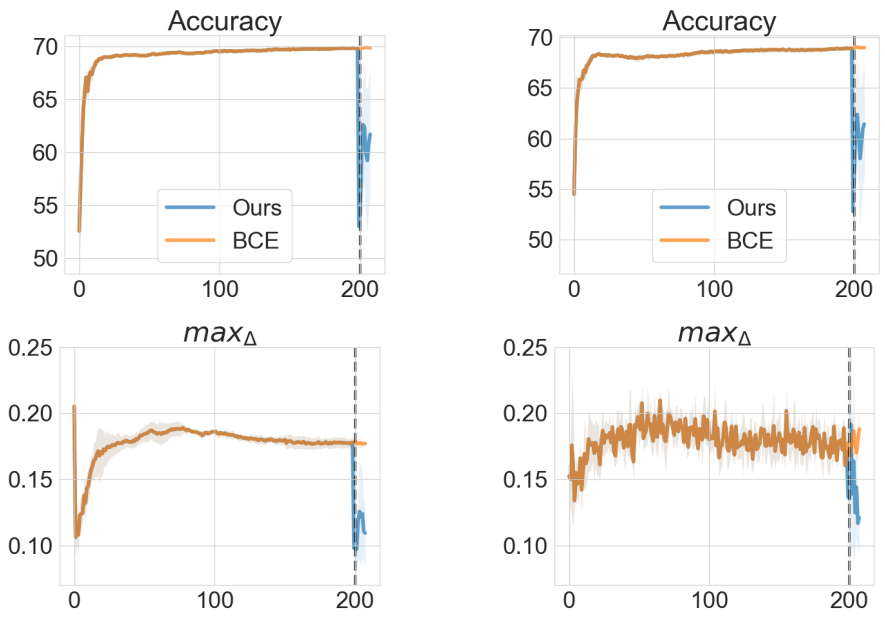
(a) Training split

(b) Validation split

Figure 6.8: `Color-MNIST` `finetune` $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ experiment. Curves indicate the mean and 95%-confidence intervals for 3 different model initializations. Validation results are also averaged over 3 cross-validation (CV) folds. The dashed line indicates the loss switch for the finetuning (here BCE to $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$).

Values for each of these hyperparameters, for each one of `Color-MNIST`, `COMPAS` and `NELS`, were chosen via grid search. The ranges we considered for each one of the key hyperparameters are as follows:

- $\gamma \in [0.1, 0.99]$ - controls the decrease in learning rate at each step taken by the scheduler.
- **weight decay** $\in [1e^{-6}, 1e^{-1}]$ - strength of penalty on magnitude of model parameters, akin to ℓ_2 regularization.
- **scheduler step size** $\in [20, 200]$ - determines the intervals of epochs at which the scheduler applies a reduction in learning rate.
- **learning rate** $\in [1e^{-8}, 1e^{-1}]$ - learning rate used by optimizer to update model parameters.
- **hidden dimensions** $\in [64, 512]$ - number of hidden dimensions of the Multi-Layered Perceptron model (`NELS` uses a single-layered logistic regression, so not applicable).



(a) Training split (b) Validation split

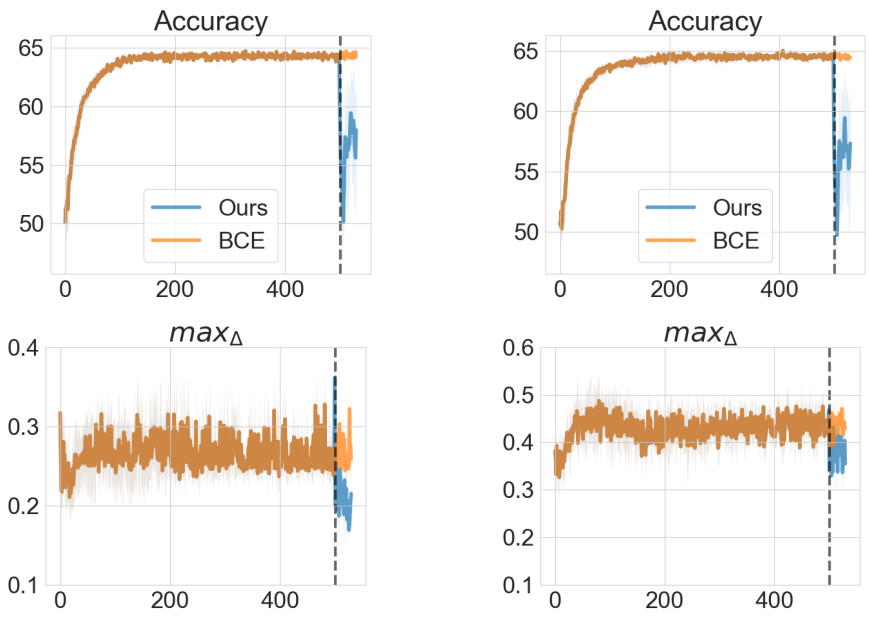
Figure 6.9: `COMPAS` `finetune` $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ experiment. Curves indicate the mean and 95%-confidence intervals for 5 different model initializations. Validation results are also averaged over 3 cross-validation (CV) folds. The dashed line indicates the loss switch for the finetuning (here BCE to $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$).

- **finetune learning rate** $\in [1e^{-8}, 1e^{-1}]$ - learning rate used once objective function changed as part of the finetuning regime.
- **finetune** $\gamma \in [0.1, 0.85]$ - γ (as above) used once objective function changed as part of the finetuning regime.

The optimizer used in all cases was Adam (Kingma and Ba, 2015).

6.7.2.2 Additional Experiments

In Section 6.5, we demonstrated the applicability of a finetuning optimization with our proposed objectives. We focused on the `COMPAS` dataset to provide a proof of concept, but elaborate here on the rest of the results that make up figure 6.5. In Figure 6.7, we additionally present results for using the finetuning approach with the $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ variant of our criteria, rather than $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$. We do so to argue that one could use either to achieve similar goals. Hyperparameter choices are overall comparable for both of these objective formulations, with an occasional difference in number of epochs run or a slight difference in finetune learning rate.



(a) Training split (b) Validation split

Figure 6.10: `NELS finetune` $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ experiment. Curves indicate the mean and 95%-confidence intervals for 5 different model initializations. Validation results are also averaged over 5 cross-validation (CV) folds. The dashed line indicates the loss switch for the finetuning (here BCE to $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$).

Color-MNIST. Similarly to the behavior we have seen for `COMPAS` in Section 6.5, in Figures 6.5 and 6.7 we see we are able to fine better balancing point between fairness violations in terms of absolute group difference in PPV and NPV (*sufficiency* violations) and TPR and TNR (*separation* violations). Notice how we are able to find better accuracy predictors, that are also better in terms of 3 out of 4 violations. This is the case when using both $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ and $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ as the finetuning objective. We also include training and validation curves in Figure 6.8.

COMPAS. We provide the training and validation curves for the `COMPAS` experiments in Figure 6.9.

NELS. The results obtained by the $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ and $\max(\Delta^{\text{suff}}_{binary}, \Delta^{\text{sep}}_{binary})$ are even closer for the `NELS` dataset. With either approach, we find a predictor that is equivalent in terms of accuracy, but offers much smaller PPV violation. We do so by trading off a relatively smaller increase of TNR, and where TPR and NPV are equivalent, as can be seen in Figure 6.5. A similar dynamic can be seen in Figure 6.7. We also include training and validation curves in Figure 6.10.

6.8 Conclusion

We have demonstrated that beyond the impossibility result emanating from the COMPAS debate, there is a possibility as well. While one cannot hold perfect calibration and still have equal error rates in settings where a perfect classifier is not available and labels are not distributed equally across groups, the approximate versions of *sufficiency* and *separation* allow us to look for better balance within the well-known trade-off. We have shown how the $\max(\Delta^{\text{suff}}, \Delta^{\text{sep}})$ criterion can be used within a Multi-Objective setting as well as a finetuning regime. We emphasize that this is one possible approach to minimizing both *sufficiency* and *separation* violations, but one could prioritize other forms of trade-offs, e.g., designing objectives that put greater emphasis on one prediction class or type of violation, based on the use case. However, we aimed to demonstrate the advantage of trying to optimize both in the general case.

7 | Pragmatic Fairness: Developing Policies with Outcome Disparity Control

7.1 Introduction

The fairness of decisions made by machine learning models involving underprivileged groups has seen increasing attention and scrutiny by the academic community and beyond. A growing body of literature has looked at the unfavorable treatment that might arise from historical biases present in the data, data collection practices, or the limits of modeling choices and techniques. Within this field of study, the vast majority of works looked at the fair prediction setting, studying the predictions given by supervised models using observed labels to emulate the historical policies that generated the datasets (e.g. Hardt et al., 2016; Woodworth et al., 2017; Corbett-Davies and Goel, 2018; Zafar et al., 2017a; Zemel et al., 2013, 2016; Goel et al., 2018). Fairness criteria suggested for this setting mainly involved enforcing relations of the outcomes with respect to membership in a sensitive group, such as independence of the outcomes across groups where dependence is deemed unfair (demographic parity (DP)), or independence of the errors across groups in the setting in which dependence of the outcomes is deemed fair (equality of opportunity (EoP) and Equalized Odds (EO) (Barocas et al., 2019)).

In this chapter, we propose a causal framework for learning optimal policies with fairness constraints that are inspired by the public health literature (Jackson and VanderWeele, 2018; Jackson, 2018, 2020). Unfairness is defined as the presence of differences in the distribution of the outcome stratified by levels of chosen sensitive attributes, where the distribution depends on a policy that a decision maker is responsible for selecting. Unlike the few works in the fair policy optimization literature, we reason about the problem by treating sensitive attributes as effect modifiers rather than causes as in the Counterfactual Fairness line of work (Kusner et al., 2017b, 2019; Chiappa, 2019), and by considering the actual effect of our actions, as opposed to planning under an idealized outcome distribution that removes the unfair contributions of the sensitive attributes (Nabi et al., 2019). Critically, we take a pragmatic view

and ask how we can best control or mitigate disparity by the introduction of the new policy, given an action space. Causal modeling is used here as a principled approach to leveraging historical data and proposing constraints to mitigate disparity.

We assume a setting in which the decision maker performs a single-stage action conditioned on pre-action sensitive attributes and covariates under the assumption that there are no unmeasured confounders, which allows the use of data collected prior to learning and optimization under an observational-data regime.

We consider two fairness goals: i) equalizing the impact of a new policy w.r.t. sensitive attributes in order to conservatively avoid introducing new disparities, motivated by mitigating individual-level differences with the baseline policy; and ii) actively reducing disparities at the population level to the extent permitted by the action space. We demonstrate how these constraints operate on two semi-simulated datasets.

7.2 Disparity Controlled Policy

Let Y be an outcome of interest, and S be *sensitive attributes*[1]: characteristics of an individual (e.g. race, gender, disabilities, sexual or political orientation) which we wish to protect against some measure of unfairness, and X are some covariates.

We consider the task of learning how to select actions A based on S and X according to a policy that maximizes the expected outcome while controlling for disparity, which we formally define in the section below. The policy is defined by the conditional distribution $p(A|S, X; \sigma_A)$ where, following Dawid et al. (2007) and Correa and Bareinboim (2020)[2], σ_A denotes a *regime indicator* (we also refer to σ_A as the policy for simplicity).

We assume the setting represented by the causal graph on the left. S and X are allowed to be associated, as indicated by the undirected edge $S - X$ and potentially directly influence Y. A can only indirectly control this influence. This setting formalizes a situation common to many real-world scenarios in which control for the association of S and Y can only be achieved to some extent through a predefined set of available actions, i.e., a given action space. The level to which we can minimize such unfair impact will therefore *depend on the choice of the action space.* We consider an observational-data regime in which the decision maker learns the optimal policy based

[1]As a reminder, while in the chapter above we introduce the membership in a sensitive group as A, we switch to S in the context of fair policy optimization, as we will need A to refer to the action node that our regime indicator/optimized policy parametrization will act on.

