Data Quality in Causal Machine Learning with Applications to Algorithmic Fairness

Jake Howarth Fawkes

St Hugh's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2025

Statement of Originality

I hereby declare that except where specific reference is made to the work of others, the content of this thesis is my own work and has not been submitted in whole or in parts for any other degree or qualification. This thesis is my own work unless otherwise stated in the authorship form at the end of the chapters.

Jake Fawkes
Hillary 2025

I dedicate this thesis to Hilary for patiently listening to my ramblings and constantly supporting me throughout the years this took to write. In my opinion, you are the best of the world.

Acknowledgements

I want to begin by giving my deepest thanks to my supervisors, Robin J. Evans and Dino Sejdinovic. Robin, I am very grateful to you for believing in me and teaching me all the way from my first statistics and probability courses to the end of a PhD. I have learned so much under your guidance and I really appreciate the patience you had for me, especially given my lack of organisation. During my PhD, I always felt I had the right balance of intellectual freedom and support, and I am grateful to you for that. Dino, thank you for showing me first hand how much fun research can be. Working with you is a great reminder of how much passion and enthusiasm you can have for Math and ML. It is a great example for all who are lucky enough to work with you. I strongly hope I have the chance to cross paths with you both again in the future.

I would like to thank others who have guided me through projects over the years, specifically Zachary C. Litpon, Amartya Sanyal, Chris Holmes, and Uri Shalit. You were all much too generous with your time and thoughts; it was a privilege to learn from you. I also want to thank the groups who were kind enough to host me at various points over my PhD. Thank you so much to Krikamol Muandet for welcoming me in Saarbrücken and being a shining example of academic curiosity. Thank you to Ali Shah, Robert Grout and the rest of the team at Accenture for giving me my first internship. Thank you to Ciarán M. Gilligan-Lee, Michael O' Riordan, Thanos Vlontzos, and Oriol Corcoll for supervising me during my time at Spotify's advanced causal inference lab. You all created an incredibly friendly environment which was a pleasure to research in. Thank you to Jason Harford, Kristina Ulicna, and the rest of Valence labs for hosting me as an intern. Working with you all reinjected me with a passion for research that was dimming at the end of my PhD.

Next and very importantly, Coucou, Shasha. Meeting, working, and laughing with you was one of the true highlights of my PhD. I am still trying to figure out a way to get you to work with me again so that I can spend more of my days talking to you. I want to say a big thank you to Nic Fishman for being a good friend and an incredible collaborator. I always greatly enjoyed hearing your thoughts on the field and discussing

things with you. Lucile, thank you very much for teaching me the Lucile method of being such a fun collaborator that people are desperate to work with you. Thank you to all the others who I have been lucky enough to work with over my PhD, specifically Robert, Desi, Omri, and Mel. It was a pleasure working with you all. Finally, thank you to the rest of Dino's group in Jef, Alan, and Veit for the many interesting ideas you all introduced me to over the PhD.

I was fortunate to have a number of great friendships throughout my PhD, which made the many years it took to complete much easier. Thank you to Patrick, Joris, Florence, JT, and the many others I was lucky enough to meet for being constant sources of fun and friendship away from study. Thank you to the other members of the best office in the stats department for making coming to work so much fun, specifically Eugenio, Dan, Xi, Linying, Chris, and Vik.

Finally, thank you so much to my family for being there for me always. Ellie, you are a constant source of joy to me, so thank you for being you. I love you very much. To Laura and the rest of the Cockhills, thank you so much for welcoming me with open arms into your family; I am forever grateful to you all for that. This feels like a long journey since getting the acceptance letter on a sunny Bath day, but you have all been there every step of the way. To my Dad, thank you for showing me what it is to be intellectually curious. To my Mum, thank you for championing and encouraging me for as long as I can remember. As far as I am concerned, this thesis is as much your achievement as mine. I love you.

Contents

1 Introduction **3**

1.1 Background on Causality . . . 3

1.1.1 Structural Causal Models . . . 4

1.1.2 Potential Outcomes and the Interventional Distribution . . . 6

1.1.3 Counterfactual Distributions . . . 6

1.1.3.1 Debate over Counterfactual Dependencies . . . 7

1.1.4 Causal Structure Learning . . . 8

1.2 Causal Approaches to Data Quality . . . 10

1.2.1 Unmeasured Confounding, Hidden Variables, and Identifiability 10

1.2.1.1 Hidden Variables and Marginalisation . . . 12

1.2.1.2 Identifiability of Causal Effects with Hidden Variables 14

1.2.2 Selection Bias . . . 18

1.2.2.1 Recovering Observational Distributions . . . 18

1.2.2.2 Recovering Causal Effects . . . 20

1.2.3 Multi-Environment Data . . . 21

1.3 Algorithmic Fairness . . . 24

1.3.1 Non-Causal Fairness Methods . . . 25

1.3.2 Causal Fairness Methods . . . 27

1.3.2.1 Counterfactual Fairness . . . 28

1.3.2.2 Path-Specific Fairness . . . 29

1.4 Thesis outline . . . 31

2 The Hardness of Validating Observational Studies with Experimental Data 35

2.1 Introduction . . . 37

2.2 Background and Notation . . . 38

2.2.1 Notation . . . 38

2.2.2 Objectives and Assumptions . . . 39

2.2.3 Related work bounding $\Delta(\mathbf{x})$ . . . 41

2.3 The Hardness of Validating Observational Studies . . . 41

2.3.1 Testing Notation and Background . . . 42

2.3.2 Setting the Testing Problem for Unmeasured Confounding . . . 43

2.3.3 Limits on Testing . . . 45

2.3.4 Implications for Sensitivity Models . . . 46

2.4 Pseudo Outcome Gaussian Processes and Uniform Error Bounds . . . 47

2.4.1 Pseudo Outcome Regression with Gaussian Processes . . . 48

2.4.2 Uniform Error Bounds . . . 49

2.5 Experiments . . . 50

2.5.1 Simulated Experiment . . . 51

2.5.1.1 Results . . . 51

2.5.2 Semi-Synthetic Experiments . . . 52

2.5.2.1 Results . . . 52

2.5.3 Additional Results . . . 53

2.6 Conclusion . . . 53

Appendices **54**

2.A Causal Assumptions . . . 54

2.A.1 Constant CATE . . . 54

2.B Hardness of Testing . . . 54

2.C Gaussian Process . . . 57

2.C.1 Closed Form Posterior Expressions . . . 58

2.C.2 Closed Form Bounds . . . 59

2.C.3 LCM Kernel and Causal Multitask Kernel of Alaa and Van Der Schaar [4] . . . 60

2.D Experiment Details . . . 61

2.D.1 Model Tuning Details . . . 61

2.D.2 Simulation Details . . . 63

2.D.3 IHDP details . . . 63

2.E Additional Results . . . 64

2.E.1 Simulated Experiment Additional Results . . . 64

2.E.2 IHDP Experiment Additional Results . . . 64

2.E.3 Robustness Results . . . 65

2.F Uniform Error Bounds . . . 66

3 Is Merging Worth It? Securely Evaluating the Information Gain for Causal Dataset Acquisition **67**

3.1 Introduction . . . 69

3.2 Problem Statement, Assumptions &Notation . . . 71

3.3 Method . . . 72

3.3.1 Quantifying Data Merge Utility through Expected Information Gain . . . 73

3.3.2 EIG Targeting CATE Parameters . . . 74

3.3.3 Procedure and Model Classes . . . 75
3.4 Privacy . . . 77
3.5 Experiments &Results . . . 78
3.5.1 Illustrative experiment . . . 79
3.5.2 Ranking experiment . . . 81
3.5.3 Multi-Party Computation Experiments . . . 81
3.6 Related Work . . . 82
3.7 Discussion . . . 83

Appendices **84**

3.A Mathematical Details . . . 84
3.A.1 Algorithms for Computing EIG . . . 84
3.A.2 Differential Privacy Definition . . . 85
3.A.3 Sensitivity of Linear Statistic . . . 86
3.B Model Details . . . 86
3.B.1 Models via Sampling . . . 87
3.B.2 Closed Form Models . . . 88
3.B.2.1 Bayesian Polynomial Regression Derivations . . . 88
3.B.3 Causal Multitask Gaussian Processes [4] . . . 89
3.B.3.1 Expected Information Gain . . . 90
3.C Experimental Details . . . 91
3.D Further experimental results . . . 93
3.E Related work: further details . . . 94

4 Selection, Ignorability and Challenges with Causal Fairness 97

4.1 Introduction . . . 99

4.1.1 Paper Outline . . . 100

4.2 Preliminaries . . . 100

4.2.1 Notation and Definitions . . . 100

4.2.1.1 Causal Definitions . . . 100

4.2.1.2 Counterfactual Fairness . . . 101

4.2.2 Introducing the Law School Example . . . 102

4.2.3 Ancestral Closure of the Sensitive Attributes . . . 102

4.3 Ignorability . . . 103

4.4 Selection . . . 104

4.4.1 Theoretical Results . . . 104

4.4.2 When will Ignorability Under Selection Hold? . . . 106

4.4.3 Explicit Violation in certain cases . . . 107

4.5 Challenges for Causal Fairness . . . 108

4.5.1 Difficulties when no model in $\mathbb{M}_{S=1}$ fits . . . 109

4.5.2 Do Structural Counterfactuals from models in $\mathbb{M}_{S=1}$ have a causal interpretation? . . . 109

4.5.3 Counterfactual Fairness and Demographic Parity . . . 110

4.5.4 Path Specific Fairness . . . 111

4.5.4.1 Identifiability of Path Specific Effects . . . 112

4.6 Conclusion . . . 112

4.7 Acknowledgments . . . 112

Appendices **113**

4.A Race and Ignorability . . . 113

4.B Proof of Proposition 21 . . . 114

4.C Proof of Proposition 22 . . . 114

4.D Faithfulness in the Twin Network . . . 115

4.E Proof of Corollary 23 . . . 117

4.F Detailing the Calculations and Datasets . . . 117

4.F.1 Adult Dataset . . . 117

4.F.2 German Credit Dataset . . . 118

4.F.3 Law School Dataset . . . 118

4.G Proof of Lemma 24 . . . 118

4.H Proof of Corollary 25 . . . 119

5 The Fragility of Fairness: Causal Sensitivity Analysis for Fair Machine Learning **120**

5.1 Introduction . . . 122

5.2 Related work . . . 123

5.3 Measurement Biases . . . 124

5.3.1 Proxy Label Bias . . . 125

5.3.2 Selection Bias: . . . 125

5.3.3 Extra-Classificatory Policy Bias . . . 127

5.3.4 Cross-Dataset Analysis . . . 127

5.4 Graphical Causal Sensitivity Analysis . . . 128

5.4.1 Causal Background . . . 128

5.4.2 Partial Identification and Sensitivity Analysis . . . 129

5.4.3 Discrete Causal Sensitivity Analysis . . . 130

5.5 Causal Sensitivity Analysis for FairML . . . 131
5.6 Experiments . . . 132
5.6.1 Recreating Fogliato et al. [81] under varying assumptions . . . 133
5.6.2 Intersection of Biases . . . 134
5.6.3 Cross-Dataset Experiments . . . 135
5.7 Codebase and Web Interface . . . 136

Appendices **137**

5.A Disparity Metric Definitions . . . 137
5.A.1 Observational Metrics . . . 137
5.A.2 ECP Parity Metric Definitions . . . 137
5.B Technical Description . . . 138
5.B.1 Structural Causal Model definition . . . 138
5.B.2 Margnalisation in DAGs . . . 138
5.B.3 Alternative Causal Graphs for Proxy Bias . . . 140
5.C Additional Results . . . 141
5.C.1 Proxy Label Results . . . 141
5.C.1.1 Plots from Fogliato et al. [81] under varying assumptions 141
5.C.1.2 Proxy Identification Results . . . 141
5.C.2 Selection Results . . . 144
5.C.2.1 Selective labels under MNAR . . . 144
5.C.2.2 Selection and Proxy Plots . . . 144
5.C.3 ECP bias results . . . 145
5.C.3.1 ECP experimental set up . . . 145
5.C.4 Causal Fairness Experiments . . . 145
5.D Details of cross dataset bias analysis . . . 147

5.E Details of cross-dataset experiment . . . 149
5.E.1 Analysis of Results . . . 150
5.E.1.1 Correlational Plots . . . 150
5.E.1.2 Cross-Dataset Analysis . . . 152
5.F Codebase and web interface . . . 157
5.G Impact Statement . . . 158

6 Conclusion, Limitations and Future Outlook **159**
6.1 Conclusion . . . 159
6.1.1 The Hardness of Validating Observational Studies with Experimental Data . . . 159
6.1.2 Is merging worth it? Securely Evaluating the Information Gain for Causal Dataset Acquisition. . . . 160
6.1.3 Selection, Ignorability, and Challenges with Causal Fairness . . . 160
6.1.4 The Fragility of Fairness: Causal Sensitivity Analysis for Fair Machine Learning . . . 161
6.1.5 Concluding Remarks . . . 161

Bibliography **162**

Abstract

The success of modern machine learning methods can be attributed to three main factors: i) The availability of increasing amounts of high quality data, ii) the sustained growth in computational resources, iii) the invention of algorithms that can reap the gains of both of these simultaneously, being specifically tailored to consume ever larger datasets on cutting-edge hardware. In this thesis, we focus on the first question of data quality in the subfield of ML that intersects with another field, causality.

Causality aims to produce a precise mathematical formulation for the age-old question of cause and effect. In doing so, it provides a framework to formally reason about interventions, counterfactuals, and when valid causal inferences can be drawn. Graphical causal inference- the setting of this thesis - represents causal systems using directed graphs, with arrows from cause to effect. These are not just convenient visualisations but are a type of statistical model, which allow us to predict the distribution after intervening on variables. This makes them particularly well suited for reasoning about data quality, as statistical biases can be viewed as intervening or conditioning in these models.

Causal machine learning began by using machine learning methods to answer questions in causal inference, particularly in high dimensional, large data settings where modern ML excels. Later, a stream of work began in the opposite direction, aiming to understand how causality can be used to alleviate some of the flaws of machine learning methods. Much of this focused on the issue of data quality, looking to understand how to make ML models that generalise beyond the data they are trained on. In this thesis, we initially focus on issues of data quality when using ML in classical causal problems, presenting two papers on this topic. Firstly, we discuss the problem of estimating causal effects from observational studies - which may be subject to unmeasured confounding - when a small amount of experimental data is available to de-bias estimates. We place theoretical limits of the effectiveness of these methodologies, and provide a Gaussian process assumption which permits valid inference. Secondly we present a paper discussing the problem of data merging for improved estimation of conditional

causal effects. We frame this as a Bayesian experimental design problem, and develop a cryptographically secure method to evaluate the expected information gain. After this we move on to the second set of questions, asking what causality can do for the field of fair machine learning.

Fair machine learning (or algorithmic fairness) is interested in understanding how ML models can be made compliant with legal anti-discrimination requirements in decision-making settings, such as employment, criminal justice, and healthcare. In order to answer this, significant effort was placed in trying to formalise mathematically what violations of these requirements would look like. Initially, this focused on measuring various statistics -such as model performance by group - hoping that problems of discrimination could be broken down into statistical disparities. However, two clear issues were found with this approach. Firstly, for every statistic it seemed possible to draw up a scenario where discrimination intuitively was/wasn't present despite the statistic saying it wasn't/was. Secondly, in most practical cases, it is impossible to be non-discriminatory or "fair" relative to multiple statistics simultaneously. This created the problem where one statistic alone couldn't capture the problem of fairness, but it was also impossible to have multiple. These concerns lead to the field of causal fairness, which aimed to solve this problem by providing such statistics with causal meaning. These works argued that discrimination is a causal effect of protected characteristics on outcomes. Framing things in this way lead to the development of numerous new fairness statistics, which crucially varied with causal context.

We present two papers in the field of causal fairness. Firstly, we draw attention to the issue that selection bias plays in this context. We argue that from the perspective of causal fairness, selection bias is almost always present in fair ML applications. We then argue that this can create significant issues for the field as a whole, as it leaves the majority of causal effects fundamentally unidentified from observational data alone. Secondly, we look at the problem of data quality in fair machine learning more generally from a causal perspective. We provide a unified causal framework for multiple measurement biases that are typically present in fair ML applications. We then use tools from causal sensitivity analysis to create general sensitivity analysis tools to reason about the impact of measurement biases in Fair ML applications.

Finally, to conclude this thesis we present some of the limitations with these works as they stand, alongside promising directions for future work.

1 Introduction

This thesis follows the format of an integrated thesis and is composed of 6 chapters. It begins by introducing the field of causality, reviewing causal approaches to data bias and algorithmic fairness from a causal perspective. The bulk of the thesis is then made up of four independently published papers, each with its own review of the literature, and appendices. Finally, we conclude by outlining the limitations of each of these works, as well as potential avenues for future extensions.

1.1 Background on Causality

We will begin by reviewing the relevant background in causal inference. We will follow the graphical tradition, pioneered by Pearl [162] and Spirtes et al. [201], however we will frame this through the SWIG framework of Richardson and Robins [175]. We make this choice so that we can easily present the work on path-specific effects using Malinsky et al. [147], which is needed to discuss path-specific fairness [158]. This also allows us to discuss alternative theories of counterfactuals, specifically single world and cross world independences Robins and Richardson [178], which is relevant for later discussions.