[2]Note that we consider a special case of soft interventions and potential outcomes, where we change the distribution $p(A|S, X)$ via the regime indicator, but do not cut the links between the intervention target A and its parents S, X.

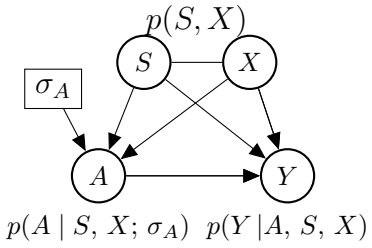


Figure 7.1: Problem setup.

on historical data collected using a *baseline policy* $\sigma_A = \varnothing$ (we indicate $p(A \mid S, X; \sigma_A = \varnothing)$ more simply with $p(A \mid S, X; \varnothing)$), rather than by taking actions (interventional-data regime). This observational-data regime choice reflects considerations such as the cost, ethical constraints, or difficulty of collecting online data, which are particularly relevant in fairness (we discuss this point in Section 7.2.2). We denote with Y_{σ_A} the *potential outcome* random variable, which represents the outcome resulting from taking actions according to $p(A \mid S, X; \sigma_A)$

Identifiability and Causal Assumptions. The structural assumptions adopted are the following independence constraints. *Assumption 1.* $\sigma_A \perp\!\!\!\perp Y | A, X, S$. *Assumption 2.* $\sigma_A \perp\!\!\!\perp S, X$. Any causal graph in which S, X, and A are sets of vertices, with the partial ordering given by $\{S \cup X, \sigma_A, A, Y\}$, is covered by our formulation as long as it satisfies these two assumptions. In particular, they imply a conditional unconfoundedness assumption of A and Y and allows the estimation of interventional quantities from observational data (see Section 7.2.1.1). We assume unmeasured confounding to focus on clear demonstration of ideas and see relaxing it as a future direction. The second assumption follows from exogeneity of regime indicators and A not causing X.

7.2.1 Formal Problem Statement

The goal of the decision maker is to find a disparity-controlled policy: a policy σ_A that maximizes utility $\mathbb{E}[Y_{\sigma_A}]$ while also controlling for disparity. To control disparity, we propose two families of constraints on σ_A which may be applicable for different use cases, depending on the choice of context and available actions. First, an *equal benefit* (`EqB`) constraint requiring that the distribution of the difference of individual outcomes, $Y_{\sigma_A} - Y_{\varnothing}$, are approximately equal across different levels s of S. This constraint would be suitable when disparity in the baseline policy is considered legitimate (like in EO for the prediction setting), and we would like to avoid the introduction of *new* disparity when designing our policy. Second, a *moderation breaking* (`ModBrk`)

constraint aiming at removing S's influence on the expected outcome $E[Y_{\sigma_A}]$ as much as possible (similarly to DP in the prediction setting), but uniquely via *what we can control*: the allocation of A as determined by $p(A|S, X; \sigma_A)$. In other words, we would like to minimize the interactions of A and S as they give rise to Y. This constraint involves population-level quantities, making the constraint fully identifiable under an unmeasured unconfoundedness assumption.

Real-world Considerations. It is an entirely problem-dependent question whether or not it is desirable to control for disparities of outcomes across levels of the sensitive attributes. For instance, if the action space consists of solely two options, to give medical treatment or not, it may be unclear why we should take into consideration a difference in rates by which individuals from different groups recover: other things being equal, we just want to maximize the numbers of lives saved. In contrast, if Y is a measure of wealth, pure wealth maximization may be judged to be harmful if disparities among groups in S are exacerbated. In this case, we may settle for a scenario with less aggregated wealth if disparities are controlled. Such value judgments are *not* to be decided algorithmically. Our goal is to provide a formalization of disparity control *if* it is judged to be desirable and to provide an estimate of *whether* different levels of control can be achieved under an acceptable loss of total expected outcome, *given* an action space that is, again, a property of the real world. For further discussion of the motivation behind the constraints, see Section 7.2.3.

Disparity Control via the `EqB` Constraint. We formalize the `EqB` constraint objective using the cumulative distribution function (cdf) $\mathcal{F}$ as

$$\arg\max_{\sigma_A} \mathbb{E}[Y_{\sigma_A}] \quad \text{s.t.} \quad \mathcal{F}(Y_{\sigma_A} - Y_{\varnothing} \mid S = s) = \mathcal{F}(Y_{\sigma_A} - Y_{\varnothing} \mid S = \bar{s}), \forall s, \bar{s}, \tag{7.1}$$

assuming discrete S. This constraint ensures that a new policy has a similar magnitude of effect across sensitive attributes. The aim is *to not increase* disparity via the new policy, rather than decreasing it. In general, the distribution of $Y_{\sigma_A} - Y_{\varnothing}$ is not identifiable, due to the impossibility of having the joint distribution of the two potential outcomes (a problem often referred to as the "fundamental problem of causal inference" in the context of atomic interventions). Therefore, similarly to bounding the unidentifiable individual treatment effect for atomic intervention (Pearl, 2000), we propose matching testable upper bounds and lower bounds of the cdf. The rationale of this setup is, assuming that we have either a currently acceptable level of disparity or that the nature of σ_A cannot reduce it, the decision maker should not increase disparity for any particular individual as compared to

individuals in a different group. Moreover, we can only constrain our policy based on what empirical evidence is provided. In the context of the individual-level contrast $Y_{\sigma_A} - Y_{\varnothing}$, this leaves us to compare bounds that can be learned from observational data.

Disparity Control via the `ModBrk` Constraint. In the `ModBrk` constraint, we propose reducing disparity in expectation, to the extent that the policy space of σ_A allows us. To better understand the limits of what can be achieved, let us consider, without loss of generality, the following representation for $\mu^Y(a, s, x) := \mathbb{E}[Y \mid a, s, x]$

$$\mu^Y(a, s, x) = f(s, x) + g(a, s, x) + h(a, x). \tag{7.2}$$

Note that this is not a linearity assumption: this is a decomposition without loss of generality and is akin to summation in the last layer of a neural network. This leads to the following factorization of $\mu^Y_{\sigma_A}(s, x) := \mathbb{E}[Y_{\sigma_A} \mid s, x] = \int_a p(Y \mid a, s, x) p(a \mid s, x; \sigma_A)$

$$\mu^Y_{\sigma_A}(s, x) = f(s, x) + g_{\sigma_A}(s, x) + h_{\sigma_A}(x),$$

where $g_{\sigma_A}(s, x) := \mathbb{E}[g(A, s, x) \mid s, x; \sigma_A]$, and $h_{\sigma_A}(x) := \mathbb{E}[h(A, x) \mid s, x; \sigma_A]$. This decomposition explicitly highlights that $\mathbb{E}[Y_{\sigma_A} \mid s, x]$ has (i) a component $f(s, x)$ that cannot be affected by σ_A at all, but which can contribute to disparity; (ii) a component $h_{\sigma_A}(x)$ that can be adjusted by σ_A to increase expected outcomes but do not affect differences due to S; and (iii) a component $g_{\sigma_A}(s, x)$ by which our choice of σ_A can influence differences that are *moderated* by S.

Considering how $\mu^Y_{\sigma_A}(s, x)$ varies with s as a measure of disparity suggests the formulation

$$\arg\max_{\sigma_A} \mathbb{E}[Y_{\sigma_A}] \quad \text{s.t.} \quad (\mathbb{E}[g_{\sigma_A}(s, X) \mid S = s] - \mathbb{E}[g_{\sigma_A}(\bar{s}, X) \mid S = \bar{s}])^2 \leq \varepsilon, \forall s, \bar{s} \tag{7.3}$$

The slack ε is chosen in practice by a combination of domain requirements and the feasibility of the problem: central to the setup is that we work with a given space of policies, which is constrained by real-world phenomena and, in general, can only do so much to reduce disparity. The idealized separation of the expected outcome into components f, g, and h is meant to highlight that. This constraint is a valuable tool in settings where we acknowledge different groups might require different treatments, but we would not like the policy to *affect* different populations differently.

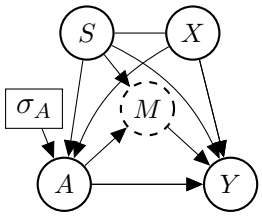
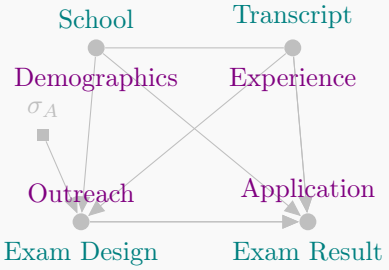


Figure 7.2: Motivating illustration of `ModBrk`.

Example use cases



To gain more intuition of scenarios in which the `EqB` constraint will be appropriate, consider the graph on the left (green labels) corresponding to an educational committee desiring to change the design of a high-school exam. It is known that the student's school of origin is associated with exam results, but the committee does not have the power to change academic disparities in the delivery and resources available at different locations: it sees its role as updating the exam without introducing further disparities among students at different schools. The policy is an update to the exam that should not discriminate against, e.g., schools without expensive extra-curricular activities or of particular predominant cultural backgrounds. The policy space is constructed to only produce exams of acceptable academic standards (otherwise, policy optimization can be achieved trivially by allowing exams to be overly easy which is not a desirable outcome for the exam committee) so that the goal becomes improving clarity to maximize exam scores, better-reflecting knowledge of all students.

Alternatively, this constraint could also be useful when one hopes to allow for legitimate dependence of the outcome on group membership. In such a setting, we would not like to change such dependence when introducing a new policy. This could be the case, e.g., when attempting to improve the accuracy of a medical device that was tailored to account for biological differences between males and females; we would like the gender difference in outcome to be kept under an update to the measurement process.

As a scenario where the `ModBrk` constraint may be appropriate, consider a company looking to change its outreach campaign A to mitigate imbalances in the demographics S and potentially associated level of experience X of job applicants Y (above graph – purple labels). The company cannot control factors such as cultural preference among applicants for industry sectors but can induce modifications such as focusing recruiting efforts on events organized by minority groups in relevant conferences, thus optimizing for the things we can control, and choosing recruiting strategies that do not interact with group membership.

Some more insights into the `ModBrk` constraint can be gained by noticing that we are operating in the (estimated) true process that follows a particular σ_A. We marginalize the intermediate process between (A, S, X) and the outcome Y, but implicitly assume that our actions change mediating events, such as M in the above graph. The extent to which we can mitigate unfairness is a property of the real-world space of policies. For instance, we cannot interfere in the direct dependence between S and Y except for the moderating effect that M might have under a new policy. This is in contrast to previous work, such as (Nabi et al., 2019), which solves a planning problem in a "projection" of the real-world process where unfair information has been removed by some criteria.

7.2.1.1 Our Causal Assumptions and Identifiability of the Objective Function

We use the regime indicator σ_A (as in Dawid et al. 2007 and Correa and Bareinboim 2020) only to indicate a parametrization of the distribution $p(A|S, X, \sigma_A)$, i.e. different regime indicators imply different distributions, and to be able to indicate that this distribution is not fixed in the causal graph. Thus, we use a special case of soft interventions, where we do not necessarily need to change the parents of A (i.e. cut links in the causal graph), but rather formulate our soft-interventions/regime indicators as a different parametrization of the policy. This is all the machinery we need for the rest of our method. We frame the `EqB` constraints directly in terms of potential outcomes, and therefore discuss identifiability using this potential outcome formalism, as commonly done in the literature (see e.g. (Dawid et al., 2007) for the relation between regime indicators and potential outcomes). Discussion of identifiability for such quantities does not require a structural causal model setting, which can be added on top if desired as a possible framing, but it will not remove any assumptions from the model nor introduce any necessary assumptions, and hence will not change the applicability or validity of the method (to dispel any misunderstanding, what we provide are causal assumptions even if an SCM is not invoked: the notion of regime indicator is a causal concept, not a probabilistic one).

As we state at the top of Section 7.2.1, the independence constraints entailed by the chain graph depicted in Fig. 7.1 ($\sigma_A \perp\!\!\!\perp Y \mid \{S, X\}$; $\sigma_A \perp\!\!\!\perp \{S, X\}$) will imply the identifiability conditions we need. Namely, we get $p(y|x, s; \sigma_A)$ from $p(y|x, s; \sigma_A) = \int p(y|x, s, a; \sigma_A)p(a|s, x; \sigma_A)\, da = \int p(y|x, s, a; \varnothing)p(a|s, x; \sigma_A)\, da$. Thus, we can identify the required objective function of our problem formulation given (an estimate of) $p(y|x, s, a; \varnothing)$ and the fact that $p(a|s, x; \sigma_A)$ is chosen by us. This is a standard result

in the literature. As a matter of fact, the very proof of the backdoor criterion uses a regime indicator, and nothing but the two conditional independencies we assume (see (Pearl, 2000) Section 3.3.1, page 80 (second edition)). Nonparametric identification boils down to deriving the implications of conditional independencies between regime indicators and random variables. This is the main lesson of (Dawid et al., 2007). The do-calculus can be seen as a sophisticated sound and complete way of deriving such results from a DAG independence model.

7.2.2 Relation to Prior Work

Most work on ML fairness has focused on learning predictions $\hat{Y}$ of Y that satisfy constraints on the relations between $\hat{Y}$ and S (Barocas et al., 2019). We are instead interested in the decision-theoretic problem of selecting a policy that gives a more favorable outcome and ask that the distribution of outcomes is fair. The solution to this problem cannot be obtained by using fair predictors. Consider, e.g., the problem of fair loan allocation: a fair predictor of loan repayment would simply change the rates at which loans were granted, thus using such a predictor as a criterion for giving someone a loan would be problematic. The appropriate approach would be to ask for a policy that maximizes the probability of repayment, subject to fairness constraints. This issue has been studied under the names of "self-fulfilling prophecy" and "delayed impact of fair ML" in Dwork et al. (2012) and Liu et al. (2018), respectively. D'Amour et al. (2020) provided a tool to explore this issue further in simulated multi-stage online settings. We are also not aiming to match a distribution of historical labels Y, but to optimize a future expected utility.

Increasing attention is being given to the problem of designing optimal policies that have some fairness guarantees. Like us, Nabi et al. (2019) uses observational data, but in order to learn a policy as if particular path-specific effects between S-A and S-Y were completely deactivated. They do not place any constraints on $p(y_{\sigma_A} \mid s, x)$, require complex counterfactual computations, and aim to achieve a notion of fair policy which targets specific causal subpaths. Critically, they once again rely on manipulation of the sensitive attribute S and do not have an equivalent to our pragmatic view, asking what we can do in this world with a set of available actions. Chohlas-Wood et al. (2021) considers a more general utility function than ours as the optimization objective but focuses on enforcing a notion similar to demographic parity on the choice of the policies, rather than considering a fairness notion on the outcomes. Still, in the observational data regime, the most similar work has been Kusner et al. (2019). Crucially, it is different in its motivation: it aims to extend Kusner et al. (2017b)

to the policy setting and thus relies on manipulation of the sensitive attribute S. Further, it consists mostly of budget treatment allocations and interference problems. It does not have our pragmatic view and does not consider individual-level effects or parameterized policy spaces that take into account a covariate vector X. Due to the difference in their disparity definition, this method would not be directly comparable to ours. Considering interventional data in the online setting, Joseph et al. (2016, 2018) derive algorithms that have "meritocratic fairness," requiring that, at every round, no arm with lower expected reward is preferred to one with a higher expected reward. This goal can be seen as an individual-level fairness metric, but it is not at odds with a policy that maximizes utility; instead, its main concern is to offer more conservative exploration.

Recently, fairness in bandit settings has received a lot of additional attention (Huang et al., 2021; Schumann et al., 2022; Patil et al., 2020). We see interventional data and online sequential settings as an interesting future direction, especially appropriate for use cases with low individual impact, such as software and internet applications that interact with users naturally and frequently, at low participation cost. However, we focus on observational-data regimes which are often more appropriate, especially when the collection of interventional data could be harmful, or when it would be unfair for individuals to essentially pay the price for exploration and calibration of learning agents.

7.2.3 Further Motivation for the Constraints

We draw the inspiration to our problem setup from works in the public health literature which directly consider and estimate the efficacy of possible interventions, within a defined action space, to the reduction or control of disparity. One such work offers the following illuminating example (Jackson and VanderWeele, 2018):

"We now describe results from decompositions that estimate how well certain interventions might reduce racial disparities in adulthood (wages, unemployment, incarceration, and health) by equalizing childhood SES and/or test scores across race."

In our notation, their action space consists of particular hypothetical interventions on socio-economical status with the objective of, e.g., minimizing differences in wages. These are real direct policy-making problems, as opposed to the indirect problem of classifier construction to emulate historical patterns, where the selection of action is implicit within the prediction algorithm. We build on Jackson and VanderWeele to consider a real actionable space σ_A, as "intervening on SES" is an idealization since socio-economical status is a construct of many moving parts for which no "do(SES)"

exists. It is not only about removing disparities in wages but improving them to the extent σ_A allows while mitigating disparities across groups. It is taken as a principle that, even though we would like to see wages increase as much as possible, it is detrimental to do so in a way that increases disparities.

Throughout the paper we speak of "mitigating" and "reducing" biases, not completely eliminating them, when using either `EqB` or `ModBrk`. That is because *eliminating bias* in general may not be possible, or the corresponding price to be paid (low rewards) may be unacceptable. Our pragmatic viewpoint means it is all a function of the real-world mechanisms that the action space provides. We simply provide a framework to examine what is possible.

7.3 Method

We now describe how the `EqB` and `ModBrk` constrained objectives (7.1) and (7.3) can be estimated from an observational dataset $\mathcal{D} = \{a^i, s^i, x^i, y^i\}_{i=1}^n$, $(a^i, s^i, x^i, y^i) \sim p(A, S, X, Y; \varnothing)$, and introduce two methods for optimizing σ_A by learning the parameters of a model for $\mu^A_{\sigma_A}(s, x) = \mathbb{E}[A \,|\, s, x; \sigma_A]$.

7.3.1 Equal Benefit Constraint

For the `EqB` constrained objective (7.1), we require knowledge of the baseline policy as well as creating upper and lower bounds on the cdf $\mathcal{F}(Y_{\sigma_A} - Y_\varnothing \,|\, S = s)$. As the goal of this work is to explore frameworks that control the fairness of outcomes, to avoid being bogged down in more involved estimation tasks, we allow ourselves some parametric assumptions[3]. In particular, we assume that the baseline policy is Gaussian and only consider Gaussian policies, with equal homogeneous variance, i.e. $p(A|s, x; \varnothing) = \mathcal{N}(\mu^A_\varnothing(s, x), V^A)$, $p(A|s, x; \sigma_A) = \mathcal{N}(\mu^A_{\sigma_A}(s, x), V^A)$. In addition, we assume that Y_{σ_A} and $Y_\varnothing$ are jointly Gaussian conditioned on S, X, with marginal means $\mu^Y_{\sigma_A}(s, x)$, $\mu^Y_\varnothing(s, x)$ and equal homogenous marginal variance V^Y.

The assumption on Y_{σ_A} and $Y_\varnothing$ enables us to bound the population cdf by maximizing/minimizing it with respect to the unknown correlation coefficient $\rho(s, x) \in [-1, 1]$, where $Cov(Y_{\sigma_A}, Y_\varnothing \mid s, x) := \rho(s, x)V^Y$. Let Φ denote the cdf of the standard Gaussian. We can write

$$\mathbb{P}(Y_{\sigma_A} - Y_\varnothing \leq z \,|\, s, x) = \Phi\left(\frac{z - \mu^Y_{\sigma_A}(s, x) + \mu^Y_\varnothing(s, x)}{\sqrt{2V^Y(1 - \rho(s, x))}}\right).$$

[3]Lei and Candès (2021) includes a nonparametric account of individual treatment effect bounds, and can be adapted here.

Defining $\mu(s,x) := \mu^Y_{\sigma_A}(s,x) - \mu^Y_{\varnothing}(s,x)$ and $f^z_{s,x}(\rho) := \Phi\left(\frac{z-\mu(s,x)}{\sqrt{2V^Y(1-\rho)}}\right)$, we can show that, for any s, x, the following bounds are the tightest:
(i) $f^z_{s,x}(x,-1) \leq \mathbb{P}(Y_{\sigma_A} - Y_{\varnothing} \leq z \,|\, x, s) \leq f^z_{s,x}(1)$, for $z - \mu(s,x) > 0$;
(ii) $f^z_{s,x}(1) \leq \mathbb{P}(Y_{\sigma_A} - Y_{\varnothing} \leq z \,|\, x, s) \leq f^z_{s,x}(-1)$, for $z - \mu(s,x) < 0$.
See Section 7.3.1 for discussion.

Using N_s to indicate the number of elements in $\mathcal{D}$ with $s^i = s$, we obtain global lower and upper bounds estimates for $\mathbb{P}(Y_{\sigma_A} - Y_{\varnothing} \leq z \,|\, s)$ as $F^L_s(z) = \frac{1}{N_s}\sum_{i:s^i=s} f^z_{s,x^i}(-\text{sign}(z-\mu(s,x^i)))$ and $F^U_s(z) = \frac{1}{N_s}\sum_{i:s^i=s} f^z_{s,x^i}(\text{sign}(z-\mu(s,x^i)))$. We operationalize the constraint by minimizing the mean squared error (MSE) of the bounds differences, i.e. our final objective is

$$\arg\max_{\sigma_A} \mathbb{E}[Y_{\sigma_A}] \quad \text{s.t.} \quad \sum_{z \in S_z} \left(||F^L_s(z) - F^L_{\bar{s}}(z)||^2_2 + ||F^U_s(z) - F^U_{\bar{s}}(z)||^2_2 \right) \leq \varepsilon, \quad (7.4)$$

$\forall s, \bar{s}$, where S_z is a grid of values. In practice, we employ an augmented Lagrangian approach, allowing us to enforce the above as an inequality constraint controlled by the slack value ε (see Section 7.3.2).

We propose to estimate the utility $\mathbb{E}[Y_{\sigma_A}]$ with the inverse probability weighting (IPW) estimator

$$\mathbb{E}[Y_{\sigma_A}] = \int_{a,s,x,y} y \frac{p(a \,|\, s, x; \sigma_A)}{p(a \,|\, s, x; \varnothing)} p(a, s, x, y; \varnothing) \approx \frac{1}{N} \sum_{i=1}^{N} y^i \frac{p(a^i \,|\, s^i, x^i; \sigma_A)}{p(a^i \,|\, s^i, x^i; \varnothing)}.$$

Our optimization approach involves two phases outlined in Fig. 7.3. We model $\mu^A_{\varnothing}(s,x)$, $\mu^A_{\sigma_A}(s,x)$ and $\mu^Y(a,s,x) = \mathbb{E}[Y \,|\, a, s, x]$ using MLP neural networks $\text{MLP}^A_{\varnothing}$, $\text{MLP}^A_{\sigma_A}$, and MLP^Y respectively. In Phase I, we learn the parameters of $\text{MLP}^A_{\varnothing}$ and MLP^Y from $\mathcal{D}$ with MSE loss between predicted and observed actions and outcome. We obtain a MAP estimate of $\mu^Y_{\varnothing}(s,x) = \int_a \mu^Y(a,s,x) p(a|s,x,\varnothing)$ as $\hat{\mu}^Y_{\varnothing}(s,x) = \int_a \mu^Y(a,s,x)\delta_{A=\mu^A_{\varnothing}(s,x)}$, where δ denotes the delta function. We estimate V^A and V^Y through averaging the MSE of target and predicted mean output from MLP^A and MLP^Y respectively. In Phase II, we learn the parameters of the policy model $\text{MLP}^A_{\sigma_A}$ that maximize objective (7.4) using the $\text{MLP}^A_{\varnothing}$ and MLP^Y trained in Phase I.