1.1.1 Structural Causal Models

The key workhorse for graphical causal inference is the structural causal model:

Definition 1. *A **Structural Causal Model** (SCM) over a set of variables* $\mathbf{V} = (V_i)_{i=1}^n$ *consists of a set of functions* $\mathcal{F} = \{f_i\}_{i=1}^n$*, a tuple of noise variables* $\boldsymbol{\epsilon} = (\epsilon_i)_{i=1}^m$ *and a distribution over the noise variables* $P(\boldsymbol{\epsilon})$ *such that:*

$$V_i = f_i(\mathrm{Pa}(V_i), \boldsymbol{\epsilon}), \tag{1.1}$$

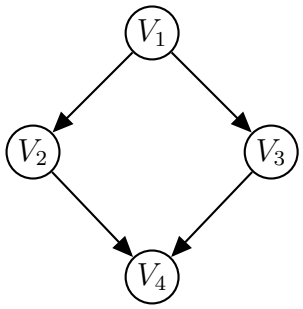
where $\mathrm{Pa}(V_i) \subset \{V_i\}_{i=1}^n$*. The distribution* $P(\boldsymbol{\epsilon})$ *is assumed to factorise as* $P(\boldsymbol{\epsilon}) = \prod_{i=1}^m P(\epsilon_i)$*. Pushing this distribution forward through the functions* $\mathcal{F}$ *we get the observational distribution* $P(\mathbf{V})$*. Given this we define the structural causal model to be the pair* $\mathcal{C} = (\mathcal{F}, P(\boldsymbol{\epsilon}))$*.*

*We define the **Causal Graph** to have a node for each element of* $\mathbf{V}$ *and a directed edge from all nodes in* $\mathrm{Pa}(V_i)$ *into* V_i*. We will assume the graph is **acyclic**[1]. and therefore refer to it as a **Directed Acyclic Graph** (DAG).*

The relationship between a structural causal model and the graph is given in the following diagram:

$$\begin{aligned} V_1 &= f_1(\epsilon_1) \\ V_2 &= f_2(V_1, \epsilon_2) \\ V_3 &= f_3(V_1, \epsilon_3) \\ V_4 &= f_4(V_2, V_3, \epsilon_4) \\ \mathcal{C} &= (\{f_1, f_2, f_3, f_4\}, P(\boldsymbol{\epsilon})) \end{aligned}$$

(a) An example of an SCM $\mathcal{C}$.



(b) The directed acyclic graph corresponding to the SCM.

Figure 1.1: Example of a DAG and the corresponding SCM.

The underlying causal structure will have implications for the observed distribution in terms of conditional independences. To see this, consider the causal graphs in Figure 1.2. We can see that for all causal structures which comply graphs (a) - (c) we will

[1]Causal models with cyclic graphs are possible but not considered in this thesis. More information can be found in [27]

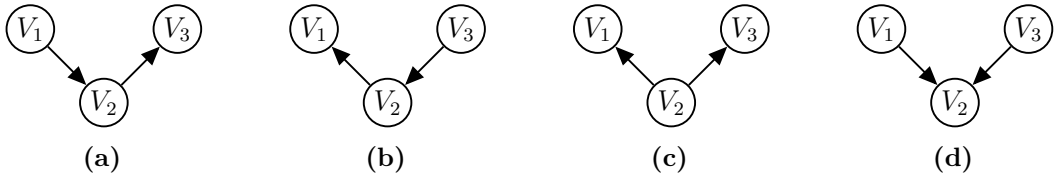


Figure 1.2: Causal graphs showing the possible conditional independence structures over three variables. For graphs (a) - (c) we have $V_1 \perp\!\!\!\perp V_3 \mid V_2$ whereas for (d) we have $V_1 \perp\!\!\!\perp V_3$ but $V_1 \not\perp\!\!\!\perp V_3 \mid V_2$.

have that $V_1 \perp\!\!\!\perp V_3 \mid V_2$. But for graph (d) we have that $V_1 \perp\!\!\!\perp V_3$ but $V_1 \not\perp\!\!\!\perp V_3 \mid V_2$. These three variable relationships can be chained together to give a rule for reading conditional independences from DAGs, which is known as **d-separation**. It is defined as follows:

Definition 2. *Consider disjoint sets of variables* $\mathbf{A}, \mathbf{B} \subseteq \mathbf{C}$*. We say an undirected-path from* $A \in \mathbf{A}$ *to* $B \in \mathbf{B}$ *is* ***blocked*** *by* $\mathbf{C}$ *if we have either of the following:*

1. *The path contains a triplet of vertices* (V_1, V_2, V_3) *such that the subgraph on this triplet is of the form (a)-(c) in Figure 1.2 and* $V_2 \in \mathbf{C}$*.*
2. *The path contains a triplet of vertices* (V_1, V_2, V_3) *such that the subgraph is of the form (d) where neither* V_2 *nor any of its children are in* $\mathbf{C}$*.*

We say $\mathbf{A}$ *is* ***d-separated*** *from* $\mathbf{B}$ *by* $\mathbf{C}$ *if all undirected paths from* $\mathbf{A}$ *to* $\mathbf{B}$ *are blocked by* $\mathbf{C}$*. We write this as* $\mathbf{A} \perp\!\!\!\perp_d \mathbf{B} \mid \mathbf{C}$*.*

Now d-separation allows conditional independences to be read off the causal graph according to the following proposition:

Proposition 1. *The d-separation* $\mathbf{A} \perp\!\!\!\perp_d \mathbf{B} \mid \mathbf{C}$ *implies the conditional independence* $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$ *for all SCM's compatible with the causal graph.*

This was first proved in Verma and Pearl [210]. Meek [149] then demonstrated that this criterion is complete so that all possible independences shared between causal models that comply with a given graph can be found using d-separation. The DAG also implies the following factorisation of the observational distribution:

$$P(\mathbf{V}) = \prod_{i=1}^{n} P(V_i \mid \mathrm{Pa}(V_i)) \tag{1.2}$$

Which is equivalent to d-separation implying conditional independence [131].

1.1.2 Potential Outcomes and the Interventional Distribution

Whilst causal models have observational implications, the key point of is that they allow us to define counterfactuals (also known as potential outcomes [183]) which can be used to define interventional distributions. These are defined as follows:

Definition 3. *For any tuple of variables* $\mathbf{A} \subset \mathbf{V}$ *we define the* ***potential outcomes*** *when setting* $\mathbf{A} = \mathbf{a}$ *recursively as follows:*

1. *For* $V_i \in \mathbf{V}$ *if* $\text{Pa}(V_i) \subset \mathbf{A}$ *we define* $V(\mathbf{a}) = f_i(\mathbf{a}_{\text{Pa}(V_i)}, \epsilon_i)$
2. *For other* V_i*, we define* $V_i(\mathbf{a})$ *recursively as:*

$$V_i(\mathbf{a}) \coloneqq V_i(\mathbf{a}_{\text{Pa}(V_i)}, \{V_j \mid j \in \text{Pa}(V_i) \setminus \mathbf{A}\}) \tag{1.3}$$

Note, we take $V_i(v_i)$ *to be the random variable* V_i*.*

The interventional distribution is then defined as $P(V(\mathbf{a}))$. The natural question is, how can we estimate the interventional distribution from observational data alone? The first key step is to notice that for a variable Y a fixed value pa_Y of $\text{Pa}(Y)$ the interventional distribution may be written as:

$$P(Y(\text{pa}_Y) \mid \text{Pa}(Y) = \text{pa}_Y) = P(Y \mid \text{Pa}(Y) = \text{pa}_Y) \tag{1.4}$$

This is the central fact which allows us to relate interventional distributions to observational ones. It allows us to write the interventional distribution, $P(\mathbf{V}(\mathbf{a}))$, in terms of the observational distribution as:

$$P(\mathbf{V}(\mathbf{a}) = \mathbf{v}) = \prod_{i=1}^{n} P(V_i \mid \text{Pa}(V_i) \setminus \mathbf{A} = \mathbf{v}_{\text{Pa}(V_i)\setminus\mathbf{A}}, \text{Pa}(V_i) \cap \mathbf{A} = \mathbf{a}_{\text{Pa}(V_i)\cap\mathbf{a}})$$
$$\text{when} \prod_{A_j \in \mathbf{A}} P(A_j = a_j \mid \text{Pa}(V_i) \setminus \mathbf{A} = \mathbf{v}_{\text{Pa}(V_i)\setminus\mathbf{A}}, \text{Pa}(V_i) \cap \mathbf{A} = \mathbf{a}_{\text{Pa}(V_i)\cap\mathbf{a}}) > 0$$

This is known as the extended G-formula [179] and a proof can be found in Richardson and Robins [175].

1.1.3 Counterfactual Distributions

Structural causal models also allow us to define distributions over counterfactual events. Counterfactuals are "what if?" questions, asking what would have happened if events

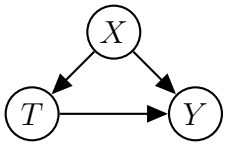
had been different, given that we observe a particular outcome. Such events may be written using the distribution over multiple different potential outcomes, for example:

$$P(Y(\tilde{\mathbf{a}}) = y, Y(\mathbf{a}) = y, \mathbf{X} = \mathbf{x}) \tag{1.5}$$

Using the SCM, we can evaluate these probabilities as follows: Let E be an event over a number of different potential outcomes and the structural causal model be given by $\mathcal{C} = (\mathcal{F}, P(\boldsymbol{\epsilon}))$. We write $\mathcal{F}(\boldsymbol{\epsilon}) = E$ if for a given value ϵ of the noise variables the event E holds. Then the probability of E is given as follows:

$$P(E) = \mathbb{E}_{P(\boldsymbol{\epsilon})} \left[\mathbb{1} \left\{ \mathcal{F}(\boldsymbol{\epsilon}) = E \right\} \right] \tag{1.6}$$

As this probability relies on the SCM, in general we would need access to the SCM to evaluate a counterfactual probability - which would be a very strong assumption. However, for some counterfactual probabilities the causal graph alone is sufficient to identify the probability. For example, consider the following causal model:



(1.7)

For all structural causal models compliant with this graph, we have that the counterfactual probability $P(Y(t) \mid T = t')$ is identified as:

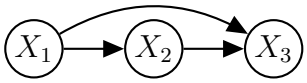
$$P(Y(t) \mid T = t') = \mathbb{E}_{P(X|T=t')} \left[P(Y \mid X, T = t) \right] \tag{1.8}$$

So long as $P(X \mid T = t') \ll P(X \mid T = t)$. That is, for measurable sets S when $P(X \in S \mid T = t) = 0$ we have $P(X \in S \mid T = t') = 0$. Imagining $T = t$ corresponds to the allocation of a medical treatment, $P(Y(t) \mid T = t')$ would be the distribution of outcomes for the untreated group, had they counterfactually received treatment.

1.1.3.1 Debate over Counterfactual Dependencies

It is worth making a note at this stage that there is some debate over the correct assumptions to be made when performing counterfactual inference. So far we have written everything under the assumptions of *non parametric structural equation models with independent errors*, also known as the **NPSEM-IE**. That is, when setting out the definition of an SCM we assumed that the distribution factorises as $P(\boldsymbol{\epsilon}) = \prod_{i=1}^{n} P(\epsilon_i)$.

This assumption leads to a large number of independences between unobserved counterfactuals. For example consider the following graph[2]:



(1.9)

Under an NPSEM-IE, we make the assumptions that for any choice of $x_1, \tilde{x}_1, x_2$ we have the following:

$$X_1 \perp\!\!\!\perp X_2(x_1) \perp\!\!\!\perp X_3(\tilde{x}_1, x_2) \tag{1.10}$$

This assumption is controversial, as it implies an independence between counterfactuals which can never be jointly observed. Therefore, we could never experimentally verify statements of the form $X_2(x_1) \perp\!\!\!\perp X_3(\tilde{x}_1, x_2)$. Such statements are known as *cross-world independences*. Robins and Richardson [178] argue that these cross-world independencies are inherently unscientific, as they cannot be falsified. Instead, they argue for an alternative set of counterfactual independences, known as the *Finest Fully Randomized Causally Interpretable Structured Tree Graph* (**FFRCISTG**) first introduced in Robins [177]. This implies that for any vector, $\mathbf{v}$, the following set of variables are all independent:

$$\left\{ V_i(\mathbf{v}_{\mathrm{Pa}(V_i)}) \mid V_i \in \mathbf{V} \right\}, \tag{1.11}$$

where $V_i(\mathbf{v}_{\mathrm{Pa}(V_i)})$ corresponds to setting the values of $\mathrm{Pa}(V_i)$ to their corresponding values in $\mathbf{v}$. Returning to the graph in 1.9 this would imply the following dependences only:

$$X_1 \perp\!\!\!\perp X_2(x_1) \perp\!\!\!\perp X_2(x_1, x_2) \tag{1.12}$$

For any choice of x_1, x_2.

1.1.4 Causal Structure Learning

In all of the above, we have taken the causal graph as somewhat of a given. Given how central it is for correctly estimating the causal effect, the natural question is where does it come from? The first main answer to this question is that the graph comes from expert knowledge on a given problem, be these from physical laws, otherwise

[2]Example taken from the lecture notes of Prof. Qingyuan Zhao.

understood mechanisms, or randomised controlled trials. It is also argued that if you try to do causal inference without the knowledge of the graph, you are not avoiding this problem, you are just *implicitly* assuming a certain causal graph. This graph will arise from the statistically estimable quantity that is claimed to represent a causal effect.

There is, however, a second option for getting the graph, which is known as causal discovery or causal structure learning. Causal discovery notes that the underlying causal structure can have implications for the observational distribution, most commonly in the form of conditional independences, as discussed in Section 1.1.1. The goal of causal discovery is to use this structure to guide us to a set of feasible causal graphs. The field is very broad and mostly lies outside of the scope of this thesis. However, for completeness we will briefly discuss the most famous method for causal discovery, the PC algorithm [200].

The PC algorithm is built upon the observation that if these is no edge between X and Y in the graph $\mathcal{G}$, we must have some set Z such that $X \perp\!\!\!\perp Y \mid Z$. The algorithm builds upon this, starting with the complete undirected graph removing the edge between X and Y whenever we find some set Z such that $X \perp\!\!\!\perp Y \mid Z$. There is a subtlety here, whilst d-separation allows us to conclude that the lack of an edge implies a conditional independence, it does mean a conditional independence implies a lack of an edge. It is possible to construct SCMs such that the impact across different paths "cancel out", creating a conditional independence in the observational distribution which is not implied by the graphical structure. Therefore, for this to step to be valid, we must assume that the conditional independences in the observational distribution are only the ones implied by the graphical structure. This assumption is known as *faithfulness*, and is essential for performing causal discovery in this way.

Once all edges have removed that can be, we are left with an undirected graph over all variables. There are now two ways to orient the edges. Firstly, note that according to the graphs in Figure 1.2, the only way we can have two variables X and Y that are adjacent to a vertex Z, but independent of each is if we have $X \to Z \leftarrow Y$. Therefore, for any triplets of variables (X, Y, Z) where X and Y to Z but not each other, we test if $X \perp\!\!\!\perp Y \mid S$ where S does not contain Z. If this holds for any S we can orient the edges as $X \to Y \leftarrow Z$. After we have completed all of these orientations, we then use the acyclicity assumption to orient any edges which would create a cycle if they pointed

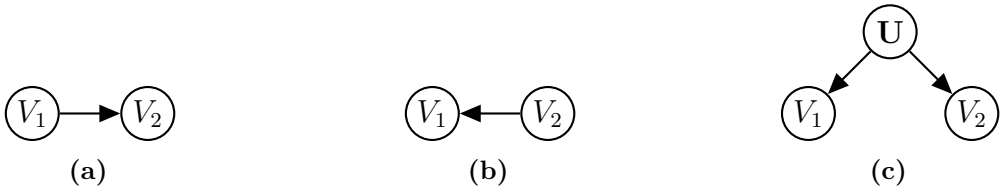


Figure 1.3: Example graphs for the cases given by Reichenbach's common cause principle.

one way. Once this procedure is complete, we are left with a graph containing both directed and undirected edges, known as a Completed Partially Directed Acyclic Graph (CPDAG). In this graph, all undirected edges could be oriented in either direction to give a valid DAG. However, the information in the observational distribution does not allow us to distinguish between any of these possibilities. For this reason, the CPDAG represents what is known as a Markov equivalence class over DAGs, where the equivalence is in the sense of permitted observational distributions.

1.2 Causal Approaches to Data Quality

We now review the literature on data quality and causality. We specifically focus on three forms of bias that are most relevant to this thesis: unmeasured confounding, selection bias, and multi-environment data.

1.2.1 Unmeasured Confounding, Hidden Variables, and Identifiability

As the focus of this thesis is causal approaches to statistical biases, it feels right to begin this section with the statistical bias that arguably gives rise to causal inference, unmeasured confounding. When you begin a PhD in causal inference, you may expect you will learn the "secret sauce" that makes the phrase "correlation is not causation" a thing of the past, meaning that no arrangement of sprinklers, rain, or wet pavement will confuse you again. In reality the secret sauce works the other way round. Most correlations would be causal if not for the problem of unmeasured confounding - that is when two variables share an unmeasured common cause.