The `EqB` Constraint Formulation. We first formulate the `EqB` constraint under a Gaussian assumption primarily for simplicity of exposition. We derive a conservative bound in Section 7.3.1 to guarantee that $F^L_s(z) \leq F^{true}_s(z) \leq F^U_s(z)$, where $F^{true}_s(z) = \mathbb{P}(Y_{\sigma_A} - Y_{\varnothing} \leq z|s)$. Note that no tighter bound exists without further assumptions.

This is a result that follows directly from the properties of a Gaussian distribution: that is, for a bivariate Gaussian (X, Y) with fixed marginals, one gets the smallest/largest value of the cdf $\mathbb{P}(X - Y \leq z)$ among all possible correlation coefficients when their correlation is -1 or 1 (depending on z and the marginal means). This can be directly verified in the equation following the second paragraph of Section 7.3.1 by simple differentiation. Thus we know the global bound, $F_s^U(z) - F_s^L(z)$.

Relying on such-tight-as-possible bounds, equating the upper and lower bounds across groups will indeed ensure similar constraints, and is the best we can do if we do not want to deal with the quality of approximation itself, but rather enforce a similarity between approximations across groups. In other words, the goal is to make the distribution of $D \equiv Y_{\sigma_A} - Y_{\varnothing}$ as independent as possible of the values of S, meaning any pair s and $\bar{s}$ should imply the same conditional distribution for D (up to an agreeable tolerance level ε). Ideally, if the conditional distribution of D was identified, the constraint would be

$$|\mathcal{P}(D \leq z \mid s) - \mathcal{P}(D \leq z \mid \bar{s})| \leq \varepsilon$$

As these distributions are *not* identified, we apply the principle of aiming at *not having any evidence against the lack of equality between these distributions*. This means making *the set of all possible* $\mathcal{P}(D \leq z \mid s)$ *and* $\mathcal{P}(D \leq z \mid \bar{s})$ *be the same for all* z. As the smallest of such sets are given by intervals, this is equivalent to making the upper bounds the same and the lower bounds the same, *not* to mix upper and lower bounds:

$$|\mathcal{P}^U(D \leq z \mid s) - \mathcal{P}^U(D \leq z \mid \bar{s})| \leq \varepsilon$$
$$|\mathcal{P}^L(D \leq z \mid s) - \mathcal{P}^L(D \leq z \mid \bar{s})| \leq \varepsilon$$

Finally, in order to reduce the number of possible Lagrange multipliers, we use the norm in (4) as an alternative to enforcing all of the constraints separately.

For extensions beyond the normality assumption, there are very standard worst-case tight bounds based on standard copula theory: see use of Fréchet bound in the linked blogpost, the nonparametric approach in (Lei and Candès, 2021).

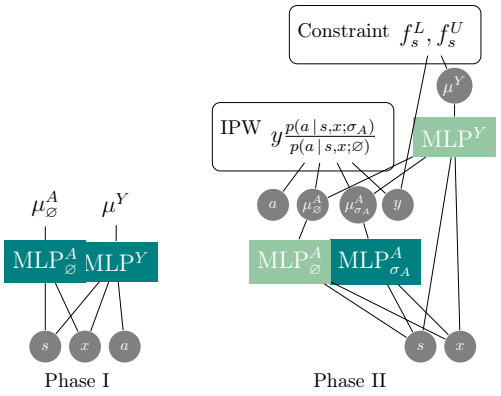


Figure 7.3: Training phases for `EqB`.

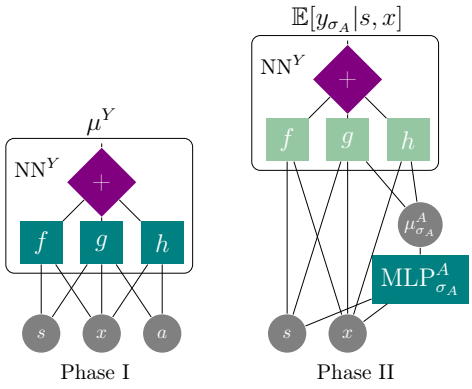


Figure 7.4: Training phases for `ModBrk`.

Gray circular nodes: inputs to the models. Teal blocks: parameter layers. Light green blocks: fixed parameters. Purple diamond nodes: additive gates.

7.3.2 Moderation Breaking Constraint

Recall the `ModBrk` constrained objective in Eq. 7.3.[4] We optimize σ_A in two phases, outlined in Fig. 7.4. We model $\mu^A_{\sigma_A}(s,x)$ using an MLP neural network MLP$^A_{\sigma_A}$, and $\mu^Y(a,s,x) = f(s,x) + g(a,s,x) + h(a,x)$ using a structured neural network NN^Y that separates into the three components f, g, h reflecting the decomposition of $\mu^Y(a,s,x)$. In Phase I, we estimate the parameters of NN^Y from $\mathcal{D}$. In Phase II, we learn the parameters of the policy model MLP$^A_{\sigma_A}$ by optimizing objective (7.3) with $\mathbb{E}[Y_{\sigma_A}] \approx \frac{1}{N}\sum_{i=1}^N \mathbb{E}[Y_{\sigma_A}|s^i,x^i] = \frac{1}{N}\sum_{i=1}^N \mu^Y(\mu^A_{\sigma_A}(s^i,x^i),s^i,x^i)$, using the NN^Y trained in Phase I[5].

Action Clipping. For the `EqB` constraint, the IPW estimator ensures coverage of $p(a|s,x;\sigma_A)$ and $p(a|s,x;\varnothing)$, as it facilitates a reweighting of the observed actions. For the `ModBrk` constraint, in which we make use of deterministic policies, we suggest ensuring an overlap of $p(a|s,x;\sigma_A)$ with $p(a|s,x;\varnothing)$ by adaptively constraining the output of the policy model MLP$^A_{\sigma_A}$ to be within an interval that resembles the one observed under the baseline policy $\varnothing$. We considered two options: (a) matching the minimal and maximal value a seen for each s,x combination (binning continuous elements as necessary) and (b): extending that interval according to the difference between the minimal and maximal A value seen with each S,X. In the experiments, we opted for constraining the output of MLP$^A_{\sigma_A}$ to be in the interval $[\min_{A_{S,X}} - \eta\text{gap}_{A_{SX}}, \max_{A_{S,X}} + \eta\text{gap}_{A_{SX}}]$,

[4]This formulation can also be extended to a setting where the group membership variable S is continuous by defining constraint via a partial derivative, $\mathbb{E}\left[\left|\frac{\partial g_{\sigma_A}(s,X)}{\partial s}\right| \;\middle|\; S=s\right] \leq \varepsilon$.

[5]If all variables are discrete, after estimating $\mu^Y(a,s,x)$ and $p(s,x)$, computing Eq. 7.3 can be cast as an LP (see Section 7.3.2).

where $\text{gap}_{A_{SX}} = \max_{A_{S,X}} - \min_{A_{S,X}}$. We enforce this interval by placing a shifted sigmoid function in the last layer of $\text{MLP}^A_{\sigma_A}$. We set $\eta = 1$ to allow some extrapolation and increase in utility.

Augmented Lagrangian Penalties. We write the constrained problem as $\min_x \max_{\lambda \geq 0} \; f(x) + \lambda(c(x) - b)$. The max returns $f(x)$ when x satisfies the constraints (as the maximum is obtained at $\lambda = 0$), and ∞ otherwise (as the maximum is at $\lambda = \infty$). There could a difficulty in optimizing the form above directly, as λ jumps from 0 to ∞ when passing through the constraint boundary. To fix this we add a penalty on making large changes to λ, obtaining $\min_x \max_{\lambda \geq 0} \; f(x) + \lambda(c(x) - b) - 1/(2\mu||\lambda - \lambda'||_2)$, where λ' are the Lagrange multipliers from the previous iteration and $\hat{\mu}$ is a penalty term iteratively increased.

The inner max can be solved in closed form as

$$\lambda = \begin{cases} 0 & \text{if} \quad k(x) \leq -\frac{\lambda'}{\mu} \\ \lambda' + \mu\, k(x) & \text{otherwise} \end{cases} \tag{7.5}$$

where k(x) := (c(x)-b), representing the inequality constraint. Plugging the above into a complete optimization problem,

$$\begin{aligned} &\min_x f(x) + \phi(k(x), \lambda', \mu) \\ &\text{s.t.} \quad \phi(k(x), \lambda', \mu) = \begin{cases} -\frac{\lambda'^2}{2\mu} & \text{if} \quad k(x) \leq -\frac{\lambda'}{\mu} \\ \lambda' k(x) + \frac{\mu\, k(x)^2}{2} & \text{otherwise} \end{cases} \end{aligned} \tag{7.6}$$

This suggests a straight-forward algorithm: at each step, minimize the above, update the value of λ via equation 7.3.2, and repeat until convergence.

The `ModBrk` Constraint as an LP. The optimization problem Eq. 7.3 can be cast as a linear program for the common case where if the variables are discrete. We decouple the estimation problem and assume that $\mu^Y(a, s, x)$ and $p(s, x)$ are known or appropriately estimated.

First, consider the case when A is binary; then, the policy may be identified by $\mu^A_{\sigma_A}(s, x) = p(A = 1 | S = s, X = x; \sigma_A)$. For convenience, define $\delta(s, x) =$

$\mu^Y(1,s,x) - \mu^Y(0,s,x)$. Then, we may evaluate

$$\mathbb{E}[Y_\pi] = \sum_{s,x} p(s,x)\left(\mu^Y(1,s,x)\mu^A_{\sigma_A}(s,x) + \mu^Y(0,s,x)(1-\mu^A_{\sigma_A}(s,x))\right) = \sum_{s,x} p(s,x)(\mu^Y(0,s,x) + \delta(s,x)\mu^A_{\sigma_A}(s,x)), \quad (7.7)$$

which simplifies the optimization problem to

$$\begin{aligned}
&\max_{\sigma_A} \sum_{s,x} p(s,x)\delta(s,x)\mu^A_{\sigma_A}(s,x) \\
&s.t. \quad \sum_x \left(p(s,x)\delta(s,x)\mu^A_{\sigma_A}(s,x) - p(s',x)\delta(s',x)\mu^A_{\sigma_A}(s',x)\right) \le \sqrt{\varepsilon} \quad \forall s, s' < s, \text{ and} \\
&\qquad \sum_x \left(p(s,x)\delta(s,x)\mu^A_{\sigma_A}(s,x) - p(s',x)\delta(s',x)\mu^A_{\sigma_A}(s',x)\right) \ge -\sqrt{\varepsilon} \quad \forall s, s' < s,
\end{aligned}$$

which is a linear program in σ_A with a linear objective and linear constraints.