The fact that correlations either arise through causal relationships or a shared common cause was historically justified using Reichenbach's common cause principle. This states that if two variables, V_1 and V_2, are dependent, either one causes the other or

they share a common cause, $\mathbf{U}$, which when conditioned on would render V_1 and V_2 independent. An example of these cases are given in Figure 1.3. In fact, this statement can be proved under the assumption that V_1, V_2 are embedded in a structural causal model:

Proposition 2 (Reichenbach's common cause principle)**.** *Suppose $V_1, V_2 \in \mathbf{V}$ and we have a structural causal model $\mathcal{C} = (\mathcal{F}, P(\boldsymbol{\epsilon}))$ which describes the relationships between the variables. If $V_1 \not\perp\!\!\!\perp V_2$ then one of the following holds:*

1. *There exists a directed path $V_1 \to \cdots \to V_2$ in the causal graph, $\mathcal{G}$.*
2. *There exists a directed path $V_2 \to \cdots \to V_1$ in the causal graph, $\mathcal{G}$.*
3. *There exists a set of variables $\mathbf{U}$ such that $V_1 \perp\!\!\!\perp V_2 \mid \mathbf{U}$ and for some $U, U' \in \mathbf{U}$ we have the directed paths $U \to \cdots \to V_1$ and $U' \to \cdots \to V_2$, in the causal graph, $\mathcal{G}$.*

Proof. As $V_1 \not\perp\!\!\!\perp V_2$ there must exist an unblocked path in $\mathcal{G}$ between V_1 and V_2. Now, let $\mathbf{U}$ be the set of variables in $\mathbf{V}$ which cannot be reached from a directed path from V_1. Now we have two cases:

1. $V_1 \not\perp\!\!\!\perp V_2 \mid \mathbf{U}$. In this case, there must still exist an unblocked path from V_1 to V_2 in $\mathcal{G}$. Let $(V_1, X_1, \cdots, X_n, V_2)$ be such a path. Now we cannot have $V_1 \leftarrow X_1$ as otherwise X_1 would be in $\mathbf{U}$ and so the path would be blocked by $\mathbf{U}$. Hence, we must have $V_1 \to X_1$. Now for all remaining vertices we must have $X_i \to X_{i+1}$ as having $X_i \leftarrow X_{i+1}$ would block the path. Therefore $(V_1, X_1, \cdots, X_n, V_2)$ is a directed path from V_1 to V_2.
2. $V_1 \perp\!\!\!\perp V_2 \mid \mathbf{U}$. If this is the case then there must be a path $(V_1, X_1, \cdots, U, \cdots, X_n, V_2)$ in $\mathcal{G}$ which is blocked by the vertex $U \in \mathbf{U}$. As $\mathbf{U}$ cannot be reached from V_1 via a directed path we must have $U \to X_m \cdots \to V_1$. Now consider the portion of the path $(U, X_{m+1}, \cdots, V_2)$. If $U \to X_{m+1}$ then all edges on this path must be $\to$ and so we have found a set $\mathbf{U}$ satisfying condition (3). If not, then let X' be the last vertex on this path such that we have a directed path $(X, \cdots, U, \cdots, V_1)$. If $X = V_2$ we have a directed path from V_2 to V_1 but if not we must have $X \in \mathbf{U}$ as if X could be reached from V_1 by a directed path the graph would contain a cycle.

□

Therefore, structural causal models can provide a theoretical justification for the common cause principle of Reichenbach. As we will later see in the selection bias section, this is not the full story on the matter, but framing everything in causal models does allow such claims to be formally verified.

Now we know that correlations are either due to causal relationships or a shared common cause (known as a *confounder*) we should be able to disentangle causal relationships from statistical correlations, putting an end to the phrase "correlation is not causation" forever. However the key challenge of causal inference comes from the problem of *unmeasured confounding*. That is when we do not observe all the common causes. If this is the case, we cannot correct for the effects of unmeasured confounders to properly estimate causal effects. In order to reason about this, we first need to consider the problem of hidden variables in causal inference.

1.2.1.1 Hidden Variables and Marginalisation

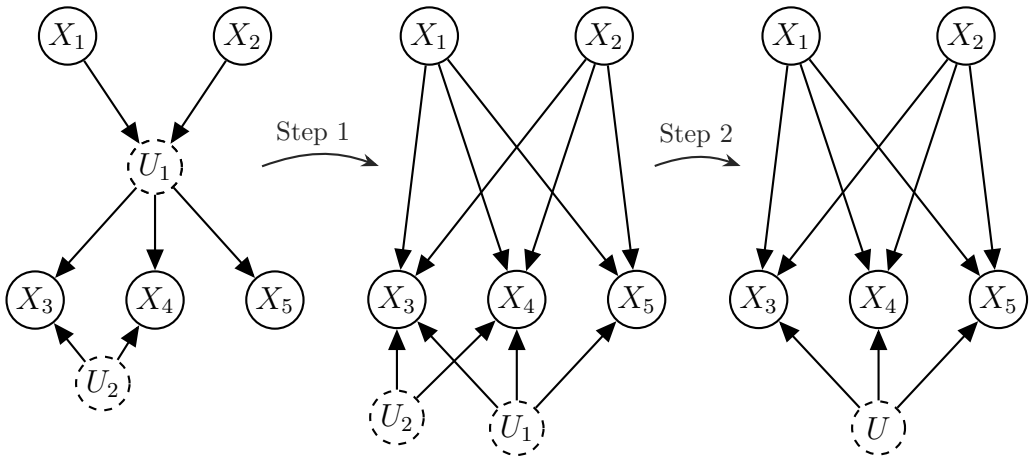


Figure 1.4: Example the marginalisation operations given in Evans [70] applied to a DAG with hidden variables.

First we will look at the problem of hidden variables from the perspective of structural causal models. Specifically, suppose the variable set can be split as $\mathbf{V} \cup \mathbf{U}$ where we only observe the variables $\mathbf{V}$. If no restrictions are placed over $\mathbf{U}$, then in theory there are infinitely many possible causal structures. As in practice we would never expect to observe every single variable that is relevant to a problem, this would preclude any realistic chance of doing causal inference. Thankfully, Evans [70] showed that any structural causal model over $\mathbf{V}$ can be reduced to one of finitely many causal models,

irrespective of the cardinality of the set of hidden variables. Specifically, an SCM in canonical form which is defined as follows:

Definition 4. *Let $\mathcal{C} = (\mathcal{F}, P(\boldsymbol{\epsilon}))$ be a structural causal model over variables $\mathbf{V} \cup \mathbf{U}$ where the variables $\mathbf{U}$ are unobserved. We say $\mathcal{C}$ is in **canonical form** if the following holds:*

1. *Each $U \in \mathbf{U}$ has no causal parents.*
2. *There are no two variables $U, U' \in \mathbf{U}$ such that $\mathrm{Ch}(U) \subseteq \mathrm{Ch}(U')$ where $\mathrm{Ch}(U)$ is the set of $X \in \mathbf{V} \cup \mathbf{U}$ such that there is an edge $U \to X$.*

Marginalisation Operation Evans [70] provide a marginalisation operation which converts an arbitrary SCM over variables $\mathbf{V} \cup \tilde{\mathbf{U}}$ to one over variables $\mathbf{V} \cup \mathbf{U}$ which is in canonical form. This consist of two steps:

1. For all $U \in \tilde{\mathbf{U}}$, add an edge $Z \to \tilde{Z}$ if the current graph contains $Z \to U \to \tilde{Z}$ and then delete any edges $Z \to U$,
2. After completing the first step for all variables in $\tilde{\mathbf{U}}$, delete any U if there exists another $\tilde{U} \in \tilde{\mathbf{U}}$ that influences all of the variables U influences.

Evans [70] showed that there is a structural causal model over the resulting graph which preserves the causal structure over the variables $\mathbf{V}$. Importantly, due to the deletion step, this model has a bounded number of unobserved variables, regardless of how large the set U is. We illustrate these steps in Figure 1.4.

Latent Projection Suppose we want to represent the causal structure over $\mathbf{V}$ without having to draw the hidden variables in $\mathbf{U}$. This is done via *Latent Projection*, first introduced by Verma [209]. This involves replacing hidden variables in a graph by generalised edges. It can be done in two ways, leading to two different types of graphs:

1. **ADMGs** [209]. Begin with a hidden variable model in canonical form. We then add a bidirected edge between $V_i \leftrightarrow V_j$ if there exists a hidden variable U such that $V_i \leftarrow U \to V_j$. After adding all such edges we remove the hidden variables, $\mathbf{U}$. This forms the class of *Acyclic Directed Mixed Graphs* (**ADMGs**).

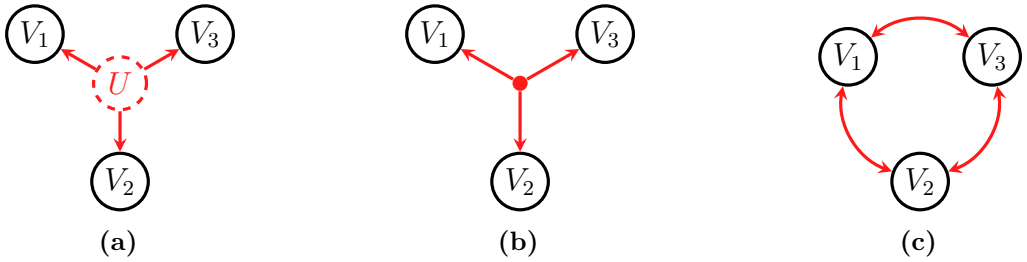


Figure 1.5: Example of the two different forms of latent projection. (b) and (c) are both latent projections of the graph (a), where (b) is the projection into the space of mDAGs and (c) is the projection into the space of ADMGs.

2. **mDAGSs** [70] Begin again a hidden variable model in canonical form. For any hidden variable $U \in \mathbf{U}$ we add a hyper-edge between all children of U and then remove U. This forms the class of *Marginalised DAGs* (**mDAGs**).

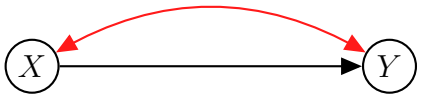
We illustrate the difference between these projections in Figure 1.5. Conditional independences can still be read off such graphs using an extension of d-separation which accounts for generalised edges, known as m-separation [174].

For both ADMGs and mDAGs, there exists a model in the projected space which has the same distribution as the hidden variable model. However, projecting into the space of ADMGs can result in the loss of some constraints implied by the hidden variable model. For example, in the hidden variable model in Figure 1.5a, Fritz [84] demonstrated that for binary V_1, V_2, V_3 we cannot have $P(V_1 = V_2 = V_3 = 1) = P(V_1 = V_2 = V_3 = 0) = \frac{1}{2}$. This constraint is lost when projecting into the space of ADMGs, unlike when projecting into the space of mDAGs. Despite this, we will work with the space of ADMGs throughout, as much of the theory of identifiability of causal effects has been developed for this case.

1.2.1.2 Identifiability of Causal Effects with Hidden Variables

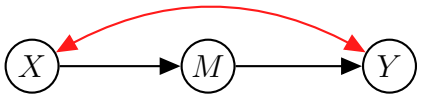
A bi-directed edge between variables $V_1 \leftrightarrow V_2$ in an ADMG corresponds to an unmeasured confounder- that is an unmeasured common cause of the variables V_1 and V_2. Unmeasured confounding represents the largest obstacle to performing valid causal inference. This is because in the presence of unmeasured confounding we cannot break down a correlation into a series of causal relationships. In the most simple example, given by the following DAG, we cannot estimate the interventional

distribution, $P(Y(x))$, without further assumptions:



(1.13)

However, there are some scenarios where if we have more details in the causal graph, we can still correctly estimate causal effects. For example, consider the alteration to the graph in 1.13, which results from an intermediate variable on the path from V_1 to V_2:



(1.14)

In this case, the interventional distribution $P(Y(x))$ is now identified as:

$$P(Y(x)) = \mathbb{E}_{P(M|X=x)}\mathbb{E}_{P(X)}[P(Y \mid M, X)], \tag{1.15}$$

a fact known as the *front-door criteria*, whose proof can be found in Pearl [162]. In general, we want to be able to understand when an expression involving potential outcomes can be written in terms of factual probabilities, and so estimated from data[3]. For this, Pearl [162] developed the do calculus which shows when two expressions involving interventional operators are equivalent given a causal graph. Here we present the po calculus of Malinsky et al. [147]. We do this as it will allow us to use potential outcomes with graphical models, which will be especially useful later for considering path-specific effects. To do this we first introduce a new, modified graph called a *single-world intervention graph* (**SWIG**) that is used to reason about interventions.

SWIGs [175] SWIGs are graphical models that are used to graphically represent the interventional distribution $\mathbf{V}(\mathbf{a})$. It is denoted by $\mathcal{G}(\mathbf{a})$ and is formed from $\mathcal{G}$ by the following steps:

1. Begin with a node for each variable $\mathbf{V}$ and then split each $A \in \mathbf{A}$ into two nodes labelled a and A, denoting the fixed and random parts respectively. We refer to this node set as $\mathbf{V}(\mathbf{a})$.

[3]There is a subtly here, which is that being able to write an expression in terms of factual probabilities is not exactly the same as being able to estimate it with finite data, due to estimators being discontinuous in the distribution. For a discussion of this issue in relation to causal inference, we refer the reader to [146].

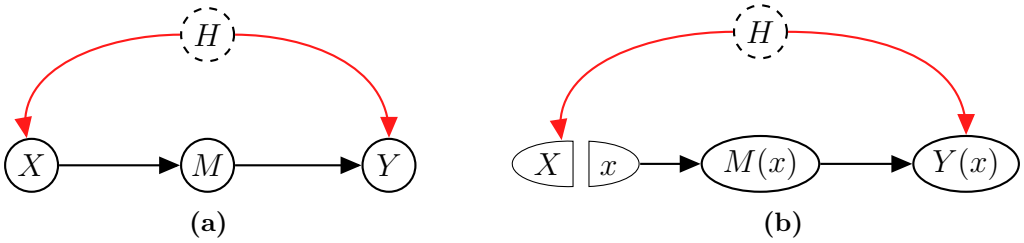


Figure 1.6: Example of a SWIG for the graph in 1.14 with the hidden variables drawn in. Figure 1.6a shows the normal graph and Figure 1.6b shows the SWIG for the intervention $X = x$.

2. For edges $V \to A$ and $A \to V$ in $\mathcal{G}$ draw edges $V \to A$ and $a \to V$ in $\mathcal{G}(\mathbf{a})$ respectively. Otherwise keep all the edges the same as $\mathcal{G}$.

3. Replace each node V in $\mathcal{G}(\mathbf{a})$ by $V(\mathbf{a}_{\text{An}_{\mathcal{G}(\mathbf{a})}(V)})$ where $\text{An}_{\mathcal{G}(\mathbf{a})}(V)$ is the set of ancestors of V in $\mathcal{G}(\mathbf{a})$ (i.e the set of nodes from which there exists a direct path to V). If $\text{An}_{\mathcal{G}(\mathbf{a})}(V) = \emptyset$ simply leave V unchanged.

We illustrate this operation for a simple DAG in Figure 1.6. Just as the observational graph implied conditional independences in the observational distribution the interventional graph implies conditional independences in the interventional distribution. This is given by the SWIG global Markov property:

Definition 5 (**SWIG Global Markov Property**). *The SWIG global Markov property states that for disjoint subsets* $\mathbf{Y}(\mathbf{a})$, $\mathbf{X}(\mathbf{a})$, *and* $\mathbf{Z}(\mathbf{a})$ *and a subset* $\mathbf{a}'$ *of* $\mathbf{a}$ *if we have:*

$$\mathbf{Y}(\mathbf{a}), \mathbf{a}' \perp\!\!\!\perp_m \mathbf{Z}(\mathbf{a}) \mid \mathbf{X}(\mathbf{a}), \tag{1.16}$$

where $\perp\!\!\!\perp_m$ *denotes d-separation*[4]*, then for some* $f(\cdot)$ *we have:*

$$P(\mathbf{Z}(\mathbf{a}) \mid \mathbf{Y}(\mathbf{a}), \mathbf{X}(\mathbf{a})) = P(\mathbf{Z}(\mathbf{a}) \mid \mathbf{X}(\mathbf{a})) \tag{1.17}$$

$$= f(Z, X, \mathbf{a} \setminus \mathbf{a}'). \tag{1.18}$$

When the data is generated according to an SCM with independent errors (or the weaker FFRCISTG model), the SWIG global Markov property holds. In fact, once we add the consistency property ($\mathbf{B}(\mathbf{a}) = b \implies V_i(\mathbf{a}, b) = V_i(\mathbf{a})$) the expression:

$$P(\mathbf{V}(\mathbf{a}) = \mathbf{v}) = \prod_{i=1}^{n} P(V_i \mid \text{Pa}(V_i) \setminus \mathbf{A} = \mathbf{v}_{\text{Pa}(V_i)\setminus\mathbf{A}}, \text{Pa}(V_i) \cap \mathbf{A} = \mathbf{a}_{\text{Pa}(V_i)\cap\mathbf{a}}),$$

[4]m-separation is the extension of d-separation to graphs with bidirected edges. It implies the same set of conditional independences as performing d-separation on the graph where every bidirected edge $V_i \leftrightarrow V_j$ is replaced by $V_i \leftarrow U \to V_j$.

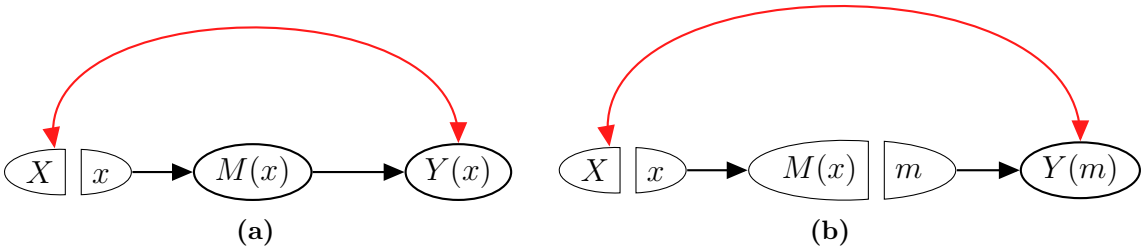


Figure 1.7: Example of a SWIG for the graph in 1.14 with the hidden variables drawn in. Figure 1.6a shows the normal graph and Figure 1.6b shows the SWIG for the intervention $X = x$.

can be viewed as a factorisation of the distribution of $P(V(\mathbf{a}))$ according to the SWIG $\mathcal{G}(\mathbf{a})$ that arises from the DAG $\mathcal{G}$. We could even see the original global Markov property and factorisation as specific cases of the SWIG global Markov property when the intervention set is empty.

The SWIG global Markov property is combined with the two additional rules to give a sound and complete way for reducing statements about involving potential outcomes into ones involving observational quantities. That is- any statement involving potential outcomes can be reduced to one with only observational properties if and only if it can be reduced using the three following rules:

$$
\begin{aligned}
1: \; & p(Y(x) \mid Z(x), W(x)) = p(Y(x) \mid W(x)) \\
& \qquad \text{if } (Y(x) \perp\!\!\!\perp_m Z(x) \mid W(x))_{\mathcal{G}(x)} \\
2: \; & p(Y(x,z) \mid W(x,z)) = p(Y(x) \mid W(x), Z(x) = z) \\
& \qquad \text{if } (Y(x,z) \perp\!\!\!\perp_m Z(x,z) \mid W(x,z))_{\mathcal{G}(x,z)} \\
3^*: \; & p(Y(x,z)) = p(Y(x)) \\
& \qquad \text{if } (Y(x,z) \perp\!\!\!\perp_m z)_{\mathcal{G}(x,z)}
\end{aligned}
$$

These are equivalent to the rules of do calculus [162] and they were introduced in Malinsky et al. [147].

As an example, we will now use the po calculus to re-derive the front door criteria from Equation 1.15. To do so we have drawn up all the relevant SWIGs in Figure 1.7.

$$P(Y(x)) = \mathbb{E}_{P(M(x))}[Y(x) \mid M(x) = m] \text{(Law of Total Expectation)} \tag{1.19}$$
$$= \mathbb{E}_{P(M(x))}[Y(x, m) \mid M(x) = m] \text{(Consistency)} \tag{1.20}$$
$$= \mathbb{E}_{P(M(x))}[Y(m) \mid M(x) = m] \text{(Rule 3 in } \mathcal{G}(x, m)) \tag{1.21}$$
$$= \mathbb{E}_{P(M(x))}[Y(m)] \text{(Rule 1 in } \mathcal{G}(x, m)) \tag{1.22}$$
$$= \mathbb{E}_{P(M(x))}[\mathbb{E}_{P(X)}[Y \mid M = m, X]] \text{(Consistency)} \tag{1.23}$$
$$= \mathbb{E}_{P(M|X=x)}[\mathbb{E}_{P(X)}[Y \mid M = m, X]] \text{(Rule 1, Consistency in } \mathcal{G}(x)) \tag{1.24}$$

1.2.2 Selection Bias

Selection bias occurs when we sample in a biased way from the target population. This is a problem as it means statistical estimates from this sample will themselves be biased and not representative of the target population. A common example of selection bias would be in collection of survey data. We only ever see responses from individuals who fill out the survey and it is unlikely that those who fill out the survey will be representative of the target population. For example, people with particularly strong opinions may be more motivated to fill out the survey to express them. Or as the joke goes, "85% of survey respondents said they enjoy the process of filling out surveys".

In causal inference we depict selection bias using a binary selection variable, denoted by S, which is added to the causal graph. The biased population we sample from corresponds to having $S = 1$ and the goal of learning under selection bias is to estimate statistics or causal effects for the whole population, including when $S = 0$. Selection nodes are depicted with dashed edges, with a selection of examples in Figure 1.8.

In the context of causal inference there are three questions we could ask about recovering from selection bias: i) When is an observational distribution recoverable from biased data? ii) When is an interventional distribution recoverable from biased data? iii) How can external data from the unbiased population help in either of these cases? We will now detail the solutions to these problems given by Bareinboim et al. [19].

1.2.2.1 Recovering Observational Distributions

First, we focus on recovering observational distributions. Generally this is a distribution of the form $P(\mathbf{Y} \mid \mathbf{X})$, where $\mathbf{X} \cup \mathbf{Y} \subseteq \mathbf{V}$ and $\mathbf{V}$ is the whole set of observed variables.

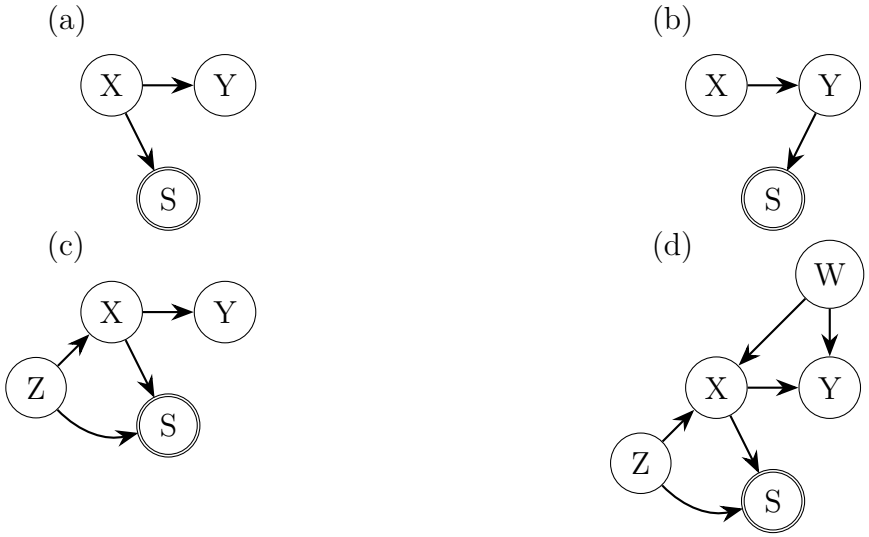


Figure 1.8: Examples of some of the most simple graphs in which $P(Y|X)$ is and isn't recoverable from selected data, taken from Bareinboim and Tian [17]. In (a), $P(Y|X)$ is recoverable whilst the changing of the edge to (b) makes it unidentifiable. Likewise, in (c) we can recover $P(Y|X)$ but adding the variable W makes it unidentifiable as conditioning on X now opens a path from S to Y through Z, X, W.

We assume we observe the selected data, so $P(\mathbf{V} \mid S = 1)$ and that the data is generated according to a structural causal model that complies with the graph $\mathcal{G}_S$ where the S denotes the fact that we have added a selection node to the graph. Therefore, the goal is to write the distribution $P(\mathbf{Y} \mid \mathbf{X})$ in terms of probabilities conditional on $S = 1$. Firstly, we can see that the condition $\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{X}$ is easily sufficient to correct for selection bias as we can write:

$$P(\mathbf{Y} \mid \mathbf{X}) = P(\mathbf{Y} \mid \mathbf{X}, S = 1), \tag{1.25}$$

which uses conditional independence to show that the distributions are equal in the selected and unselected data. Bareinboim et al. [19] show that $\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{X}$ turns out to be necessary and sufficient for the recovery of the distribution from observational data:
Proposition 3 (Bareinboim et al. [19])**.** *The distribution* $P(\mathbf{Y} \mid \mathbf{X})$ *is recoverable from* $P(\mathbf{V} \mid S = 1)$ *if and only if* $\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{X}$.

It is worth noting that there are specific cases where additional assumptions can allow for recovery of the observational distribution more broadly. For example, Evans and Didelez [72] give a scenario where knowledge of independence constraints in the unbiased distribution can lead to identifiability for discrete variables.

Recoverability with External Data In some scenarios we may have access to external unbiased data which can be used to recover the conditional distribution $P(\mathbf{Y} \mid \mathbf{X})$ more broadly than when $\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{X}$. For example, we may have access to census data which gives information on a variety of demographic characteristics such as age, gender, or income distribution. The hope is that this could be used in conjunction with more specific survey data to get unbiased estimates of the observational distribution. The following proposition gives an example of this:

Proposition 4. *Suppose we observe samples from the unselected distribution for variables* $\mathbf{X}, \mathbf{C}$ *where* $\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{X}, \mathbf{C}$. *Then the distribution* $P(\mathbf{Y} \mid \mathbf{X})$ *can be written as:*

$$P(\mathbf{Y} \mid \mathbf{X}) = \mathbb{E}_{P(\mathbf{C}|\mathbf{X})}\left[P(\mathbf{Y} \mid \mathbf{X}, \mathbf{C}, S = 1)\right], \tag{1.26}$$

Which can be estimated with samples from $P(\mathbf{V} \mid S = 1)$ *and* $P(\mathbf{X}, \mathbf{C})$.

1.2.2.2 Recovering Causal Effects

Now we have discussed the recoverability of observational distributions from selected distributions, we move on to the question of interventional distributions. Firstly, we can see that the rules of do (or po) calculus lead directly to a statement about recovering interventional distributions:

Corollary 5 (Bareinboim and Tian [17])**.** *An interventional distribution,* $P(Y(x))$*, is recoverable from selected data generated by the graph* $\mathcal{G}$ *if and only if it can be reduced by the rules of do calculus to an expression involving observational probabilities which are themselves recoverable by Proposition 3.*

Whilst this gives a condition for the recoverability of an interventional distribution, it is not clear when it is or is not satisfied for a given graph and distribution. This is because do calculus can lead to many different expressions in terms of observational probabilities, and it may be the case that only one of these expansions is recoverable from selected data. In general, searching over all sets would be exponentially costly in the size of the graph. However, for DAG models, Bareinboim and Tian [17] are able to reduce this to a complete condition for the recoverability of the interventional distribution:

Proposition 6. *Let* X, Y *be disjoint subsets of* $\mathbf{V}$. *The interventional distribution* $P(Y(x))$ *is recoverable if and only if for every* $V_i \in \mathrm{An}_{\mathcal{G}_{\mathbf{V}\setminus X}}(Y)$ *where* $\mathrm{An}_{\mathcal{G}_{\mathbf{V}\setminus X}}(Y)$ *are*

the set of ancestors of Y in the subgraph with vertices $\mathbf{V} \setminus X$, there is no direct path to S. In this case the distribution can be written as:

$$P(Y(\mathbf{x})) = \int_{D \setminus Y} \prod_{i \in D} P(v_i \mid \mathrm{Pa}(v_i), S = 1) \tag{1.27}$$

Where $D = \mathrm{An}_{\mathcal{G}_{\mathbf{V} \setminus X}}(Y)$.

Bareinboim and Tian [17] then present an algorithm for recovery of causal effects under selection bias in ADMGs, which Correa et al. [49] prove is complete. This means it is precisely known when a causal effect is recoverable from selection biased data given the ADMGs.

Recoverability with External Data On the other hand, complete conditions for the recoverability of causal effects under selection bias with external data are not known. As such, we will present a brief review of the literature in this area. Correa and Bareinboim [47] were the first to present a general approach for recovering causal effects from selection bias with access to external data, presenting a variation on the standard adjustment formula. This was then extended further by Correa et al. [48] to be applicable in ADMGs. Finally, Correa et al. [49] provided an algorithm for the recovery of causal effects with unbiased data which extends both of these and is the current state of the art. Their method relies on splitting the distribution into Q factors, which are the distributions on subgraphs of vertices that can be reached by bi-directed paths. The algorithm then aims to ascertain the identifiability of each Q factor.

1.2.3 Multi-Environment Data

One of the main challenges to performing valid causal inference is having correct knowledge of the causal structure. The standard method for finding the causal structure is a mixture of expert knowledge and causal discovery, as detailed in Section 1.1.4. However this cannot always be relied upon. This is especially true if the data is very high dimensional, as this makes causal discovery expensive and increases the probability of errors. One solution to this problem is the method of invariant causal prediction [163] which leverages multi-environment data, and is the focus of this section.

Multi-environment data is data from different experimental settings or contexts. In order for this to be beneficial for the learning of causal relationships there must be

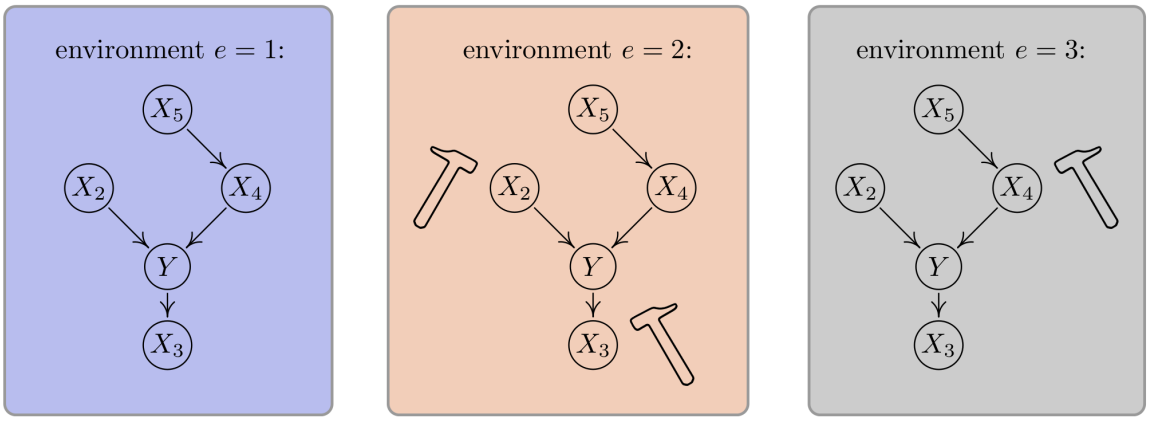


Figure 1.9: In multi-environment data setting we often imagine different environments and experimental settings as corresponding to interventions on a subset of variables. This figure is taken from Peters et al. [163].

some shared causal structure across sites. Bühlmann [32] and Peters et al. [163] argue that this is to be expected, stating that there is a deep relationship between invariance and causality. These works make the case that the causal relationships are *precisely* the relationships that we should expect to be invariant to changes in environment. As an example, the laws of physics should remain constant regardless of where an experiment is performed. Therefore, we should expect that if we find a particular catalyst speeds up a chemical reaction in Australia, the same thing should be true in Antarctica. How much of an impact the catalyst has may change, but that it has an impact should remain true. On the other hand, spurious correlations such as the colour of the experimenters lab coat should not be correlated with the outcome of the experiment over a large number of different settings. However, it may be if we only have a small number of experiments. For example if the only person who can initially get an experiment to work is a particularly extravagant scientist in a red coat then the colour of the coat will be a correlate to experimental outcome. If we were to repeat this experiment enough times over, we would expect this spurious correlation to disappear.

The goal of invariant causal prediction is to make use of this variation across environments to discover the direct causes of an outcome - graphically speaking the causal parents. As discussed above these are described at the correlations with the outcome which are *consistently observed* across experimental settings. The original invariant prediction paper of Peters et al. [163] makes this precise using linear structural causal

models. Specifically they make the following invariance assumption:

Assumption 1. *Suppose we have data (X^e, Y^e) generated from a collection of environments $e \in \mathcal{E}$. We assume there exists a vector of coefficients $\gamma^* = (\gamma_1, \cdots, \gamma_p)$ such that for all $e \in \mathcal{E}$:*

$$Y^e = \mu + X^e\gamma^* + \epsilon^e, \tag{1.28}$$

Where μ is an intercept term, ϵ^e is a random noise variable drawn from some distribution F_ϵ with zero mean and finite variance, and the distribution over X^e is arbitrary. We let $S^* = \{k : \gamma_k^* \neq 0\}$.

This can be implied by certain causal assumptions, but is notably weaker than having a full causal structure over (Y^e, X^e). As an example of a set of causal assumptions which imply assumption 1, we have the following:

Proposition 7. *Suppose we have a linear structural causal model which generates the distribution across environments and leads to a causal graph $\mathcal{G}_E$ which includes the environment indicator E. Suppose further that E has no causal parents and there is no edge from E to Y. Then Assumption 1 is satisfied with $S^* = \mathrm{Pa}(Y)$.*

Proof. This follows directly from the definition of causal parents and linear structural causal models. □

Now let $\mathcal{E}$ be a set of environments. Depending on how similar the experimental settings are, (γ^*, S^*) may not be the only sets which satisfy this. Therefore, for each set S we say it is a set *plausible causal predictors* if there exists a γ, μ with $\gamma_i \neq 0 \iff i \in S$ and F_ϵ such that we cannot reject the null hypothesis:

$$Y^e = \mu + X^e\gamma + \epsilon^e, \tag{1.29}$$

for $\epsilon^e \sim F_\epsilon$. Write $H_{0,S}(\mathcal{E})$ for this null hypothesis. We then define the set of *identifiable causal predictors* to be the intersection of all sets S that are plausible causal predictors relative to experiments $\mathcal{E}$. Or written in terms of $H_{0,S}(\mathcal{E})$:

$$S(\mathcal{E}) \coloneqq \bigcap_{S:\, H_{0,S}(\mathcal{E}) \text{ true}} S. \tag{1.30}$$

Now, as S^* by definition satisfies assumption 1 it is a plausible causal predictor for the set of experiments $\mathcal{E}$. Therefore, we have that $S^* \subset S(\mathcal{E})$. If we have a statistical test for $H_{0,S}(\mathcal{E})$ then we can construct an estimate for $S(\mathcal{E})$ as follows

$$\hat{S}(\mathcal{E}) := \bigcap_{S:\, H_{0,S}(\mathcal{E}) \text{not rejected}} S. \tag{1.31}$$

Proposition 8. *Assume that the estimator $\hat{S}(\mathcal{E})$ is constructed according to (1.31) with a valid test for $H_{0,S}(\mathcal{E})$ for all sets $S \subseteq \{1, \dots, p\}$ at level α in the sense that for all S, $\sup_{P:\, H_{0,S}(\mathcal{E})\ true} P[H_{0,S}(\mathcal{E})\ rejected] \leq \alpha$. Consider now a distribution P over (Y, X) and consider any γ^* and S^* such that Assumption 1 holds. Then, $\hat{S}(\mathcal{E})$ satisfies*

$$P\Big[\hat{S}(\mathcal{E}) \subseteq S^*\Big] \geq\ 1 - \alpha.$$

Peters et al. [163] then give a variety of tests based on regression coefficients which achieve this. This allows us to use data from multiple experiments to construct datasets which contain a subset of the direct causes with high probability. This was extended to non-linear causal models by Heinze-Deml et al. [97] and to more varied experimental settings by Mooij et al. [154].