To generalize to the non-binary case, we define, by a slight abuse of notation, $\mu^A_{\sigma_A}(a,s,x) = p(A=a|S=s,X=x;\sigma_A)$ and evaluate the objective as

$$\mathbb{E}[Y_\pi] = \sum_{s,x,a} p(s,x)\mu^Y(1,s,x)\mu^A_{\sigma_A}(a,s,x),$$

leading to the linear program

$$\begin{aligned}
&\max_{\mu^A_{\sigma_A}(a,s,x)} \sum_{s,x,a} p(s,x)\mu^Y(a,s,x)\mu^A_{\sigma_A}(a,s,x) \\
&s.t. \quad \sum_{a,x} p(s,x)\left(\mu^Y(a,s,x)\mu^A_{\sigma_A}(a,s,x) - p(s,x)\mu^Y(a,s,x)\mu^A_{\sigma_A}(a,s,x)\right) \\
&\qquad - \sum_{a,x} p(s',x)\left(\mu^Y(a,s',x)\sigma_A(s',x,a) - p(s,x)\mu^Y(a,s',x)\mu^A_{\sigma_A}(a,s',x)\right) \le \sqrt{\varepsilon} \\
&\forall s, s' < s, \\
&\qquad \sum_{a,x} p(s,x)\left(\mu^Y(a,s,x)\mu^A_{\sigma_A}(a,s,x) - p(s,x)\mu^Y(a,s,x)\mu^A_{\sigma_A}(a,s,x)\right) \\
&\qquad\quad - \sum_{a,x} p(s',x)\left(\mu^Y(a,s',x)\sigma_A(s',x,a) - p(s,x)\mu^Y(a,s',x)\mu^A_{\sigma_A}(a,s',x)\right) \ge -\sqrt{\varepsilon} \\
&\forall s, s' < s, \\
&\qquad \sum_a \mu^A_{\sigma_A}(a,s,x) = 1 \ \forall s,x, \quad \text{and} \quad \mu^A_{\sigma_A}(a,s,x) \ge 0 \ \forall a,s,x.
\end{aligned}$$

The last two constraints ensure that the policy is a valid conditional probability distribution.

7.4 Experiments

We evaluated the `EqB` and `ModBrk` constraint methods on the New York City Public School District (NYCSchools) dataset compiled in Kusner et al. (2019), which we augmented to include actions; and further tested `ModBrk` on the Infant Health and Development Program (IHDP) dataset, specifically the real-data example examining dosage effects, described in Section 6.2 of Hill (2011) (we could not use this dataset for `EqB` due to its reliance on parametric assumptions). The code reproducing the results will be made available. Next, we briefly discuss the datasets and defer further details to Section 7.5.

The Action-augmented NYCSchools Dataset. We adopted the same sensitive attribute and covariates as Kusner et al. (2019) and augmented the dataset with our own generated actions and outcomes. We created continuous actions corresponding to funding-level decisions as $A = (w_{SX}^T SX)^2 + \max(0, w_X^T X) + \mathcal{N}(0.5, .016)$, where w_{SX}, w_X are sampled uniformly in $[0, 1]$. We generated outcomes representing the percent of students who took college entrance exams (SAT/ACT) as $Y = 20E + \beta^\intercal [S, X, SX, A, XA, SA, SAX] + \gamma^\intercal [X, A, AX] + \mathcal{N}(1, .016)$, where E is the original percent of students taking the SAT/ACT exams (pre-college entry) appearing in the dataset.

The IHDP Dataset. The IHDP dataset describes a program that targeted low-birth-weight, premature infants, providing them with intensive high-quality child care and home visits from a trained provider. As continuous action A we used the self-selected number of participation days in the program. The outcome Y corresponds to the child's score attainment in cognitive tests at age three. As sensitive attribute S, we considered the mother's race (white vs. non-white). We re-reinterpret the original setting into a resource allocation problem as follows: rather than as a self-decision, we suggest thinking of the number of days in treatment as external, e.g., by assigning different individuals to different lengths of participation. In this case, the `ModBrk` constraint goal is to break the moderation of the allocation of days in the program by the group membership, such that the resource allocation policy is not responsible for the difference in test scores attained by different groups.

7.4.1 Results

We evaluated `EqB` and `ModBrk` by comparing these methods with: (1) optimizing the policy with no disparity constraint (Unconstrained), (2) optimizing the policy without using S (Drop S), (3) using constant actions (Const_A), and (4) using the baseline

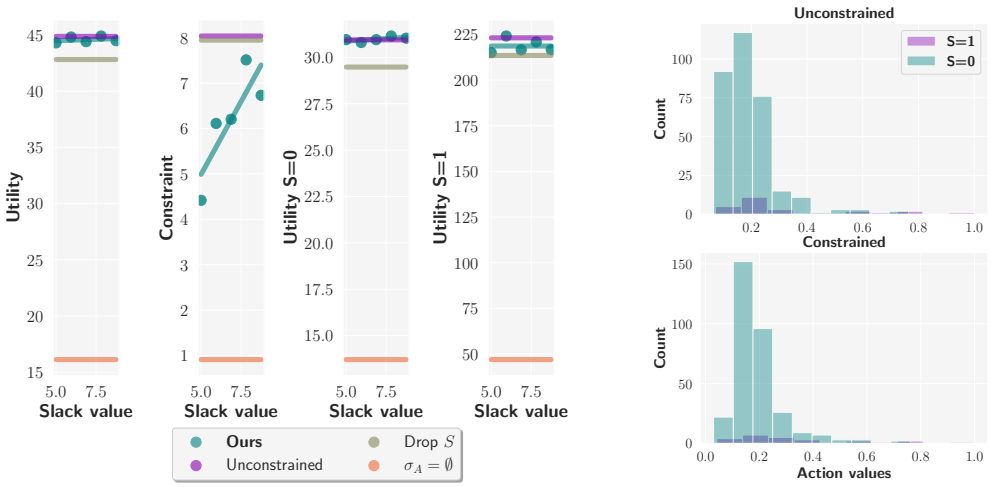


Figure 7.6: EqB on NYCSchools. (Left) influence of threshold for constraint on the value of the objective and fairness constraint violations. (Right, top) recommended actions by group for unconstrained optimization. (Right, bottom) recommended actions by group for constrained optimization, with slack of fairness violation set at 0. policy ($\sigma_A = \varnothing$). As discussed in Section 7.2, previous work on fair policy or action allocation differs from our setting and objective, and therefore cannot be meaningfully compared. We omit the Const_A baseline for EqB, as the IPW estimator is not suitable for *delta* distributions. The minimal ε slack values in the plots that follow correspond to the smallest constraint value that can be achieved with our optimization strategy; the maximal ε slack value is one that would closely match the constraint violation under unconstrained optimization.

The EqB Constraint Method. We present results for EqB on the NYCSchools dataset in Fig. 7.6. As we increase the tolerance on fairness violations, in the form of higher slack value ε in Eq. 7.4, we observe an increase in constraint value with almost no change in overall utility $\mathbb{E}[Y_{\sigma_A}]$. Observing the utility values broken down by group membership ($\mathbb{E}[Y_{\sigma_A}|S = 1]$ vs. $\mathbb{E}[Y_{\sigma_A}|S = 0]$), we see also no significant change in utility coming from either group. Here $S = 1$ corresponds to a majority-white student body in the NYCschools dataset and action corresponds to (simulated) government funding level decisions. The result shows that our method learns a policy that assigns actions ensuring equal benefit without sacrificing the utility of either group. Note that we also see an unusually high estimate of utility for $S = 1$. This could be explained by the IPW estimator not extrapolating well to unseen data, as the $S = 1$ group only consists of 7% of an already small dataset. On the right-hand side of Fig. 7.6 we observe similar recommended action histograms for the unconstrained (top right) and constrained (bottom right) cases. This shows that our method decreases disparity with similar budget constraints. Conceptually, this is possible because we

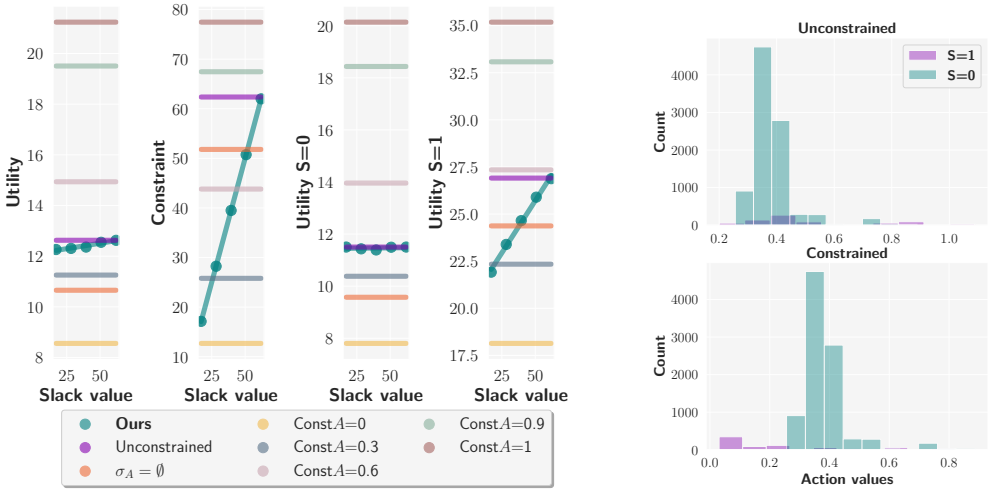


Figure 7.7: `ModBrk` on NYCSchools. (Left) influence of threshold for constraint on the value of the objective and fairness constraint violations. (Right, top) recommended actions by group for unconstrained optimization. (Right, bottom) recommended actions by group for constrained optimization, with slack of fairness violation set at 17.

are not attempting to reduce existing disparity from observed data. From a technical perspective, the similarity in the action histogram is partially due to the IPW objective which automatically aligns recommended actions with observed ones to achieve higher weighting. To validate our approach, we compute the ground truth constraint value through counterfactual realization of the policy model's predicted action mean in the data generating process and compute the distributional difference of $Y_{\sigma_A} - Y_{\varnothing}|S$ for different sensitive attribute groups. Notice also that our method succeeds in increasing the utility compared to the baseline policy. Although the baseline policy has the lowest constraint, this is expected as we are comparing distributions of $Y_{\sigma_A} - Y_{\varnothing} \,|\, S$, where $\sigma_A = \varnothing$. Dropping S has a slight increase in constraint and a decrease in utility compared to the unconstrained setting. This could be explained, as our policy model performs better by taking into account S to break the indirect association between S and Y. On the other hand, this constraint is an approximation and we would like to extend beyond its computation over a grid of z values, see Section 7.3.1. Moreover, our method shows a promising result: we can ensure `EqB` with a similar budget and no sacrifice from either group.

The `ModBrk` Constraint Method. The results for `ModBrk` on the NYCSchools and IHDP datasets are presented in Fig. 7.7 and in Fig. 7.8 respectively[6]. For both datasets, as we increase the tolerance on fairness violations, in the form of

[6]Dropping S had no influence on utility and constraint values, so we did not include it in the figures above to avoid confusion due to overlap of values.