1.3 Algorithmic Fairness

We now move on to the background on Algorithmic Fairness. This field aims to understand and mitigate the discriminatory impact that algorithms - based on ML/AI or otherwise - can have when used as part of decision-making processes. It was shot into the academic and public consciousness through Propublica's finding that algorithms used in the US justice system were more likely to incorrectly rank black inmates as high risk of reoffending compared to white inmates [9]. The field grew from this to discuss issues of potential algorithmic discrimination in healthcare, employment, and more standard ML applications like image classification or language modelling. Across all of these fields the goal was roughly the same. We have some algorithm which makes decisions or predictions about a set of individuals. This can be quite broad, varying from prison inmates and an algorithm used to decide parole, or an algorithm that selects which adverts people see on social media. These individuals can be split into groups based on a protected characteristic such as race, gender, or disability. As it is illegal to discriminate based on these characteristics, and algorithms can reproduce historical

discrimination in their training data , we need methods to understand and measure algorithmic discrimination, which can be in turn used to mitigate these problems.

This is usually mathematically formulated in the following way:

- We have access to a training dataset of individuals, where for each individual we observe a set of covariates denotes by X, a set of protected characteristics denoted by A, and an outcome of interest, denoted by Y.
- The goal is to train a model to output some prediction $\hat{Y}$ of Y. We will use our training dataset to fit this model.
- We would like to ensure the model cannot be said to discriminate on the basis of the characteristics A. Said differently, we want to ensure that the model is *fair* relative to A.

The final point was mainly approached by defining a large number of "fairness metrics" which are statistics that aim to measure the level of discrimination in the outputs of a model. We will now detail some of the most popular approaches to this, both causal and non causal.

1.3.1 Non-Causal Fairness Methods

We first survey some of the non-causal approaches to this problem. The first and most widely used fairness statistics revolve around independence statements, often referred to as parity conditions. The most common definitions are detailed below, with how they are measured for binary $Y, \hat{Y}, A$:

Demographic Parity Definition: $\hat{Y} \perp\!\!\!\perp A$

Measured as: $\left|P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1)\right|$

False Positive Rate Parity Definition: $\hat{Y} \perp\!\!\!\perp A \mid Y = 0$

Measured as: $\left|P(\hat{Y} = 1 \mid A = 0, Y = 0) - P(\hat{Y} = 1 \mid A = 1, Y = 0)\right|$

False Negative Rate Parity Definition: $\hat{Y} \perp\!\!\!\perp A \mid Y = 1$

Measured as: $\left|P(\hat{Y} = 1 \mid A = 0, Y = 1) - P(\hat{Y} = 1 \mid A = 1, Y = 1)\right|$

Positive Predictive Parity Definition: $Y \perp\!\!\!\perp A \mid \hat{Y} = 1$

Measured as: $\left|P(Y = 1 \mid A = 0, \hat{Y} = 1) - P(Y = 1 \mid A = 1, \hat{Y} = 1)\right|$

Negative Predictive Parity Definition: $Y \perp\!\!\!\perp A \mid \hat{Y} = 0$

Measured as: $\left|P(Y = 1 \mid A = 0, \hat{Y} = 1) - P(Y = 1 \mid A = 1, \hat{Y} = 0)\right|$

Equalized Odds Definition: $\hat{Y} \perp\!\!\!\perp A \mid Y$

Measured as: $\max\left\{\text{FPR}(Y, A, \hat{Y}), \text{FNR}(Y, A, \hat{Y})\right\}$ or $\text{FPR}(Y, A, \hat{Y}) + \text{FNR}(Y, A, \hat{Y})$ for false positive rate (FPR) and false negative rate (FNR) given above.

Predictive Parity Definition: $Y \perp\!\!\!\perp A \mid \hat{Y}$

Measured as: $\max\left\{\text{NPP}(Y, A, \hat{Y}), \text{PPP}(Y, A, \hat{Y})\right\}$ or $\text{NPP}(Y, A, \hat{Y}) + \text{PPP}(Y, A, \hat{Y})$ for negative predictive parity (NPP) and positive predictive parity (PPP) given above.

Large values of any of these statistics are indicative that at some stage of the process, a difference has formed between the two groups. This difference could be in the process that formed the data, the training of the model, the deployment onto the target population, or a variety of other factors. As such it is not exactly clear if a large value of any one of these statistics corresponds directly to a discriminatory impact from the model. One naive approach to solving this problem would be to demand that all of these statistics are small, so we aim to train the predictor $\hat{Y}$ such that the groups look equal under all the metrics. However, as the following result shows, in all but very restrictive cases this is not possible:

Proposition 9 (Kleinberg et al. [120])**.** *We have the following impossibility results:*

- *If $Y \not\perp\!\!\!\perp A$ then we cannot have predictive and demographic parity simultaneously.*
- *If we have a strictly positive probability distribution with $Y \not\perp\!\!\!\perp A$ then we cannot have predictive parity simultaneously.*

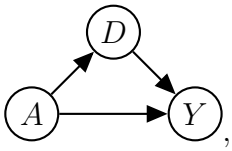
Proof. This follows relatively easily from the graphoid axioms of conditional independence [55]. Specifically the first point follows directly from the contraction axiom and the second from the intersection axiom. □

In most common applications, we cannot expect to have $Y \perp\!\!\!\perp A$. This means we would have to make a choice regarding which of the parity metrics we should satisfy to be "fair" in the given context. Again it is not clear which one of these to choose, or if there even is a correct choice. Many arguments have been made for the strengths and weaknesses of each of these parity metrics in common fairness scenarios [173] with no clear answers emerging. As a result of these problems, alternative metrics were developed which aimed to provide a more overarching view of fairness and prevent practitioners from having to choose between the parity metrics. Whilst there were other non causal approaches such as individual fairness [64] and envy-free fairness [199], we will focus on causal fairness as it is more relevant to this thesis.

1.3.2 Causal Fairness Methods

Inspired by some of the shortcomings of statistical fairness definitions, a new set of approaches to fairness were developed which incorporated causality. These works argued that discriminatory impact is better understood as some kind of causal impact of the protected characteristics of an individual on the predictions of a model. As an example of how causal reasoning may be helpful for fairness problems, we first consider the analysis presented in Pearl [162] of the Berkeley gender discrimination case.

In 1973, the University of Berkeley admitted 44% of male applicants for graduate study but only 35% of female applicants. Over fears that this would lead to the university being sued, they hired a group of statisticians to analyse the data in order to understand the evidence it provided of discrimination in the admissions process [25]. They found that when admissions was segregated by department, there was little to no difference in admissions rate- if anything, women were admitted at a slightly higher rate. This is a classic example of Simpsons paradox, where the apparent relationship between admissions and gender disappears or reverses after conditioning on a third variable, in this case department. Pearl [162] represent this using the following DAG:



, (1.32)

where A is the gender of an applicant, D is the choice of department, and Y is the admissions decision. Stated in this way, Pearl argues the original finding focuses on

the direct causal effect of A on Y, written as:

$$\mathbb{E}\left[Y(a) - Y(a')\right]. \tag{1.33}$$

The switch to conditioning on department can be framed as now targeting the *direct effect* of A on Y, removing the impact that flows through the department choice D. This can be written as:

$$\mathbb{E}\left[Y(a, D(a)) - Y(a', D(a))\right], \tag{1.34}$$

and it is then identified as:

$$\mathbb{E}_{P(D|A=a)}\left[\mathbb{E}\left[Y \mid D, a\right]\right] - \mathbb{E}_{P(D|A=a)}\left[\mathbb{E}\left[Y \mid D, a'\right]\right], \tag{1.35}$$

which corresponds directly to the final target of the study in that we now stratify by department. Viewing things from a causal perspective makes the arguments involved more clear. The argument is that if the causal impact of gender on admission only flows through department and not as a direct impact of gender on outcome it does not constitute discrimination. It is worth noting that this argument can be disagreed with, and often is. However by writing things in terms of causal graphs we have been able to give a more precise statement of the argument involved. This framing is closely related to a notion of path-specific fairness, which we will discuss later. But first, we discuss the most famous causal notion of fairness, counterfactual fairness.

1.3.2.1 Counterfactual Fairness

Intuitively when we think about fairness, we often frame problems of discrimination through the lens of a counterfactual. Arguments such as "this would not have happened if they were a man" or "this only happened because this person is black" are often used to understand if an incident is discriminatory. Counterfactual fairness [128] aims to use causal inference to make such statements precise. It is defined as follows:

Definition 6. *We say a prediction $\hat{Y}$ is counterfactually fair if we have:*

$$P(\hat{Y}(a) \mid X = x, A = x) = P(\hat{Y}(a') \mid X = x, A = x), \tag{1.36}$$

For all x, a, a'.

This definition aims to solidify the notion that the probability of any given outcome should be the same in the observed case as in a counterfactual world where an individuals sensitive attribute had been different, given everything else was kept constant. Counterfactual fairness is strongly related to the notion of actual causation [92], a field which aims to understand what causes a particular event.

In practice, as this definition relies on counterfactual statements we cannot estimate the degree to which a predictor satisfies counterfactual fairness without access to a structural causal model. However, there are some special settings where we can guarantee counterfactual fairness without access to a structural causal model, for example:

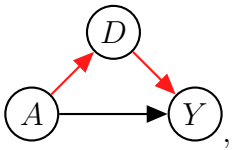
Proposition 10 (Kusner et al. [128])**.** *Let* $\hat{Y} = f(\text{Nd}(A))$ *where* $\text{Nd}(A)$ *are the non descendants of* A*. Then* $\hat{Y}$ *is counterfactually fair.*

We will discuss counterfactual fairness extensively in Chapter 4.

1.3.2.2 Path-Specific Fairness

Whilst counterfactual fairness was able to provide formal causal definitions of fairness,in its basic form it cannot account for the argument made by Pearl [162] at the start of this section regarding the Berkeley gender discrimination case. The crucial part that is missing is the argument that some causal influences are discriminatory, whilst others are not. Counterfactual fairness presents every causal influence of the protected attribute on the prediction as problematic. To account for "fair" causal influences on predictions, we need the notion path-specific effects.

Path Specific Effects Path-specific effects or mediated effects occur when we set a particular variable to one value for one set of paths and to a different value for another set of paths. For example, revisiting the Berkeley example with the following graph



, (1.37)

we have that the causal impact we were interested in is the effect that flows through the black edge alone and not the red path. To make this precise, we have the following definition of a path-specific effect, take from Malinsky et al. [147]:

Definition 7. *Fix a set of causal paths π where each path intersects A exactly once. Now for two values a, a' of A we define the potential outcome, $V_i(\pi, a, a')$, resulting from setting $A = a$ on the paths π for any $V_i \in V$ as follows:*

$$V_i(\pi, a, a') \coloneqq a \text{ if } V_i \in A \tag{1.38}$$

$$V_i(\pi, a, a') \coloneqq V_i\left(\left\{V_j\left(\pi, a, a'\right) \mid V_j \in \ \mathrm{Pa}_i^{\pi}\right\}, \left\{V_j\left(a'\right) \mid V_j \in \ \mathrm{Pa}_i^{\bar{\pi}}\right\}\right) \tag{1.39}$$

where Pa_i^{π} are the parents of V_i along a path in π and $\mathrm{Pa}_i^{\bar{\pi}}$ is the set of other causal parents.

So once again for path-specific effects we have defined the potential outcome via recursive substitution. Now having introduced path-specific counterfactuals, we now discuss the important notion of edge consistency:

Definition 8. *A counterfactual $V_i(\pi, a, a')$ is said to be edge inconsistent if the expression in Equation 1.39 contains counterfactuals of the form $V_j(a_k, \cdots)$ and $V_j(a'_k, \cdots)$ for some variable V_j. If the counterfactual is not edge inconsistent, it is said to be edge consistent.*

In fact, edge consistency is exactly the condition that is required for the identifiability of counterfactuals:

Theorem 11 (Shpitser and Tchetgen Tchetgen [195])**.** *A path-specific counterfactual $V_i(\pi, a, a')$ is identifiable under a given causal graph $\mathcal{G}$ if and only if it is edge consistent. In which case it is given by:*

$$p\left(V\left(\pi, a, a'\right)\right) = \prod_{i=1}^{K} p\left(V_i \mid a \cap \mathrm{pa}_i^{\pi}, a' \cap \mathrm{pa}^{\bar{\pi}}, \mathrm{Pa}_i^{\mathcal{G}} \backslash A\right). \tag{1.40}$$

Malinsky et al. [147] demonstrated that this is equivalent to standard identification of interventional distributions in a particular extended causal graph. In the extended causal graph, $\mathcal{G}^e$, we have an extra node for every edge, which has the same parent and child as the original edge. Interventions then correspond to setting these edge value to be a different value of its parent, allowing for multiple different interventional values of a given variable to be taken according to paths. Any edge consistent path-specific intervention can be framed as a normal intervention in this graph. This allows the standard PO calculus and algorithms for identifying conditional causal effects to be used as complete algorithms for the identification of conditional path-specific effects. As example of an extended causal graph is given in Figure 1.10.

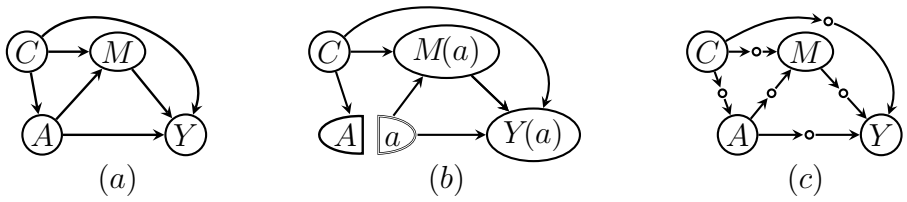


Figure 1.10: (a) A simple causal DAG $\mathcal{G}$, with a treatment A, an outcome Y, a vector C of baseline variables, and a mediator M. (b) A SWIG $\mathcal{G}(a)$ derived from (a) corresponding to the world where A is intervened on to value a. (c) An extended graph $\mathcal{G}^e$ derived from (a).

Path Specific Fairness Definitions Having now given the definition of path-specific effects, we can now define path-specific fairness as follows:

Definition 9 (Nabi and Shpitser [158])**.** *A prediction $\hat{Y}$ is said to be fair relative to a set of causal paths π if we have:*

$$\mathbb{E}\left[\hat{Y}(\pi, a, a')\right] - \mathbb{E}\left[\hat{Y}(a')\right] = 0 \tag{1.41}$$

for all a, a'. Or said differently, if the path-specific effect of the protected attribute along π is zero.

Although it may seem that this definition is restricted by the need to specify a set of paths π, this is actually viewed as a strength of the method. This allows us to incorporate normative judgments on fair / unfair influences, as Pearl does for the Berkeley gender discrimination case, into the set of paths π. Path-specific fairness was also extended to not just effects but path-specific counterfactual fairness in Chiappa [41].

1.4 Thesis outline

This thesis contains 4 published papers and a conclusion section. Even though they did not make it into this thesis, I have been lucky enough to be part of a number of other publications during my PhD, which are listed here:

- Doubly Robust Kernel Statistics for Testing Distributional Treatment Effects . **Jake Fawkes**, Robert Hu, Robin J. Evans, Dino Sejdinovic. Published at TMLR [74].

- The Role of Learning Algorithms in Collective Action. Omri Ben-Dov*, **Jake Fawkes***, Samira Samadi, Amartya Sanyal. Published at ICML 2024 [29].
- Returning the favour: when regression benefits from probabilistic causal knowledge. Shahine Bouabid*, **Jake Fawkes***, Dino Sejdinovic. Published at ICML 2023 [29].

With a brief introduction, the four papers featured in this thesis are:

The Hardness of Validating Observational Studies with Experimental Data. ***Jake Fawkes****, Michael O'Riordan, Athanasios Vlontzos, Oriol Corcoll, Ciarán Mark Gilligan-Lee.* Published at AISTATS 2025 [78].

In this paper, we discuss the problem of using small quantities of experimental data to partially resolve the issue of unmeasured confounding in large observational studies. Existing methods in this field aim to do this by either debiasing the observational study, or using the experimental data to benchmark the level of unmeasured confounding. We theoretically analyse these both of these approaches, showing there are fundamental limitations to trying to do either of them in an assumption-free manner. We identify a lack of smoothness in the unknown correction function as the cause of this problem. Based upon this, we propose a novel Gaussian process based approach, which is able to produce a region that contains the true treatment effect with high probability.

Is merging worth it? Securely Evaluating the Information Gain for Causal Dataset Acquisition. ***Jake Fawkes****, Lucile Ter-Minassian*, Desi Ivanova, Uri Shalit, Chris Holmes.* Published at AISTATS 2025 [75].

This work proposes a novel problem of selecting from a number of datasets which one best to merge with your current dataset, based on an improvement in estimation of conditional treatment effects. We aim to construct a methodology that can not only select datasets, but be evaluated in a secure and privacy-preserving manner. To evaluate the improvement from selecting a given dataset, we draw inspiration from Bayesian experimental design, aiming to evaluate the expected information gain in causal model parameters if we were to select a particular site. In order to ensure this can be done securely and privately, we make use of multi-party communication- a cryptographic technique which allows functions to be evaluated where a number of parties have secret

inputs- and differential privacy jointly. We build this methodology on top of a number of the most popular Bayesian CATE estimation techniques, specifically polynomial regression, Bayesian causal forests, and causal multi-task Gaussian processes.

Selection, Ignorability, and Challenges with Causal Fairness. ***Jake Fawkes****, Robin Evans, Dino Sejdinovic.* Published at the Conference on Causal Learning and Reasoning 2022 [76].

This chapter is the first of two discussing the problem of data quality in fair machine learning. In this work we focus on the problem selection bias presents to causal fairness methods. We argue that because many of the traits we are often interested in being fair in relation to affect outcomes from the very early stages of people lives, we are almost always working with data that is subject to heavy selection bias. We argue that this is a particular problem for causal fairness as it means that DAG like models cannot correctly capture any of the causal quantities of interest. Despite this, many methods in the field of causal fairness continue to use DAGs. We argue that continuing to do so in light of the problems due to selection bias leaves models with little to no causal interpretation.

The Fragility of Fairness: Causal Sensitivity analysis for Fair Machine Learning. ***Jake Fawkes*****, Nic Fishman*, Mel Andrews, Zachary C. Lipton* Published at the Neurips 2024 track on Datasets and Benchmarks [77].

This work extends the results of the previous chapter, looking at the problems that measurement biases and data quality present for the field of fair machine learning as a whole. To do this, we focus on three biases in particular: i)Proxy Label Bias , ii) Selection bias, and iii) Extra Classificatory policies- that is policies in the control of firms which impact outcomes they aim to measure. In order to gain understanding of the issues these biases present to the field of fair machine learning, we survey a selection of the most popular datasets in the field, aiming to asses if they fall victim to any of these biases. We find that 100% of the datasets express at least one of these biases, and 60% contain all three simultaneously. In light of this, we propose the use of causal sensitivity analysis for practitioners to reason about the effects that such biases can have on the evaluation of statistical fairness metrics. Leveraging recent advances in discrete causal sensitivity analysis, we construct sensitivity analysis tools which

can simultaneously handle all of the measurement biases discussed in the paper. We apply these tools to reproduce existing results in the literature on sensitivity analysis for fairness, before demonstrating that we can vary assumptions and extend existing results.

2

The Hardness of Validating Observational Studies with Experimental Data

Abstract

Observational data is often readily available in large quantities, but can lead to biased causal effect estimates due to the presence of unobserved confounding. Recent works attempt to remove this bias by supplementing observational data with experimental data, which, when available, is typically on a smaller scale due to the time and cost involved in running a randomised controlled trial. In this work, we prove a theorem that places fundamental limits on this "best of both worlds" approach. Using the framework of impossible inference, we show that although it is possible to use experimental data to *falsify* causal effect estimates from observational data, in general it is not possible to *validate* such estimates. Our theorem proves that while experimental data can be used to detect bias in observational studies, without additional assumptions on the smoothness of the correction function, it can not be used to remove it. We provide a practical example of such an assumption, developing a novel Gaussian Process approach to construct intervals which contain the true treatment effect with high probability, both inside and outside of the support of the experimental data. We demonstrate our methodology on both simulated and semi-synthetic datasets and make the code available.

2.1 Introduction

It is often said that randomised controlled trials (RCTs) are the gold standard for establishing causal relationships [94, 88], and estimating treatment effects [11, 87]. However, in many cases, it is prohibitively costly, slow, or even unethical to run experiments that are large enough to accurately estimate such effects. Meanwhile, observational data is abundant, being significantly easier and cheaper to obtain. Unfortunately, such data is often subject to unmeasured confounding. This leaves treatment effects unidentifiable, and naively attempting to use the observational data despite this will lead to biased estimates of causal effects.

As a response to these problems, there has been substantial recent research effort to develop methodology for combining experimental and observational data sources, aiming to get the strengths of each [46, 138, 207, 109]. These approaches aim to use the experimental data to de-bias the observational data [225, 112], falsify observational studies [106, 105], or benchmark the level of unmeasured confounding [56, 57].

In this work, we prove a selection of fundamental limitations of this approach. We use the framework of impossible inference [13, 24], a popular tool in econometrics [35]. This field studies when it is possible for a non-trivial hypothesis tests to exist for a problem, where trivial hypothesis tests are those which are unable to distinguish *any* alternative from the null. We apply this to show that whilst non-trivial tests exist to *falsify* estimates from observational studies, we cannot *validate* heterogeneous treatment effects estimates using experimental data without additional assumptions. In terms of benchmarking confounding with a causal sensitivity model, our result corresponds to the statement that it is possible to form valid *lower bounds* of the sensitivity parameter but that it is impossible to form non-trivial upper bounds—again, absent additional assumptions.

Our hardness proof relies on the lack of smoothness in the unknown correction function. Therefore, as an example of an assumption that does permit this type of inference, we take the corrective function to be a sample from a Gaussian Process, a probabilistic function family with inherent smoothness guarantees. Developing this into a workable and practical methodology leads us to create a novel Gaussian Process based approach to learning from pseudo-outcomes [116], which correctly accounts for the unwieldy error distribution of pseudo-outcomes. We experimentally evaluate our approach against other Gaussian Process effect estimation approaches, showing strong improvement

in predictive performance and uncertainty calibration. In short, to summarise our contributions:

- A proof of the limits of current approaches that use a learned corrective term to reconcile experimental and observational data.
- A demonstration that the smoothness properties of Gaussian Processes circumvent the assumptions required for the above proof and provide guarantees that Gaussian Processes give intervals which contain the treatment effect over the whole observational support with high probability.
- A novel Gaussian Process method to learn from inverse propensity weighted pseudo-outcomes—which may be of independent interest.
- Extensive experimentation validating the aforementioned novel method on synthetic and semi-synthetic data.

2.2 Background and Notation

2.2.1 Notation

We let the random variables X, T, and Y represent the covariates, treatment, and outcomes, with domains $\mathcal{X}$, $\{0,1\}$, and $\mathbb{R}$ respectively, and use $\mathbf{x}$, t, and y to denote realisations of the variables. We suppose we have two datasets, the observational $\mathcal{D}_e = \{\mathbf{x}_i^o, t_i^o, y_i^o\}_{i=1}^{n_o}$ and the experimental $\mathcal{D}_o = \{\mathbf{x}_i^e, t_i^e, y_i^e\}_{i=1}^{n_e}$ which are drawn from distributions $P_e(\mathbf{x}, t, y)$ and $P_o(\mathbf{x}, t, y)$ respectively, with $\mathcal{D} = \mathcal{D}_o \cup \mathcal{D}_e$ denoting the full dataset of size $n = n_o + n_e$. We will use a variable E to denote if we are in the observational or experimental regime, so that, for example, $P_e(\mathbf{x}, t, y) = P(\mathbf{x}, t, y \mid E = e)$. Letting $\star \in \{o, e\}$ we use $\mathcal{X}^\star \subset \mathcal{X}$ for the support of $P_\star(\mathbf{x})$ and assume that $\mathcal{X}^e \subset \mathcal{X}^o$. Vectors of observations and treatment in a dataset, $\mathcal{D}_\star$, are denoted by $\mathbf{y}_\star$ and $\mathbf{t}_\star$ respectively, while $\mathbf{X}_\star$ refers to the data matrix, so that $\mathbf{y}_\star = (y_i)_{i=1}^{n_e}$, $\mathbf{t}_\star = (t_i)_{i=1}^{n_e}$, and $\mathbf{X}_\star = (\mathbf{x}_i)_{i=1}^{n_e}$. We let $\omega_\star(\mathbf{x})$ be defined as:

$$\omega_\star(\mathbf{x}) = \mathbb{E}[Y|X = \mathbf{x}, T = 1, E = \star] \tag{2.1}$$

$$- \mathbb{E}[Y|X = \mathbf{x}, T = 0, E = \star], \tag{2.2}$$

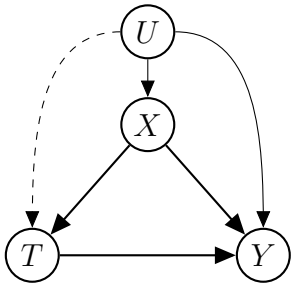


Figure 2.1: Causal Structure for generating the experimental and observational datasets. Dashed edges are only present in the observational dataset, whilst all others are present and fixed across both datasets.

be the difference between expected conditional outcomes for each treatment conditional on $E = \star$ and $X = \mathbf{x}$. We use potential outcomes [182], so that $Y(t)$ represents the outcome from setting $T = t$[1].

2.2.2 Objectives and Assumptions

Throughout we focus on estimating the *Conditional Average Treatment Effect* (CATE) for the observational datasets, which is given by:

$$\tau(\mathbf{x}) \coloneqq \mathbb{E}\left[Y(1) - Y(0) | X = \mathbf{x}, E = o\right]. \tag{2.3}$$

We assume the datasets are generated according to the causal structure in Figure 2.1, where dashed edges are present only in the observational dataset. Importantly, this implies $Y(t) \perp E \mid X = \mathbf{x}$, which ensures that the CATE is fixed across environments as:

$$\tau(\mathbf{x}) = \mathbb{E}\left[Y(1) - Y(0) | X = \mathbf{x}, E = o\right] \tag{2.4}$$

$$= \mathbb{E}\left[Y(1) - Y(0) | X = \mathbf{x}, E = e\right]. \tag{2.5}$$

We demonstrate this in Appendix 2.A.1.

For the experimental study we assume that treatment is randomised according to a known propensity score $\pi(\mathbf{x}) = P_e(T = 1 \mid X = x)$ which we assume to satisfy *strict overlap* [66]. That is we assume that there exists a $\delta > 0$ such that:

$$\delta < \pi(\mathbf{x}) < 1 - \delta \text{ for all } \mathbf{x} \in \mathcal{X}^e. \tag{2.6}$$

[1]We make use of the SWIG framework to combine causal graphical models with potential outcomes. More details can be found in Richardson and Robins [175].

Under the additional assumption of consistency ($Y(t) = Y$ when $T = t$) we have that the CATE is *identified* [162, 175] within the support of the experimental dataset, and given by:

$$\tau(\mathbf{x}) = \omega_e(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{X}^e. \tag{2.7}$$

This is not the case in the observational dataset, where the hidden confounding induced by U means that $\tau(\mathbf{x}) \neq \omega_o(\mathbf{x})$ in general. We use $\Delta(\mathbf{x})$ to denote this gap as:

$$\Delta(\mathbf{x}) = \tau(\mathbf{x}) - \omega_o(\mathbf{x}). \tag{2.8}$$

So that if $\Delta(\mathbf{x}) = 0$, an unbiased estimate of $\omega_o(\mathbf{x})$ from the observational study is an unbiased estimate of the true CATE. Throughout, we will assume a model $\hat{\omega}_o(\mathbf{x})$ has been fit for $\omega_o(\mathbf{x})$ from the observational sample and let $\hat{\Delta}(\mathbf{x}) \coloneqq \tau(\mathbf{x}) - \hat{\omega_o}(\mathbf{x})$.

The idea is that $\Delta(\mathbf{x})$ should be simpler than the true CATE function, $\tau(\mathbf{x})$. Under such circumstances, it should be more efficient to use to the experimental dataset to estimate or bound $\hat{\Delta}(\mathbf{x})$ and combine it with $\hat{\omega_o}(\mathbf{x})$ than use the small experimental dataset to learn the CATE directly[225]. Ideally, we would want these to be extendable from the support of the experimental distribution to the support of the observational distribution [112], potentially by incorporating bounds on the correction function as opposed to point estimates. This would give us expression for CATE over all of $\mathcal{X}^o$.

Finally, we introduce the IPW pseudo-outcome [116, 53], which throughout we will only refer to relative to the experimental distribution, as follows:

Definition 10 (IPW Pseudo-Outcome)**.** *The IPW Pseudo-Outcome is given by:*

$$\tilde{Y} \coloneqq \left(\frac{T - \pi(X)}{\pi(X)\,(1 - \pi(X))} \right) Y \tag{2.9}$$

Where $\pi(X) = P(T = 1 \mid X, E = e)$*.* $\tilde{Y}$ *has the property that* $\mathbb{E}[\tilde{Y}|X = \mathbf{x}, E = e] = \tau(\mathbf{x})$*.*

We assume that $\pi(X)$ is known, which is common for a well conducted randomised controlled trial.

2.2.3 Related work bounding $\Delta(\mathbf{x})$

Our setting is first considered in Kallus et al. [112], where they assume that $\Delta(\mathbf{x})$ is linear in order to allow for extrapolation from the experimental sample. By taking $\Delta(\mathbf{x}) = \beta_0^\top \mathbf{x}$ and assuming β_0 is identifiable from experimental data, they can obtain an estimate $\Delta(\mathbf{x})$ that generalises beyond the experimental sample by performing a linear regression of $\{\mathbf{x}_i^e\}_{i=1}^{n_e}$ onto $\{\tilde{y}_i^e - \hat{\omega}_o(\mathbf{x})\}_{i=1}^{n_e}$. Kallus et al. [112] prove that as the number of experimental and observational samples tend to infinity this will converge to the true CATE at a faster rate than using the experimental sample alone. This has been extended to the semiparametric [225] and nonparametric case [221], however extrapolation still requires the function to be uniquely identified from the experimental study.

A strongly related area of work aims to use data from the experimental study to test causal effect estimates from observational studies. One approach aims to falsify causal estimates from observational studies using experimental data [105, 106]. This work converts a variety of assumptions regarding the validity of the observational study, consistency of the CATE across studies and external validity of the RCT into testable statistical hypothesis, which can then be falsified. Another approach aims to estimate how much unmeasured confounding must be present in an observational study for it to be consistent with the RCT [56, 57]. This is achieved using causal sensitivity models [181], which uses a single parameter to control the strength of unmeasured confounding. The goal is then to use the RCT to lower bound this parameter. A significant portion of work in both these areas utilises non-parametric tests of conditional moment restrictions [155].

2.3 The Hardness of Validating Observational Studies

We now provide some theoretical limits on using experimental data to measure the level of unmeasured confounding in an observational study. We do this using the framework of impossible inference [24], a popular tool in econometrics [35]. This field studies when it is possible for a non-trivial hypothesis tests to exist for a problem, where trivial hypothesis tests are those which are unable to distinguish *any* alternative from the null.

Impossible inference has been applied within causal inference to show the hardness of conditional independence testing [191]. In our case, the conditional independence $Y \perp E \mid X, T$ would imply a total lack of unmeasured confounding as:

$$\begin{aligned} P_o(Y \mid X = \mathbf{x}, T = t) &= P_e(Y \mid X = \mathbf{x}, T = t, E = e) \\ &= P(Y(t) \mid X = \mathbf{x}, E = e). \end{aligned}$$

Therefore, the hardness of this testing problem already demonstrates that there are no non-trivial tests for full unconfoundedness in the observational distribution.

However, this still doesn't preclude us from being able to estimate CATE in an unbiased manner from the observational dataset. Or, failing full estimation, bounding the correction function, $\hat{\Delta}(\cdot)$ to return intervals for CATE. Therefore, in this section, we apply the same techniques to focus on what experimental data allows us to test for in term unbiasedness of CATE estimates. This translates to estimating or bounding $\hat{\Delta}(\mathbf{x})$ with confidence guarantees, which we do via the following definition:

Definition 11. *We say that that $\hat{\Delta}(\cdot)$ is **controlled** by functions $\bar{f}, \underline{f} : \mathcal{X} \to \mathbb{R}$ if we have:*

$$\hat{\Delta}(\mathbf{x}) \in \left[\underline{f}(\mathbf{x}), \bar{f}(\mathbf{x})\right] \;\; a.s \text{ for } x \sim P_o(\mathbf{x}) \tag{2.10}$$

Where $\underline{f}(\mathbf{x}) \leq \bar{f}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$.

If the unmeasured confounding were controlled by $\underline{f}(\mathbf{x}) = \bar{f}(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$ then we would have that the CATE estimate from the observational study is unbiased to the true CATE. Going further than this, if we knew functions that controlled the confounding, then we could use them alongside the CATE estimate from the observational study to give intervals which contain the true CATE. Therefore, the goal is to understand how we can use the observational study to learn the functions, $\bar{f}, \underline{f}$, that control the confounding.

2.3.1 Testing Notation and Background

First, we will assume a fixed propensity score function $\pi : \mathcal{X} \to [0, 1]$. Now we define $\mathcal{E}_{M,\pi}$ to be the set of distributions over (X, T, Y) which are absolutely continuous in X with respect to Lebesgue measure, bounded above in ℓ_∞ norm by M, and whose propensity score is given by π. As before, the domains of T, Y, E are $\{0, 1\}, \mathbb{R}$ and

$\{o, e\}$ respectively. For this section, we will use the notation $\mathbb{P}_P$ to denote a probability taken with respect to $P \in \mathcal{E}_{M,\pi}$.

A potentially randomised test ψ_n is a measurable function which takes in a dataset, $\mathcal{D}$, a random variable $U \sim U[0,1]$ that represents the randomness of the test—a choice of permutations in permutation testing, for example—and outputs a result in $\{0,1\}$ where 1 corresponds to a rejection. So we write $\psi_n(\mathcal{D}, U)$ for the result of the test ψ_n with dataset $\mathcal{D}$ and U as input.

Suppose we observe an experimental dataset $\mathcal{D}$ sampled i.i.d from a distribution $P_0 \in \mathcal{E}_{M,\pi}$ and wish to test the null hypothesis $H_0 : P_0 \in \mathcal{N} \subset \mathcal{E}_{M,\pi}$ against the alternative hypothesis $H_1 : P_0 \in \mathcal{A} \subset \mathcal{E}_{M,\pi}$. Then we have the following important definitions:

Definition 12. *Let ψ_n be a randomised test which takes in a dataset $\mathcal{D}$ of size n. We say ψ_n has **level** α at size n if we have $\sup_{P \in \mathcal{N}} \mathbb{P}_{\mathcal{D} \sim P^n}(\psi_n(\mathcal{D}, U) = 1) \leq \alpha$. For an alternative distribution $P \in \mathcal{A}$, we define $\mathbb{P}_{\mathcal{D} \sim P^n}(\psi_n(\mathcal{D}, U) = 1)$ as the **point-wise power** against P at size n.*

Ideally, we would want a test to have power against as many alternatives as possible, preferably uniformly so that $\inf_{P \in \mathcal{A}} \mathbb{P}_{\mathcal{D} \sim P^n}(\psi_n(\mathcal{D}, U) = 1) \to 1$. For non-parametric hypothesis, restrictions on the alternative such as smoothness conditions are often required to achieve this [14].