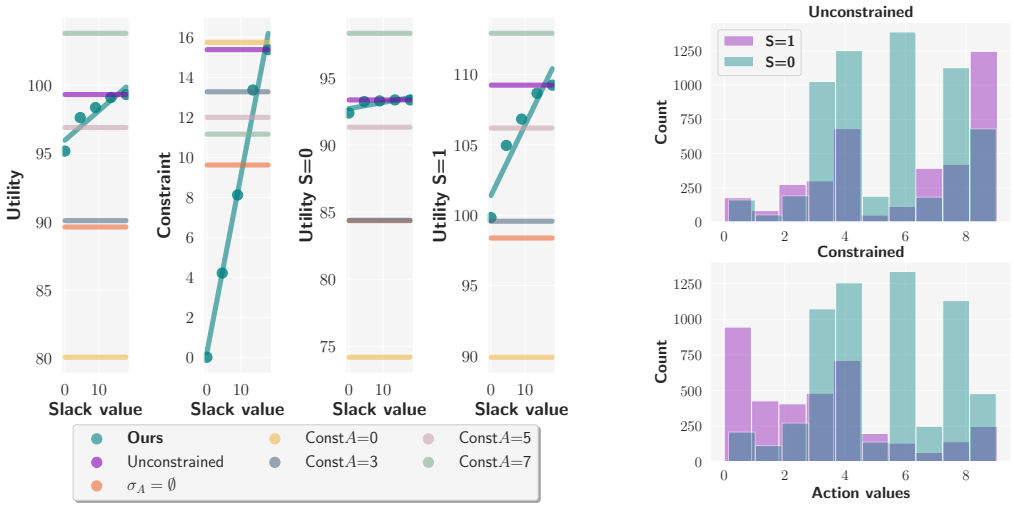


Figure 7.8: IHDP `ModBrk` constraint. (Left) influence of threshold for constraint on the value of the objective and fairness constraint violations. (Right, top) recommended actions by group for unconstrained optimization. (Right, bottom) recommended actions by group for constrained optimization, with slack of fairness violation set at 0.01.

higher slack value ε in Eq. 7.3, we see a major increase in constraint value while allowing some increase in utility $\mathbb{E}[Y_{\sigma_A}]$. Observing the utility values broken down by group membership ($\mathbb{E}[Y_{\sigma_A}|S=1]$ vs. $\mathbb{E}[Y_{\sigma_A}|S=0]$), we see most of this increase of utility driven by the privileged group, $S=1$ (a majority-white student body in the NYCSchools dataset, and white mothers for the IHDP dataset). This is to be expected given that we are trying to minimize interactions involving $S=1$ and A, i.e. where the membership in the privileged group inflates utility values, via higher g values. These higher g values for the $S=1$ group also translate into higher utility values for $S=1$ according to the baseline policy, as can be seen from the orange line, indicating baseline policy actions in the corresponding figures. Notice that the choice of an appropriate slack ε here is a matter of trade-off based on user preferences of utility vs. constraint value, and will depend on the specific dataset. We can gain deeper insight into the working of our approach by inspecting the histogram of recommended actions at the end of the policy model training. For both datasets, we see that applying the constraint with a small ε value results in mapping more $S=1$ members to lower action values, compared to the $S=0$ group, indicating that to enforce the constraint more tightly means allocating less intervention budget to the privileged group while giving out as many high actions to $S=0$ as possible under the action clipping setting described previously. Notice that in both settings we succeed in increasing the utility compared to the baseline policy. Recall that we employ action clipping to ensure some

overlap with the baseline actions for estimation purposes, but this means that we are bound to some extent to the interval of actions seen for each s, x in the baseline policy (this trade-off can be explored via the choice of η value, see Section 7.3.2). One could also increase η to achieve higher utility at the cost of increased budget and sacrifice in coverage. This is also why we cannot achieve utility values that are as high as the highest Const_A in the figures; we ask how best to distribute actions within the effective budget.

7.5 Additional Dataset Details

To the best of our knowledge, both datasets were made public under CC0 license.

The Action-augmented NYCSchools Dataset. As sensitive attribute S, we consider the majority race of students in the school (white vs. non-white), while as covariates X we consider: (1) if the school offers advanced placement or international baccalaureate classes, (2) whether the school offers calculus courses, and (3) the number of full-time counselors employed (fractional values indicate part-time work). We then create continuous actions corresponding to funding level decisions as $A = (w_{SX}^T SX)^2 + \max(0, w_X^T X) + \mathcal{N}(0.5, \sigma)$, where w_{SX}, w_X are sampled uniformly in $[0, 1]$, and $\sigma = 0.4$ (to allow variability, but not encourage negative values, given that the mean is set at 0.5). We then scale A to lie in [0,1]. We generate the outcome Y that represents the percent of students who take the college entrance exams in the US (SAT/ACT), as $Y = 20E + \beta^\intercal [S, X, SX, A, XA, SA, SAX] + \gamma^\intercal [X, A, AX] + \mathcal{N}(1, \sigma)$, where E is the original percent of students taking the SAT/ACT exams (pre-college entry) appearing in the dataset. We multiply E by 20 to bring it within the same scale as the other components in the structural equation. β and γ are sampled from a normal distribution, with higher mean values for coefficients affecting terms involving A, to ensure the outcome is noticeably responsive to changes in A. In the original dataset, E is the outcome and therefore can be interpreted as a function of S and X. The original dataset includes 345 datapoints. For the `ModBrk` constraint, we use a bootstrapped version (by re-sampling S, X, E with replacement) which increases the number of datapoints to 20,000.

The IHDP Dataset. The IHDP dataset consists of data from a program that targeted low-birth-weight, premature infants, providing them with intensive high-quality child care and home visits from a trained provider. The continuous action

A is self-selected number of participation days in the program. The outcome Y is child score attainment in cognitive tests at age three. As sensitive attribute S, we consider mother's race (white vs. non-white). After removing duplicate encoding and NaN values, we are left with 31 covariates, 22 categorical and 9 continuous. We use a bootstrapped version (by re-sampling S, X with replacement, and obtaining corresponding A, Y predictions from two black-box models trained on the original S, Y in the dataset) which increases the number of samples to 20,000.

We re-reinterpret the original setting into a resource allocation problems as follows: rather than as a self-decision, we suggest to think of the number of days in treatment as external, e.g., voucher allocation or by assigning different individuals to different lengths of participation. In this case, the `ModBrk` constraint goal is to break the moderation of the allocation of days in program by the group membership (withe vs non-white mothers), such that the resource allocation policy is not responsible for the difference in test scores attained by different groups.

7.6 Additional Experimental Details

All experiments are implemented using the Pytorch library (Paszke et al., 2019). The `EqB` experiments were run on a standard personal Mac, using CPUs. The `ModBrk` experiments were run on a single Microsoft Azure NVISIA Tesla P100 GPU server (NC6s v2 instance). We provide the corresponding environment setups in the accompanying code.

After performing hyper-parameter tuning, we ran the experiments with the following configurations.

7.6.1 `EqB` NYCSchools

Our stage I model consists of two MLPs, where $\text{MLP}^A_\varnothing$ learns $\mu^A_\varnothing$ and MLP^Y learns μ^Y as described in Section 7.3.1. Both MLPs share same structure as in we use one linear layer and one ReLu layer Agarap (2018), with the hidden dimension set to 128 (for each component). We train the stage I model for 500 epochs for MLP^Y and 1000 epochs for $\text{MLP}^A_\varnothing$, with a learning rate set at 0.01 and 0.0001 respectively, and the optimizer is Adam Kingma and Ba (2015). We verify the training of our stage I models by checking R^2 score with predicted and the target. We also check the predictions of the trained model MLP^Y for constant actions against the original model's. We omit to check baselines with higher constant action values due to IPW estimator is hard to extrapolate to unseen data.

Stage II model was defined with one linear layer and one ReLU layer, with the hidden dimension set to 10 (for each component). In order to speed up its convergence, we standardize the output of Stage II model with constant values. We train stage II model for 1000 epochs, with learning rate 0.01.

7.6.2 `ModBrk` NYCSchools

Our stage I model is an MLP composed of additive f, g, h components, as described in Section 7.3.2. For this setting, we use three linear layers, and two ReLu Non-linearities Agarap (2018), with the hidden dimension set to 256 (for each component). We train the stage I model for 3000 epochs, with a learning rate set at 0.005, and the optimizer is Adam Kingma and Ba (2015). We verify the training of our stage I model by checking explained variance and R^2 scores on a held-out set. We also check the predictions of the trained model for constant actions against the original model's.

Stage II model was defined with 3 linear layers and 2 ReLU nonlinearities, with a single batchnorm layer following the first input layer. We also include a shifted sigmoid at the end of the network to enforce our action clipping. For this setting, the hidden dimension was set to 64, the learning rate was 0.001, and we used the Adam optimizer Kingma and Ba (2015). Some scheduling was included for smoother training. For both, we use a single batch of size 20,000.

7.6.3 `ModBrk` IHDP

We use the same model architectures as above for this dataset, with the following differences. For the stage I model, we use only 2 linear layers and 1 ReLU non-linearity Agarap (2018), with the hidden dimensions set at 512. This model was trained for 1000 epochs. We compare constant action predictions compared to those of a standard black box model trained for $E[Y|S, X, A]$. The learning rate used for stage I was 0.001, with a small weight decay of 0.1 for regularization.

For stage II, we use 50 hidden dimensions, a learning rate of 0.0005 and no scheduler. We train for 3000 epochs.

7.7 Conclusion

We introduced a framework to learn fair policies given access to observational data and an action space in question. Taking a pragmatic view, we asked what is the best utility that can be achieved with the provided action space, while controlling two notions

of disparity: one focusing on ensuring equal benefit with respect to a baseline policy, and the other focusing on mitigating a possible moderation effect involving group membership and the policy. We see this work as a first conceptual contribution in defining pragmatic fair impact policies and envisage various possible future directions, including extending the proposed methods beyond binary groups, to a multi-stage policy setting, to handle unmeasured confounding and to online optimization. We would also like to extend the bounding method in the `EqB` constraint to non-parametric cases via Frechet bounds.

8 | Conclusion

The works presented herein all reflected two core principles.

- The application of ML methods to aid causal inference, and
- The promise causal inference has for ML and its trustworthiness: enabling fairer, more robust, and more interpretable systems.

In conclusion, this body of work has explored various aspects of causality and machine learning, with a focus on the estimation of causal effects, interpretability of models, and fairness in algorithmic decision-making. In Chapter 3, we presented a method for estimating the Average Treatment Effect (ATE) of a binary intervention under highly limited knowledge of the data generation process.

In Chapter 4, we investigated the effect of crude interventions on complex, high-dimensional objects, such as text, images, or networks. We proposed a two-step procedure that uncovers mediation mechanisms from the interventions, through the complex objects, and into the outcome of interest.

In Chapter 5, we proposed a unifying framework that connects the landscape of local explanations literature via causal concepts. We showed that different methods for model interpretability, including feature attributions, rule lists, and counterfactuals, can all be expressed in terms of the causal concepts of the probability of sufficiency and necessity of interventions. We also proposed a sound and complete algorithm for the enumeration of explanatory factors.

In Chapter 6, we examined the issue of fairness in algorithmic decision-making and proposed a framework that seeks a balance between error rate and calibration-style violations of fairness. We showed that the previous impossibility result – which stated that exact equal error rates and exact calibration cannot be simultaneously achieved in a less-than-perfect model – can be refined and used to achieve a better balance between error rate and calibration-style violations of fairness.

Finally, in Chapter 7, we expanded our view on fairness to policy optimization settings and proposed a framework called pragmatic fairness. It optimizes a policy for maximum expected utility while controlling disparity by controlling the policy itself. We introduced two approaches to disparity control, one that looks to avoid

introducing new differences in effects across groups by the introduction of the new policy compared to a baseline, and the other which aims to reduce existing disparities. Overall, this body of work contributes to the growing literature on causality and ML and provides insights into the estimation of causal effects, interpretability of models, and fairness in algorithmic decision-making.

While causal principles and approaches in ML offer greater clarity and principled approaches, they often rely on the knowledge of an underlying structure tying together variables that give rise to dependencies in datasets. This can become a restrictive assumption and inhibit their application in practice. Throughout this work we looked at what can be done under partial or minimal knowledge of an underlying causal structure, to try and accommodate different levels of causal knowledge and strength of assumptions, and make the adoption of methods that borrow from this line of thinking easier and more immediately possible.

8.1 Future Directions

It is an interesting and worthwhile undertaking to extend the works presented in this thesis by relaxing some of their underlying assumptions. The strength of assumptions needed for the successful application of causal inference methods in ML is one of the main challenges to its wider adoption and application.

How often can one identify, in practice, instrumental variables, and auxiliary variables (as required by the method introduced in Chapter 3)? How likely are assumptions around the lack of existence of unobserved confounders to hold in practice (as is required by the methods in Chapters 4 and 7)? The works included in this thesis were all aiming to relax assumptions needed by existing solutions, but the pursuit of looser assumptions is always much needed in the realm of causal inference.

Another direction for pursuit is the extension of parametric parts of works presented here (Chapter 7) to non-parametric directions. The Equal Benefit criterion suggested in that chapter invites a general question of interest - when is it worth it to make a parametric assumption that is justifiable to derive tighter bounds, versus the adoption of a non-parametric approach that would still produce appropriately tight bounds? These trade-offs would be of interest for this work and beyond.

8.2 Discussion

During the writing of this thesis, ML has kept making impressive leaps forward, into ever greater public attention of scrutiny. The dispersion and growing popularity of large language models (LLMs) and generative models shine a new and brighter light on some of the topics explored in this thesis.