A particular problem given by sets $(\mathcal{N}, \mathcal{A})$ is known as **untestable** if for all tests the point-wise power is bounded by the level for any alternative. In this case, there exists no test that can distinguish the null from *any* alternative, therefore the null has to be restricted for there to be an informative test.

2.3.2 Setting the Testing Problem for Unmeasured Confounding

We now apply the above to testing for bias in treatment effect estimation due to unmeasured confounding. Fixing the observational estimate $\hat{\omega}(\cdot)$ and so the correction function $\hat{\Delta}(\cdot)$, we define the following sets of distributions:

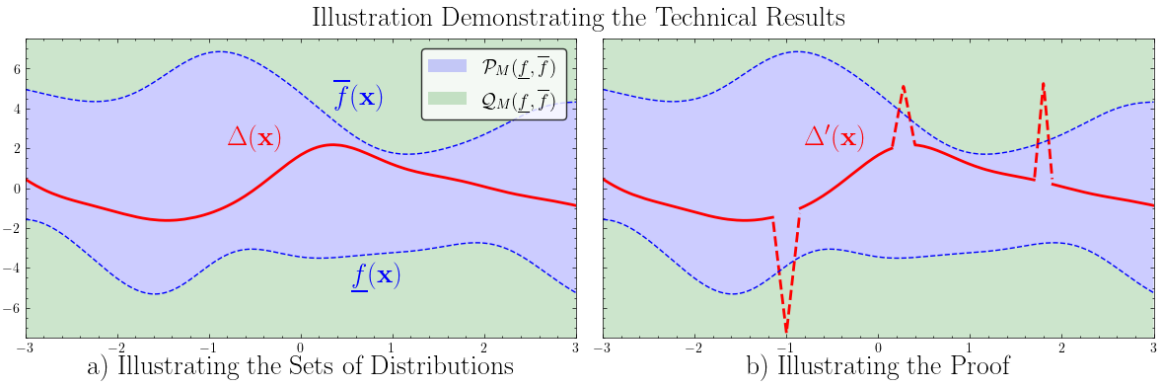


Figure 2.2: Illustration of both the sets of distributions and proof of the technical result in Section 2.3. The first figure demonstrates the sets of distributions $\mathcal{P}_{M,\pi}(\underline{f},\bar{f}), \mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$. $\mathcal{P}_{M,\pi}(\underline{f},\bar{f})$ is the set of distributions where $\Delta(\cdot)$ is always contained in the blue region, and $\mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$ is the set of all other distributions, so those where $\Delta(\cdot)$ leaves the blue region. To prove the hardness of validating observational study estimates, we show that for any $P \in \mathcal{P}_{M,\pi}(\underline{f},\bar{f})$ we can find distributions $Q \in \mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$ that are arbitrarily close by adding spikes as in Figure b.

Definition 13. *Let* $\underline{f},\bar{f} : \mathcal{X} \to \mathbb{R}$. *We define:*

$$\mathcal{P}_{M,\pi}(\underline{f},\bar{f}) = \left\{ P \in \mathcal{E}_{M,\pi} : \ \hat{\Delta}(\cdot) \textit{ controlled by } \underline{f},\bar{f} \ \right\}$$
$$\mathcal{Q}_{M,\pi}(\underline{f},\bar{f}) = \mathcal{E}_{M,\pi} \setminus \mathcal{P}_{M,\pi}$$

Which way round to test? Now with both sets of distributions , the question is what to take as the null and alternative? We have the following choices:

$$\text{Test 1:} \left\{ H_0 : P \in \mathcal{P}_{M,\pi}(\underline{f},\bar{f}), \quad H_1 : P \in \mathcal{Q}_{M,\pi}(\underline{f},\bar{f}), \right.$$
$$\text{Test 2:} \left\{ H_0 : P \in \mathcal{Q}_{M,\pi}(\underline{f},\bar{f}), \quad H_1 : P \in \mathcal{P}_{M,\pi}(\underline{f},\bar{f}), \right.$$

Rejecting under test 1 is more standard, and would correspond to *falsifying the hypothesis* that the observational study has a level of confounding controlled by $\underline{f},\bar{f}$. This would mean finding statistical evidence that bias in the observational CATE estimate is not contained in the intervals given by $(\underline{f},\bar{f})$.

However, failing to reject under test 1 *does not provide evidence* that the confounding is controlled by $(\underline{f},\bar{f})$. We may fail to reject because of other reasons, such as a lack of data or the test having limited power against the true distribution. In an ideal world, we would like to reject the hypothesis that confounding is above a certain level. Fixing

this level using some $\underline{f}, \bar{f}$, this would correspond to test 2, where the data is used to reject the hypothesis that $\hat{\Delta}$ is not controlled by $\underline{f}, \bar{f}$.

The formulation of test 2 is strongly related to bioequivalence testing[2] [218, 45]. In this field, the goal is to find statistical evidence that one medical treatment works almost equivalently to another, where there is some tolerance specified due to working with finite samples. This is used to approve generic drugs, which have the same active ingredient as a branded drug but can only be sold once the branded drug's patent expires. Here the aim is to ensure the generic drug works as well as the branded one, and so consumers can use them interchangeably.

In our context, we have a similar goal, in that we would like to show that the observational CATE estimate with the adjustment from the RCT is "good enough" up to some tolerance. We specify this tolerance by functions, $(\underline{f}, \bar{f})$, that control Δ as in Definition 11.

Following this discussion we provide the following definitions, we refer to tests of type 1 as **falsification tests** and tests of type 2 as **equivalence test**.

2.3.3 Limits on Testing

Having laid out the two testing problems in Section 2.3.2, we now apply the impossible inference framework detailed in Section 2.3.1 to these problems. Firstly, we demonstrate that whilst equivalence testing is more aligned with the aim of the field, the problem as set out is untestable:

Theorem 12. *Fix any $\underline{f}, \bar{f} : \mathcal{X} \to \mathbb{R}$ and let ψ_n be an equivalence test with null $Q_M(\underline{f}, \bar{f})$ and alternative $\mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$. If the level of this test is, α we have that:*

$$\mathbb{P}_P(\psi_n = 1) \leq \alpha, \tag{2.11}$$

for any $P \in \mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$. That is ψ_n does not have power against any alternative.

This shows that *any* equivalence test which has level α against the null will fail to distinguish *any* alternative. This means that there is no hypothesis test that can confirm from data that the difference function Δ is controlled by any pair of functions $(\underline{f}, \bar{f})$. We visualise the proof of Theorem 12 in Figure 2.2. The idea is that for any distribution $P \in \mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$, we can construct a series of distributions $\{Q_i\}_{i=1}^{\infty} : Q_i \in \mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})$

[2] Also referred to as simply equivalence testing.

that tends to P in total variation distance. Therefore, not statistical procedure can distinguish the two.

This result has implications for falsification tests:

Corollary 13. *For fixed $\underline{f}, \bar{f} : \mathcal{X} \to \mathbb{R}$, any falsification test ψ_n with null $\mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$ and alternative $\mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})$ has:*

$$\inf_{Q \in \mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})} \mathbb{P}_Q(\psi_n = 1) \leq \alpha, \tag{2.12}$$

where α is the level of ψ_n.

This means that for any n, any falsification test will always fail to distinguish some set of alternatives from the null with power distinctly above the level. However, the next result shows there are alternatives which can be distinguished from the null in falsification tests:

Proposition 14. *There exists a distribution $Q \in \mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})$ such that:*

$$\mathrm{TV}(Q, \mathrm{co}(\mathcal{P}_{M,\pi}(\underline{f}, \bar{f}))) \geq \beta, \tag{2.13}$$

for some $\beta > 0$ where $\mathrm{co}(\mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$ is the convex hull of $\mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$ and TV is the total variation distance. Following Bertanha and Moreira [24], this guarantees that there is a test with $\beta + \alpha$ against Q where α is the level of the test.

This demonstrates that, unlike validation, falsification is possible relative to certain alternatives.

2.3.4 Implications for Sensitivity Models

Finally, we apply the results in Section 2.3.3 to approaches that make use of causal sensitivity models to measure the degree of unmeasured confounding. First, defining a generalised sensitivity model as follows:

Definition 14. *A **sensitivity model** is a parameterised set of pairs of functions:*

$$\left\{(\underline{f}_\gamma, \bar{f}_\gamma) : \underline{f}_\gamma, \bar{f}_\gamma : \mathcal{X} \to \mathbb{R}, \gamma \in [\Gamma_0, \Gamma_1]\right\}, \tag{2.14}$$

where for fixed $\mathbf{x}$, $\underline{f}_\gamma(\mathbf{x}), \bar{f}_\gamma(\mathbf{x})$ are continuously decreasing/increasing respectively in γ. Moreover, that $\underline{f}_{\Gamma_0}(\mathbf{x}), \bar{f}_{\Gamma_0}(\mathbf{x}) = 0$ and $\underline{f}_{\Gamma_1}(\mathbf{x}), \bar{f}_{\Gamma_1}(\mathbf{x})$ are $-M$ and M respectively. For a distribution in $P \in \mathcal{E}_M$ we define:

$$\Gamma(P) = \inf\left\{\gamma : (\underline{f}_\gamma, \bar{f}_\gamma) \text{ controls } \Delta(\cdot)\right\} \tag{2.15}$$

Viewed this way, a sensitivity model is a way of constructing intervals around the confounded CATE, $\omega_o(\mathbf{x})$, that contain the true CATE. Moreover, previous work in this area can be seen as aiming to use the experimental dataset to perform inference on $\Gamma(P)$. Specifically, De Bartolomeis et al. [56, 57] both look to use the experimental data to construct probabilistic lower bounds on $\Gamma(P)$. We now apply results Section 2.3.3 to constructing confidence intervals for $\Gamma(P)$, showing non-trivial upper bounds are not possible:

Theorem 15. *Fix a sensitivity model and let $\mathcal{D}$ be a dataset sampled from $P^{(n)}$ where $P \in \mathcal{E}_{M,\pi}$. Let $\left[\underline{C}(\mathcal{D}), \bar{C}(\mathcal{D})\right]$ be a confidence interval for $\Gamma(P)$ in that it satisfies the following coverage requirement:*

$$\inf_{P \in \mathcal{E}_{0,M}} \mathbb{P}_{\mathcal{D} \sim P^{(n)}} \left(\Gamma(P) \in C(\mathcal{D}_n)\right) \geq 1 - \alpha \tag{2.16}$$

Then $\bar{C}(\mathcal{D}) = \Gamma_1$ with probability $1 - \alpha$. That is, there are no non-trivial upper bounds on $\Gamma(P)$.

This demonstrates that in contrast to the lower bound case, non-trivial probabilistic upper bounds on $\Gamma(P)$ are not possible without further assumptions on the set of distributions, $\mathcal{E}_M$. This creates difficulties in using sensitivity models to benchmark unmeasured confounding, as a lower bound represents the smallest amount of confounding that can explain the data. Therefore, it *does not* allow us to say with confidence that a treatment effect is contained in some interval.

2.4 Pseudo Outcome Gaussian Processes and Uniform Error Bounds

The results presented in Section 2.3 show that without further assumptions, we cannot produce intervals which contain the true CATE with high probability. The proof relied on constructing arbitrary peaks in the correction function, which was possible due to a lack of smoothness. As an example of an assumption that can permit this type of inference, we leverage Gaussian process (GPs) which come with inherent smoothness constraints. We then adapt uniform error bounds for Gaussian processes [80, 133] in order to get functions which control $\Delta(\cdot)$.

Before doing this, we develop a Gaussian process approach to learning CATE from pseudo-outcomes. To this best of our knowledge, this is first example of such a method.

This is important as it allows us to learn the difference function directly from an estimate of $\omega_o(\mathbf{x})$ in the observational study. Alternative causal Gaussian process approaches, such as Alaa and Van Der Schaar [4], would require access to an estimate of $\mathbb{E}[Y \mid X = \mathbf{x}, T = t, E = o]$, to learn a separate correction for each t. By using the pseudo-outcome approach, we sidestep this issue and so allow users full flexibility on how to model $\omega_o(\mathbf{x})$.

2.4.1 Pseudo Outcome Regression with Gaussian Processes

We now turn to designing a GP based pseudo-outcome approach. pseudo-outcomes are designed so that the minimiser of the pseudo-outcome mean squared error is the same as the minimiser of the mean squared error under the true unobserved CATE. If the propensity score is correctly specified, the pseudo mean square error will converge optimally to the true CATE mean square error [116]. However, GP based methods are fit via maximum likelihood, with closed form solutions to the posterior requiring Gaussian errors. This creates a problem for applying them directly to pseudo-outcomes, which have distinctly non Gaussian errors. Specifically, they may be written as:

$$\tilde{Y} = \tau(X) + \epsilon \tag{2.17}$$

Where $\mathbb{E}[\epsilon \mid X] = 0$ but $\mathbb{E}[\epsilon \mid X, T] \neq 0$. As we show in Appendix 2.C, this breaks Gaussianity assumptions even in cases where the errors on Y are Gaussian. However, the next proposition suggests a simple correction term which allows us to recover well-behaved errors:

Proposition 16. *There exists a function $\phi : \mathcal{X} \to \mathbb{R}$ such that the IPW pseudo-outcome can be written as:*

$$\tilde{Y} = \tau(X) + \left(\frac{T - \pi(X)}{\pi(X)\left(1 - \pi(X)\right)} \right) \phi(X) + \tilde{\epsilon}$$

Where $\mathbb{E}[\tilde{\epsilon} \mid X, T] = 0$ and $\tilde{\epsilon} \mid T, X$ is Gaussian if the original errors on Y are.

Using the decomposition provided by this proposition we may model this using independent Gaussian Processes for $\hat{\Delta}$ and ϕ, so $\hat{\Delta} \sim \mathrm{GP}(0, k_\theta), \phi \sim \mathrm{GP}(0, l_\eta)$. This is equivalent to using a vector valued GP [7] for multitask regression, where we take the three tasks to be predicting the pseudo-outcome when $T = 0$, predicting the

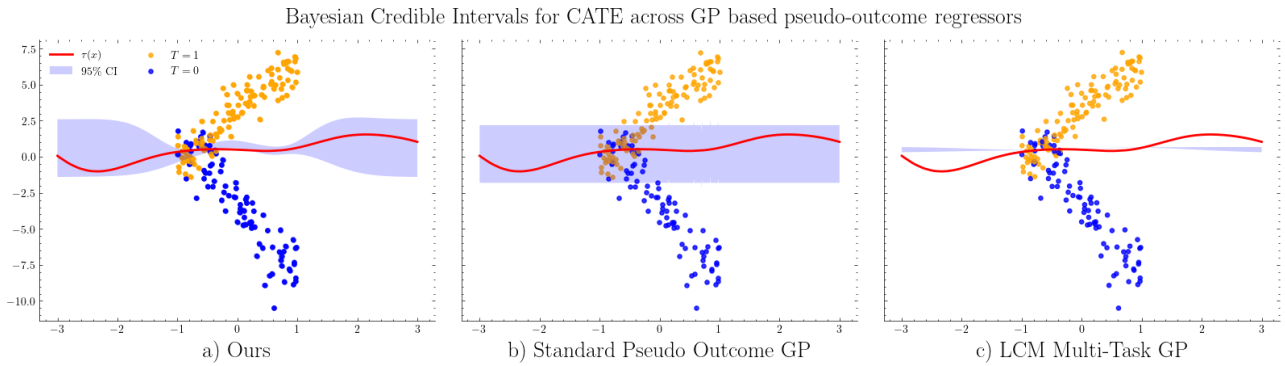


Figure 2.3: A particularly pathological example of the behaviour we observe for each method in our simulated experiment of Section 2.5.1. For the standard GP, hyperparameter optimisation leads to uninformative predictions as it cannot account for close $\mathbf{x}$ values with seemingly no correlation. For the trained LCM, we get strong predictive performance but poor uncertainty quantification, especially out of distribution. Our approach gets the best of both scenarios, with strong predictive performance and calibrated uncertainty out of distribution.

pseudo-outcome when $T = 1$, and finally predicting the CATE. This corresponds to using the following LCM multitask kernel [7]:

$$\mathbf{K} = \mathbf{a}\mathbf{a}^\top k_\theta + \mathbf{b}\mathbf{b}^\top l_\eta \tag{2.18}$$

$$\mathbf{a} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \tag{2.19}$$

$$\mathbf{b} = \begin{bmatrix} \frac{-1}{1-\pi(X)} & \frac{1}{\pi(X)} & 0 \end{bmatrix} \tag{2.20}$$

From here, we can compute the posterior for $\hat{\Delta}$ given the pseudo-outcomes whilst marginalising out the effect of ϕ. That can all be done in closed form, and gives a posterior of the form $\hat{\Delta}(\mathbf{x}) \sim GP(\tilde{\Delta}_{\mathcal{D}_e}(\cdot), k_{\mathcal{D}_e}(\cdot,\cdot))$ with expressions for $\tilde{\Delta}_{\mathcal{D}_e}(\cdot), k_{\mathcal{D}_e}(\cdot,\cdot)$ in Appendix 2.C.