On the one hand, the use of ML for causal inference tasks may get supercharged further. When it comes to causal discovery as a starting point for causal effect estimation, discovering already agreed-upon causal structures may become easier than ever with the help of LLMs (for an early investigation of this direction, see Kıcıman et al. (2023)).

Causal effect estimation and personalized treatments may be enhanced via the use of LLMs for data discovery, but also by the promise of generative models for the development of more faithful and varied synthetic data for evaluation and simulation.

On the other hand, the question of the relevance of causal methods for the enhancement of ML, especially its trustworthiness, takes on greater urgency.

Do some of the failure modes discussed in the introduction, e.g., those stemming from the violation of the i.i.d assumption during deployment, become less relevant when training sets encompass large swaths of all data available on the web? And how would this dynamic change in turn when, for the development of future models, more and more training data would be the hardly filtered output of models rather than unmediated human authors?

Reinforcement Learning from Human Feedback (Lambert et al., 2022) proved to be an important component in helping LLMs provide compelling outputs to users. One could view this technique as an attempt to infuse Large Language Models with some domain knowledge or basic understanding of a world model via what is acceptable to human labelers. Could this real-world knowledge be enhanced, and training methods become more efficient by the injection of more explicitly causal knowledge, or a clearer mechanistic understanding of cause-and-effect, instead of the costly and time-consuming process of gathering human feedback?

There are already well-documented examples of failure modes characteristic of this new generation of models. Some of them are variations on the same theme: LLMs still hallucinate content that seems likely but is not factual; generative models for image generation often struggle with the depiction of finer details of human features like hands or legs, or scripture elements like characters and letters. Although different

explanations of these phenomena may emerge, we remain curious about the clarity causal explanations may offer to these failure modes.

Finally, this new generation of models still leaves us wanting on some of the most crucial aspects of trustworthy ML:

- Models are still rather opaque (despite the feeling a 'chat with a machine' may give – how do we verify model outputs mechanistically with models of this scale?),
- Current models still pose great fairness risks (especially with training details remaining in obscurity,
- Evaluation at scale becoming more elusive with stochastic linguistic outputs,
- Training data not being fully representative globally),
- Data efficiency remains a faraway goal (when models in practice become larger in the number of parameters and scale of training data).

We very much look forward to the next generation of causal and trustworthy ML literature that will emerge to tackle these and other questions posed by these rapidly evolving technologies.

Bibliography

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif. Intell.*, 298:103502, 2021.

Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

Joshua D. Angrist and Guido W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995. doi: 10.1080/01621459.1995.10476535. URL `https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476535`.

J Angwin, J Larson, S Mattu, and L Kirchner. Machine bias. Technical report, ProPublica, 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2020.

Kellyn F. Arnold, Laurie Berrie, Peter W. G. Tennant, and Mark S. Gilthorpe. A causal inference perspective on the analysis of compositional data. *Int. J. Epidemiol.*, 49:1307–1313, 2020.

Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

Varsha Bansal. Uber's facial recognition is locking indian drivers out of their accounts, 2022. URL https://www.technologyreview.com/2022/12/06/1064287/ubers-facial-recognition-is-locking-indian-drivers-out-of-their-accounts/.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci.*, 113(27):7345–7352, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Solon Barocas, Andrew D Selbst, and Manish Raghavan. The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons. In *FAT**, pages 80–89, 2020.

Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Miruna Oprescu, and Vasilis Syrgkanis. Estimating the long-term effects of novel treatments. *Advances in Neural Information Processing Systems*, 34:2925–2935, 2021.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in Artificial Intelligence*, pages 606–615. PMLR, 2020.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M F Moura, and Peter Eckersley. Explainable machine learning in deployment. In *FAT**, pages 648–657, 2020.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly, 2009.

Martijn Blaauw, editor. *Contrastivism in Philosophy*. Routledge, New York, 2013.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? *CoRR*, abs/1912.01094, 2019. URL `http://arxiv.org/abs/1912.01094`.

Leo Breiman. Random Forests. *Machine Learning*, 45(1):1–33, 2001. ISSN 1098-6596. doi: 10.1017/CBO9781107415324.004.

Peter Bühlmann. Invariance, Causality and Robustness. *Statist. Sci.*, 35(3):404–426, 2020.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Robin Burke. Multisided fairness for recommendation, 2017.

Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 202–214. PMLR, 23–24 Feb 2018. URL `https://proceedings.mlr.press/v81/burke18a.html`.

Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3995–4004, 2017.

L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth K. Vishnoi. An algorithmic framework to control bias in bandit-based personalization, 2018.

L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 319–328, New York, NY, USA, 2019. Association for Computing Machinery. ISBN

9781450361255. doi: 10.1145/3287560.3287586. URL https://doi.org/10.1145/3287560.3287586.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning, 2015.

Krzysztof Chalupka, Tobias Bischoff, Pietro Perona, and Frederick Eberhardt. Unsupervised discovery of el nino using causal feature learning on microlevel climate data, 2016a.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Multi-level cause-effect systems. In *Artificial Intelligence and Statistics*, pages 361–369, 2016b.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44:137–164, 2017.

Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective, 2019.

Hugh Chen, Joseph D Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

Irene Chen, Fredrik D. Johansson, and David Sontag. Why is my classifier discriminatory?, 2018.

Minmin Chen, Kilian Q Weinberger, and John Blitzer. Co-training for domain adaptation. In *Nips*, volume 24, pages 2456–2464. Citeseer, 2011.

Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments, 2019.

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.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Alex Chohlas-Wood, Madison Coots, Emma Brunskill, and Sharad Goel. Learning to be fair: A consequentialist approach to equitable decision-making. *CoRR*, abs/2109.08792, 2021.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

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.

Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10093–10100, 2020.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning*, pages 2185–2195. PMLR, 2020.

Alexander D'Amour and Alexander Franks. Deconfounding scores: Feature representations for causal effect estimation with weak overlap. *arXiv preprint arXiv:2104.05762*, 2021.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women, 2018. URL `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`.

A Philip Dawid et al. Fundamentals of statistical causality. *RSS/EPSRC Grad Train Progr*, 279:1–94, 2007.

Philip Dawid. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1):39–77, 2021.

Arnoud A. W. M. de Kroon, Danielle Belgrave, and Joris M. Mooij. Causal discovery for causal bandits utilizing separating sets. *arXiv*:2009.07916, 2020.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pages 592–603, 2018.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Convergent algorithms for (relaxed) minimax fairness. *arXiv preprint arXiv:2011.03108*, 2020.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *CoRR*, abs/2108.04884, 2021. URL `https://arxiv.org/abs/2108.04884`.

M Donini, L Oneto, S Ben-David, J Shawe-Taylor, and M Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, volume 32. Neural Information Processing Systems (NIPS), 2018.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2803–2813. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/dutta20a.html`.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, pages 214 – 226, 2012.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133. PMLR, 2018.

Daniel Eaton and Kevin Murphy. Exact bayesian structure learning from uncertain interventions. In *Artificial intelligence and statistics*, pages 107–114, 2007.

Michael D. Ekstrand and Daniel Kluver. Exploring author gender in book rating and recommendation, 2020.

D. Entner, P. Hoyer, and P. Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. *JMLR W&CP: AISTATS 2013*, 31:256–264, 2013a.

Doris Entner, Patrik Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Artificial Intelligence and Statistics*, pages 256–264, 2013b.

C. Fernández-Loría, F. Provost, and X. Han. Explaining data-driven decisions made by AI systems: The counterfactual approach. *arXiv* preprint, 2001.07417, 2020.

Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *SIGMOD*, 2021.

Juan L. Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. *Advances in Neural Information Processing Systems*, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

A. Gelman. How to think about instrumental variables when you get confused. `https://statmodeling.stat.columbia.edu/2009/07/14/how_to_think_ab_2/`, 2009. Accessed: 02-02-2021.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, 2006.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural information Processing Systems*, 2019.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Mark Goldstein, Jörn-Henrik Jacobsen, Olina Chau, Adriel Saporta, Aahlad Manas Puli, Rajesh Ranganath, and Andrew Miller. Learning invariant representations with missing data. In *Conference on Causal Learning and Reasoning*, pages 290–301. PMLR, 2022.

Sachin Grover, Chiara Pulice, Gerardo I. Simari, and V. S. Subrahmanian. Beef: Balanced English explanations of forecasts. *IEEE Trans. Comput. Soc. Syst.*, 6(2): 350–364, 2019.

P. Richard Hahn, Jared S. Murray, and Carlos Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects, 2019.

Joseph Y Halpern. *Actual Causality*. The MIT Press, Cambridge, MA, 2016.

Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. Part I: Causes. *Br. J. Philos. Sci.*, 56(4):843–887, 2005a.

Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. Part II: Explanations. *Br. J. Philos. Sci.*, 56(4):889–911, 2005b.

Karen Hao. Facebook's ad algorithms are still excluding women from seeing jobs, 2021. URL `https://www.technologyreview.com/2021/04/09/1022217/facebook-ad-algorithm-sex-discrimination/`.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf`.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models, 2018.

Miguel A Hernán and James M Robins. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. Taylor & Francis, 2019. ISBN 9781420076165. URL `https://books.google.co.il/books?id=_KnHIAAACAAJ`.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Advances in Neural Information Processing Systems*, 2020.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. Stat. Theory Appl.*, 6(2):65–70, 1979.

Nabil Hossain, John Krumm, and Michael Gamon. "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 133–142, 2019.

Wen Huang, Kevin Labille, Xintao Wu, Dongwon Lee, and Neil Heffernan. Achieving user-side fairness in contextual bandits. *arXiv preprint arXiv:2010.12102*, 2020.

Wen Huang, Lu Zhang, and Xintao Wu. Achieving counterfactual fairness for causal bandit. *arXiv preprint arXiv:2109.10458*, 2021.

Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pages 217–224, 2006.

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5 (9):1–10, 2010.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *AAAI*, pages 1511–1519, 2019.

Steven J Ingels. *National Education Longitudinal Study of 1988: Base Year: Parent Component Data File User's Manual*. US Department of Education, Office of Educational Research and Improvement, 1990.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1617–1626. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/jabbari17a.html`.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. 2020.

John W. Jackson. On the interpretation of path-specific effects in health disparities research. *Epidemiology*, 29(4):517–520, 2018.

John W. Jackson. Meaningful causal decompositions in health equity research: definition, identification, and estimation through a weighting framework, 2020.

John W. Jackson and Tyler J. VanderWeele. Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology (Cambridge, Mass.)*, 29 (6):825, 2018.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem, 2019.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 862–872. PMLR, 22–25 Jul 2020. URL `https://proceedings.mlr.press/v115/jiang20a.html`.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *arXiv preprint arXiv:1605.07139*, 2016.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Meritocratic fairness for infinite and contextual bandits. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163, 2018.

Jean Kaddour, Qi Liu, Yuchen Zhu, Matt J. Kusner, and Ricardo Silva. Graph intervention networks for causal effect estimation. *arXiv* preprint, 2106.01939, 2021.

Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

Daniel Kahneman and Dale T. Miller. Norm theory: Comparing reality to its alternatives. *Psychol. Rev.*, 93(2):136–153, 1986.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*, 2020.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020a.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions, 2020b.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *arXiv preprint arXiv:2006.06831*, 2020c.

Michael Kearns and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, 2019.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, 30:656–666, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2015. URL `http://arxiv.org/abs/1412.6980`. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-029-3. doi: 10.4230/LIPIcs.ITCS.2017.43. URL `http://drops.dagstuhl.de/opus/volltexte/2017/8156`.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27, 2018.

Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14369–14379. Curran Associates, Inc., 2019. URL `http://papers.nips.cc/paper/\9581-characterization-and-learning-of-causal-graphs-\with-latent-variables-from-soft-interventions.pdf`.

Ronny Kochavi and Barry Becker. Adult income dataset, 1996. URL `https://archive.ics.uci.edu/ml/datasets/adult`.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 202–207. AAAI Press, 1996.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Indra Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *ICML*, pages 5491–5500, 2020.

M. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30:4066–4076, 2017a.

Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. Making decisions that reduce discriminatory impacts. In *International Conference on Machine Learning*, pages 3591–3600, 2019.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017b.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality, 2023.

Himabindu Lakkaraju and Osbert Bastani. "How do I fool you?": Manipulating user trust via misleading black box explanations. In *AIES*, pages 79–85, 2020.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In *AIES*, pages 131–138, 2019.

Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*, 2022. https://huggingface.co/blog/rlhf.

Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, 65(7):2966–2981, 2019.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1), 2016.

Finnian Lattimore, Tor Lattimore, and Mark D. Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, pages 1181–1189, 2016.

S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

Claire Lazar Reich and Suhas Vijaykumar. A Possibility in Algorithmic Fairness: Can Calibration and Equal Error Rates Be Reconciled? In Katrina Ligett and Swati Gupta, editors, *2nd Symposium on Foundations of Responsible Computing (FORC 2021)*, volume 192 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 4:1–4:21, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-187-0. doi: 10.4230/LIPIcs.FORC.2021.4. URL `https://drops.dagstuhl.de/opus/volltexte/2021/13872`.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Sanghack Lee and Elias Bareinboim. Structural Causal Bandits: Where to Intervene? In *Advances in Neural Information Processing Systems*, volume 31, pages 2568–2578, 2018a.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems*, pages 2568–2578, 2018b.

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4164–4172, Jul. 2019a. doi: 10.1609/aaai.v33i01.33014164. URL https://ojs.aaai.org/index.php/AAAI/article/view/4320.

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 4164–4172, 2019b.

Sanghack Lee and Elias Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8565–8576. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/61a10e6abb1149ad9d08f303267f9bc4-Paper.pdf.

Felix Leeb, Yashas Annadani, Stefan Bauer, and Bernhard Schölkopf. Structural autoencoders improve representations for generation and transfer, 2020.

E.L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, New York, Third edition, 2005.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference for Regression. *J. Am. Stat. Assoc.*, 113(523): 1094–1111, 2018.

Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistics Society, Series B*, 83: 911–938, 2021.

Sam Levin. A beauty contest was judged by ai and the robots didn't like dark skin. *The Guardian*, 8:2016, 2016. URL https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people.

David Lewis. Causation. *J. Philos.*, 70:556–567, 1973.

Peter Lipton. Contrastive explanation. *Royal Inst. Philos. Suppl.*, 27:247–266, 1990.

Zachary Lipton. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, 2018.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158, 2018.

Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060. PMLR, 2019.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C. Parkes. Calibrated fairness in bandits, 2017.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations, 2019.

Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too relaxed to be fair. In *International Conference on Machine Learning*, pages 6360–6369. PMLR, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774. 2017.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL*, pages 142–150, 2011.

J.L. Mackie. Causes and conditions. *Am. Philos. Q.*, 2(4):245–264, 1965.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Andrej Aderhold, Kyle R Allison, Richard Bonneau, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nat. Methods*, 9(8):796–804, 2012.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR, 2020.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR, 23–24 Feb 2018. URL `https://proceedings.mlr.press/v81/menon18a.html`.

Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *CD-MAKE*, pages 17–38. Springer, 2020.

George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.*, 101(2):343–352, 1955.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Münich, 2021. URL `https://christophm.github.io/interpretable-ml-book/`.

J. M. Mooij, D. Janzing, J. Peters, and B. Schölkopf. Regression by dependence minimization and its application to causal inference. *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*, pages 745–752, 2009.

Ramaravind K. Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. *arXiv* preprint, 2011.04917, 2020a.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAT**, pages 607–617, 2020b.

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 2020.

R. Nabi and I. Shpitser. Semiparametric causal sufficient dimension reduction of high dimensional treatment. *arXiv* preprint, 1710.06727, 2017.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *AAAI Conference on Artificial Intelligence*, pages 1931–1940, 2018.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *International Conference on Machine Learning*, pages 4674–4682, 2019.

Nina Narodytska, Aditya Shrotri, Kuldeep S Meel, Alexey Ignatiev, and Joao Marques-Silva. Assessing heuristic machine learning explanations with model counting. In *SAT*, pages 267–278, 2019.

Kirtan Padh, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. Addressing fairness in classification with a model-agnostic multi-objective algorithm. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 600–609. PMLR, 27–30 Jul 2021. URL `https://proceedings.mlr.press/v161/padh21a.html`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035. 2019.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y. Narahari. Achieving fairness in the stochastic multi-armed bandit problem, 2020.

J. Pearl. Myth, confusion, and science in causal analysis. *UCLA Cognitive Systems Laboratory, Technical Report (R-348)*, 2009a.

Judea Pearl. [bayesian analysis in expert systems]: comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

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

Judea Pearl. Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, 121:93–149, 1999.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Judea Pearl. Causal inference in statistics: An overview. 2009b.

Judea Pearl. *Causal Diagrams and the Identification of Causal Effects*, page 65–106. Cambridge University Press, 2009c. doi: 10.1017/CBO9780511803161.005.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

J. Peters, P. Buhlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals (with discussion). *Journal of the Royal Statistical Society: Series B*, 78:947–1012, 2016a.

J. Peters, D. Janzig, and B. Scholkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016b. doi: 10.1111/rssb.12167. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12167`.

Picker Institute Europe. National Health Service national staff survey, 2014. 2015. URL `http://doi.org/10.5255/UKDA-SN-7776-1`.

Drago Plecko and Elias Bareinboim. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf`.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

Yanou Ramon, David Martens, Foster Provost, and Theodoros Evgeniou. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C. *Adv. Data Anal. Classif.*, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, pages 1527–1535, 2018a.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *ACL*, pages 856–865, 2018b.

Thomas S Richardson and James M Robins. Single world intervention graphs: a primer. In *Second UAI workshop on causal structure learning, Bellevue, Washington*. Citeseer, 2013a.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013b.

James M Robins and Thomas S Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, 84:103–58, 2010.

Tom Ron, Omer Ben-Porat, and Uri Shalit. Corporate social responsibility via multi-armed bandits. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 26–40, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445868. URL https://doi.org/10.1145/3442188.3445868.

P. Rosenbaum. *Observation and Experiment: an Introduction to Causal Inference*. Harvard University Press, 2017.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*, 2017.

D. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, pages 322–331, 2005.

Michael Ruchte and Josif Grabocka. Efficient multi-objective optimization for deep learning. *CoRR*, abs/2103.13392, 2021. URL https://arxiv.org/abs/2103.13392.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28, 2019.

Bernhard Schölkopf, Alexander J Smola, and Klaus-Robert Müller. Kernel Principal Component Analysis. In *Advances in Kernel Methods: Support Vector Learning*, pages 327–352, Cambridge, MA, 1999. MIT Press.

Candice Schumann, Zhi Lang, Nicholas Mattei, and John P. Dickerson. Group fairness in bandit arm selection, 2022.

Rajen Shah and Jonas Peters. The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *Ann. Statist.*, 48(3):1514–1538, 2020.

Lloyd Shapley. A value for n-person games. In *Contributions to the Theory of Games*, chapter 17, pages 307–317. Princeton University Press, Princeton, 1953.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006a.

Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pages 437–444, 2006b.

Ilya Shpitser and Judea Pearl. Complete Identification Methods for the Causal Hierarchy. *J. Mach. Learn. Res.*, 9:1941–1979, 2008.

Ricardo Silva and Robin Evans. Causal inference through a witness protection program. *The Journal of Machine Learning Research*, 17(1):1949–2001, 2016.

Julien Simon. Large language models: A new moore's law, 2021.

Tom Simonite. The best algorithms struggle to recognize black faces equally. wired, july 22, 2019. URL `https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/`.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, pages 4593–4605, 2019.

Apache SpamAssassin, 2006. URL `https://spamassassin.apache.org/old/publiccorpus/`. Accessed 2021.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search.* MIT Press, 2000a.

Peter Spirtes and Richard Scheines. Causal inference of ambiguous manipulations. *Philos. Sci.*, 71:833–845, 2004.

Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search.* Springer-Verlag, New York, 1993.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000b.

John D Storey. The optimal discovery procedure: A new approach to simultaneous significance testing. *J. Royal Stat. Soc. Ser. B Methodol.*, 69(3):347–368, 2007.

Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *ACM*, New York, 2019.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/sundararajan17a.html`.

KCB Tan et al. Appropriate body-mass index for asian populations and its implications for policy and intervention strategies. *The lancet*, 2004.

Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 2020.

H Theil. Economic forecasts and policy, 1958.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Ann. Math. Artif. Intell.*, 28(1-4):287–313, 2000.

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

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Ser. B Methodol.*, 58(1):267–288, 1996.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *FAT**, pages 10–19, 2019.

Mark J van der Laan and Richard JCM Starmans. Entering the era of data science: targeted learning and the integration of statistics and computational data analysis. *Advances in Statistics*, 2014, 2014.

T. VanderWeele and I. Shpitser. A new criterion for confounder selection. *Biometrics*, 64:1406–1413, 2011a.

Tyler VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015.

Tyler J. VanderWeele and Miguel A. Hernán. Causal inference under multiple versions of treatment. *J. Causal Inference*, 1:1–20, 2013.

Tyler J VanderWeele and Thomas S Richardson. General theory for interactions in sufficient cause models with dichotomous exposures. *Ann. Stat.*, 40(4):2128–2161, 2012.

Tyler J VanderWeele and James M Robins. Empirical and counterfactual conditions for sufficient cause interactions. *Biometrika*, 95(1):49–61, 2008.

Tyler J. VanderWeele and Ilya Shpitser. A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413, 2011b. ISSN 0006341X, 15410420. URL `http://www.jstor.org/stable/41434446`.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests, 2021.

Two Sigma Ventures. The promise and perils of large language models, 2022. URL `https://twosigmaventures.com/blog/article/the-promise-and-perils-of-large-language-models/`.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard J. Law Technol.*, 31(2):841–887, 2018.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation, 2020.

David S Watson and Luciano Floridi. The explanation game: a formal framework for interpretable machine learning. *Synthese*, 2020.

J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Trans. Vis. Comput. Graph.*, 26(1):56–65, 2020.

Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf`.

Janine Witte and Vanessa Didelez. Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal*, 61(5):1270–1289, 2019. doi: 10.1002/bimj.201700294. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201700294`.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York, 2003.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.

Philip G Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

Yongkai Wu, Lu Zhang, and Xintao Wu. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*, pages 3356–3362, 2019a.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, pages 3404–3414, 2019b.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression, 2020.

Forest Yang, Moustapha Cissé, and Sanmi Koyejo. Fairness with overlapping groups. *CoRR*, abs/2006.13485, 2020. URL `https://arxiv.org/abs/2006.13485`.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide

Web Conferences Steering Committee, apr 2017a. doi: 10.1145/3038912.3052660. URL https://doi.org/10.1145%2F3038912.3052660.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Yujia Li Max Welling Richard Zemel, Christos Louizos, and Kevin Swersky. The variational fair autoencoder. In *Proceedings of the international conference on learning representations*, 2016.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Junzhe Zhang and Elias Bareinboim. Online reinforcement learning for mixed policy scopes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=sjaQ2bHpELV.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning*, pages 819–827, 2013.

Xin Zhang, Armando Solar-Lezama, and Rishabh Singh. Interpreting neural network judgments via minimal, stable, and symbolic corrections. In *Advances in Neural Information Processing Systems*, page 4879–4890, 2018.

X. Zheng, B. Aragam, P. Ravikumar, and E. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems 31*, pages 9472–9483, 2018a.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 9492–9503, Red Hook, NY, USA, 2018b. Curran Associates Inc.

Ziwei Zhu, Xia Hu, and James Caverlee. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1153–1162, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3271795. URL https://doi.org/10.1145/3269206.3271795.