2.4.2 Uniform Error Bounds

We now turn to the main purpose of using Gaussian Processes in that we provide a set of assumptions under which our method we can learn functions that control $\Delta(\cdot)$ from experimental data. We do this by adapting the uniform error bounds for GPs from Lederer et al. [133] to our specific model. First, providing the assumption which makes inference possible:

Assumption 2. *The unknown $\hat{\Delta}$ and ϕ are samples from Gaussian processes with kernel k and l respectively, i.e $\hat{\Delta}(\cdot) \sim \mathrm{GP}(0, k)$ and $\phi \sim \mathrm{GP}(0, l)$. Further we assume*

$\mathcal{X}_o$ is compact, the errors have distribution $\mathcal{N}(0, \sigma_t^2)$ given $T = t$, k has Lipschitz constant L_k, and $\hat{\Delta}$ has a Lipschitz constant $L_{\hat{\Delta}}$.

This implies the following bounds on the $\hat{\Delta}(\cdot)$:

Theorem 17. *Let the posterior for $\hat{\Delta}(\cdot)$ from the GP model defined in Section 2.4.1 be given pointwise by $\mathcal{N}(\tilde{\Delta}(\mathbf{x}), \sigma^2(\mathbf{x}))$ where $\tilde{\Delta}, \sigma : \mathcal{X}_o \to \mathbb{R}$ and let $\hat{\tau}(\mathbf{x}) = \hat{\omega}(\mathbf{x}) + \tilde{\Delta}(\mathbf{x})$. Then, under assumption 2, for fixed $\delta \in (0,1), \tau \in \mathbb{R}^+$ we have:*

$$P(|\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})| \leq B(\mathbf{x}) \ \ \forall \mathbf{x} \in \mathcal{X}_o) > 1 - \delta$$

$$B(\mathbf{x}) = \sqrt{2 \log\left(\frac{M(\tau, \mathcal{X}_o)}{\delta}\right)} \sigma(\mathbf{x}) + \gamma(\tau, \mathbf{X}_e, L_k, L_\Delta)$$

Where $M(\tau, \mathcal{X}_o)$ is the τ covering number of $\mathcal{X}_o$, defined as the minimum number of spherical balls of radius τ needed to cover $\mathcal{X}_o$, and $\gamma(\tau, \mathbf{X}_e, L_k, L_\Delta)$ is defined in Appendix 2.C.

There is not a uniformly optimal value of τ due to dependence on other constants. However, if our covariate space is a hypercube of length r with dimension d, we have that $M(\tau, \mathcal{X}_o) \leq (1 + \frac{r}{\tau})^d$ and as we show in Appendix 2.C, we have that $\gamma(\tau, \mathbf{X}_e, L_k, L_\Delta) = o(\tau^{\frac{1}{2}})$. This ensures we can always recover log dependency in δ and τ. Further, the kernel choice gives us knowledge of L_k and allows us to probabilistically bound L_f as in Lederer et al. [133] and shown in Appendix 2.C.

2.5 Experiments

For the experiments, we demonstrate the improvements in predictive performance and calibrated uncertainty provided by our pseudo-outcome GP approach when compared against other causal GP approaches. We compare against fitting a naive standard GP, and a GP with a multitask LCM kernel to pseudo-outcomes. This second approach which can be viewed as a scaled version of the causal multitask Gaussian process of [4],which represents the state of the art in GP's for CATE estimation[3]. For all models we tune the free hyperparameters using gradient descent on the marginal log likelihood, details in Appendix 2.D.1. For results on the coverage of uniform error bounds for our model specifically, see 2.F.

[3]For more on the comparison between the LCM GP and Causal Multitask GP's see Appendix 2.C.3

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.77 \pm 0.01}$	0.785 ± 0.04	$\mathbf{3.31 \pm 0.02}$
Naive GP	2.05 ± 0.02	$\mathbf{0.796 \pm 0.04}$	3.65 ± 0.03
LCM	1.91 ± 0.01	0.303 ± 0.11	1.09 ± 0.06

Table 2.1: Results for the simulated experiment in Section 2.5.1 with $d = 10$ and $n_e = 1000$, averaged over 200 runs. Our approach leads to both the best predictive performance and well calibrated uncertainty, achieving a similar coverage to the standard GP with smaller predictive intervals.

2.5.1 Simulated Experiment

Firstly, we use an adaptation of the simulated provided in Kallus et al. [112]. We let the experimental and observational covariate distribution be $\mathcal{U}([-1, 1]^d)$ and $\mathcal{U}([-3, 3]^d)$ respectively, where d is the covariate dimension, and $T \sim \text{Ber}(\frac{1}{2})$. The observational outcomes are simulated using a quadratic in the first component of X and T with normal noise. For the experimental outcomes, we simulate use the same polynomial but add a sample from a GP. We do this in order to test our methodology in setting where assumptions are satisfied. Full details are given in Appendix 2.D.

Finally, as we focus on assessing the GP portion of the model, we use the true $\omega_o(\mathbf{x})$ for this experiment. This represents a case where n_o is so large, we approach fitting a perfect model. We later relax this.

2.5.1.1 Results

We present results for this experiment with $d = 10$ and $n_e = 1000$ in Table 2.1, alongside an illustrative figure for $d = 1$ and $n_e = 200$ in Figure 2.3. In Table 2.1 we compare the mean squared error to the true CATE, the coverage of 95% Bayesian credible intervals, and the width of these intervals. We use Bayesian credible intervals to form a fair comparison for uncertainty quantification, as uniform error bounds are only available for our model. Additional results including varying dimension, varying sample size, and extrapolation beyond the experimental sample in Appendix 2.E.1.

Across all settings, our results show the following trends: i) The naive GP shows poor predictive performance and is totally uninformative in some settings. This is because it is unable to correctly optimise hyperparameters to capture the highly variable noise distribution in the pseudo-outcomes, and so it reverts to the prior ii) The trainable LCM multitask kernel has good predictive performance but poorly calibrated uncertainty, This is because the hyperparameter optimisation either leads to overly confident or

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.10 \pm 0.04}$	$\mathbf{0.831 \pm 0.008}$	$\mathbf{2.66 \pm 0.02}$
Naive GP	2.19 ± 0.13	0.752 ± 0.010	3.38 ± 0.03
LCM	1.39 ± 0.05	0.828 ± 0.010	3.00 ± 0.05

Table 2.2: Results for the IHDP setting described in Section 2.5.2 with $n_e = 400$ averaged over 100 runs. We again find that our method has the best predictive performance and most informative coverage intervals, in the sense that they contain the true CATE with high probability whilst also being significantly smaller.

overly wide credible intervals, depending on the dimension and the number of samples. Further, this optimisation leads it to overfit training data and extrapolate poorly. iii) Our approach is able to incorporate both strong predictive performance and calibrated uncertainty, with intervals that have the coverage guarantees of other methods that produce much wider intervals.

2.5.2 Semi-Synthetic Experiments

To assess our method in a more realistic setting, we use the Infant Health and Development Program (IHDP) dataset [144], similarly to Hussain et al. [106, 105]. The IHDP dataset comes from a randomised controlled trial, and it contains $n = 985$ samples and a 28 dimensional covariate distribution with 7 continuous covariates. The dataset comes with a treatment allocation, but outcomes need to be simulated.

We form our data from the IHDP dataset as follows: for the observational sample, we uniformly sample the covariates and treatment with replacement until reaching the desired sample size. For the experimental study, we do a weighted sampling to ensure a covariate shift and then randomly sample the treatment from $T \sim \text{Ber}(p)$. For the outcomes in the observational dataset, we simulate from a sparse linear model in X and T. We do this as we want to emulate a scenario where we have relatively low error in estimating $\omega_o(\mathbf{x})$ due to a large n_o, but we do not want to repeat the small dataset multiple times. Finally, we again simulate the difference between observational and experimental distributions with a GP as in Section 2.5.1. Full details available in Appendix 2.D.3.

2.5.2.1 Results

We present the results for the experiment with $n_e = 400$ in Table 2.2, with other results for varying sample size, treatment proportion and out of distribution generalisation

in Appendix 2.E.2. We can see that our proposed methodology outperforms both alternatives in terms of predictive performance and calibrated uncertainty, having the joint best coverage with the smallest intervals. The advantage becomes even more clear when extrapolating beyond the experimental study to the observational study, as shown in Table 2.E.2 in Appendix 2.E.2. In this case, the trained GP predicts overly broad intervals due to over-fitting the hyperparameters to the experimental data.

2.5.3 Additional Results

Finally, we highlight some additional results available in the Appendices. In Appendix 2.E.1, we present our simulated experiment over different dimensionality and sample sizes, in Appendix 2.E.2 we present the IHDP study for different treatment proportions, and for a varying quality of fit of the observational model and in Appendix 2.E.3 we present results for both scenarios, varying the difference function so that it is not a GP. Finally, in Appendix 2.E.3, we provide results on our uniform error bounds for our proposed GP model.

2.6 Conclusion

It is well known in causal inference that there are no observational tests for unmeasured confounding. In this work, we showed that even with experimental data, there are fundamental limits to testing for unmeasured confounding. We showed that in order to validate observational studies, one needs to make assumptions on the smoothness of the correction function. Following this, we developed a Gaussian Process approach to learning from pseudo-outcomes, and assumptions arising from this model which produce intervals that contain the CATE with high probability.

Acknowledgements

JF gratefully acknowledges funding from the EPSRC.

Appendix

2.A Causal Assumptions

2.A.1 Constant CATE

Drawing the graph with an environment node in Figure 2.A.1, we can read off the graph using the SWIG framework [175] for conditional indecencies of potential outcomes from graphical models that $Y(t) \perp E \mid X$. This implies:

$$\mathbb{E}\left[Y(t) \mid X, E = e\right] = \mathbb{E}\left[Y(t) \mid X, E = o\right] \tag{2.21}$$

And so constant CATE.

2.B Hardness of Testing

Theorem 12. *Fix any $\underline{f}, \bar{f} : \mathcal{X} \to \mathbb{R}$ and let ψ_n be an equivalence test with null $\mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})$ and alternative $\mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$. If the level of this test is α we have that:*

$$\mathbb{P}_P(\psi_n = 1) \leq \alpha, \tag{2.22}$$

for any $P \in \mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$. That is ψ_n does not have power against any alternative.

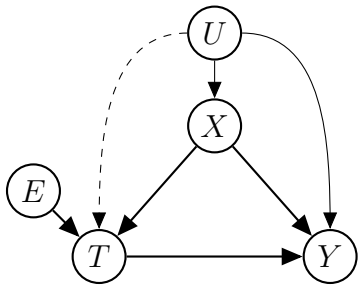


Figure 2.A.1: Causal Structure for generating the experimental and observational datasets with environment node drawn in.

Proof. Following Romano [180], we need to show that the null is dense in the alternative in total variation distance. This corresponds to showing that for $P \in \mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$ we can find a sequence of distributions $\{Q_n\}_{n=1}^{\infty} \in \mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})$ such that $d_{\text{TV}}(P, Q_n) \to 0$.

Firstly, note that as Y is bounded by M we must have that $\left|\hat{\Delta}(\mathbf{x}) + \hat{\omega}_o(\mathbf{x})\right| \leq 2M$ which implies $-(\hat{\omega}_o(\mathbf{x}) + 2M) \leq \hat{\Delta}(\mathbf{x}) \leq 2M - \hat{\omega}_o(\mathbf{x})$. So in order for the bounds to be non vacuous we need $\bar{f}(\mathbf{x}) < 2M - \hat{\Delta}(\mathbf{x}) \leq \hat{\omega}_o(\mathbf{x})$ or $\underline{f}(\mathbf{x}) > -(\hat{\omega}_o(\mathbf{x}) + 2M)$ on some set of positive measure. Assume wlog that we have $\bar{f}(\mathbf{x}) < 2M - \hat{\omega}_o(\mathbf{x})$ and then let $\mathcal{A}_n =\subset \mathcal{X}$ be a measurable set such that $P(\mathcal{A}_n) = \epsilon_n$ where $0 < \epsilon_n \leq \frac{1}{n^2}$ such that this holds. We can find such a set as P is absolutely continuous in $\mathbf{x}$ with respect to the Lebesgue measure.

Now define R_n as a distribution over (X, T, Y) where X is uniform on $\mathcal{A}_n$, conditional distribution of T given X coming from π and the distribution over Y is such that the true CATE is $2M$. Now we let:

$$Q_n \coloneqq \frac{n-1}{n} P + \frac{1}{n} R_n \tag{2.23}$$

Now consider the expectation of $\tilde{Y}$ under this distribution, given $X \in \mathcal{A}_n$:

$$\mathbb{E}_{Q_n}[\tilde{Y} \mid X \in \mathcal{A}_n] = \mathbb{E}_P[\tilde{Y} \mid X \in \mathcal{A}_n] \mathbb{P}\left((\tilde{Y}, X) \sim P \mid X \in \mathcal{A}_n\right) \tag{2.24}$$

$$+ \mathbb{E}_{R_n}[\tilde{Y} \mid X \in \mathcal{A}_n] \mathbb{P}\left((\tilde{Y}, X) \sim R_n \mid X \in \mathcal{A}_n\right) \tag{2.25}$$

We have the following:

$$\mathbb{P}\left((\tilde{Y}, X) \sim P \mid X \in \mathcal{A}_n\right) = \frac{\mathbb{P}\left(X \in \mathcal{A}_n \mid (\tilde{Y}, X) \sim P\right) \mathbb{P}\left((\tilde{Y}, X) \sim P\right)}{\mathbb{P}\left(X \in \mathcal{A}_n\right)} \tag{2.26}$$

$$= \frac{\epsilon_n \frac{n-1}{n}}{\epsilon_n \frac{n-1}{n} + \frac{1}{n}} \tag{2.27}$$

$$= \frac{\epsilon_n (n-1)}{\epsilon_n (n-1) + 1} \tag{2.28}$$

And:

$$\mathbb{P}\left((\tilde{Y}, X) \sim Q_n \mid X \in \mathcal{A}_n\right) = 1 - \mathbb{P}\left((\tilde{Y}, X) \sim P \mid X \in \mathcal{A}_n\right) \tag{2.29}$$

$$= \frac{1}{\epsilon_n (n-1) + 1} \tag{2.30}$$

So that:

$$\mathbb{E}_{Q_n}[\tilde{Y} \mid X \in \mathcal{A}_n] = \frac{\epsilon_n(n-1)\mathbb{E}_P[Z \mid X \in \mathcal{A}_n]}{\epsilon_n(n-1)+1} + \frac{\mathbb{E}_{Q_n}[Z \mid X \in \mathcal{A}_n]}{\epsilon_n(n-1)+1} \tag{2.31}$$

$$= \frac{\epsilon_n(n-1)\mathbb{E}_P[\tilde{Y} \mid X \in \mathcal{A}_n]}{\epsilon_n(n-1)+1} + \frac{2M}{\epsilon_n(n-1)+1} \tag{2.32}$$

Now as $\epsilon_n(n-1) \to 0$ as $n \to \infty$ we have $\mathbb{E}_{P_n}[\tilde{Y} \mid X \in \mathcal{A}_n] \to 2M$. As we have $\hat{\Delta}(\mathbf{x}) = \mathbb{E}_P[\tilde{Y} \mid \mathbf{x}] - \hat{\omega}_o(\mathbf{x})$, we have that $\hat{\Delta}(\mathbf{x}) \to 2M - \hat{\omega}_o(\mathbf{x})$ for $\mathbf{x} \in \mathcal{A}_n$. Therefore, by taking the cutoff sequence $\{Q_n\}_{n \geq k}^{\infty}$ for some k we have a sequence of distributions such that $\hat{\Delta}(\mathbf{x})$ expectation is greater than $\underline{f}(\mathbf{x})$ on a set of positive measure and that $d_{\text{TV}}(Q, P_n) \to 0$. □

Corollary 13. *For fixed $\underline{f}, \bar{f} : \mathcal{X} \to \mathbb{R}$, any falsification test ψ_n with null $\mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$ and alternative $\mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})$ has:*

$$\inf_{Q \in \mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})} \mathbb{P}_Q(\psi_n = 1) \leq \alpha, \tag{2.33}$$

where α is the level of ψ_n.

Proof. This follows as a direct result of the fact that the alternative is dense in the null, as shown in the previous proposition □

Proposition 14. *There exists a distribution $Q \in \mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})$ such that $\text{TV}(Q, \text{co}(\mathcal{P}_{M,\pi}(\underline{f}, \bar{f}))) \geq \beta$ for some $\beta > 0$ where $\text{co}(\mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$ is the convex hull of $\mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$. Following Bertanha and Moreira [24], this guarantees that there is a test with $\beta + \alpha$ against Q where α is the level of the test.*

Proof. As in the proof of Theorem 12 assume wlog that we have $\bar{f}(\mathbf{x}) < 2M - \hat{\omega}_o(\mathbf{x})$ for some measurable set $\mathcal{A}$. Now define a distribution $Q \in \mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})$ as follows:

- X is uniform on $\mathcal{A}$.
- $T|X \sim \pi(X)$.
- $Y|T = M(2T-1)$.

Now let $\tilde{A} = \text{supp}(Q)$ be the support of Q. We have that $\mathbb{E}_P[\tilde{Y}|(X,T,Y) \in \tilde{\mathcal{A}}] = 2M$ for any distribution P. Now for $P \in \mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})$ we have that $\mathbb{E}_P[\tilde{Y}|(X) \in \mathcal{A}] \leq \bar{f}(\mathbf{x}) <$

$2M - \hat{\omega}_o(\mathbf{x})$. Now we have:

$$\mathbb{E}_P[\tilde{Y}|X \in \mathcal{A}] = \mathbb{E}_P[\tilde{Y}|(X,T,Y) \in \tilde{\mathcal{A}}]\mathbb{P}_P((X,T,Y) \in \tilde{\mathcal{A}} \mid X \in \mathcal{A}) \tag{2.34}$$

$$+ \mathbb{E}_P[\tilde{Y}|(X,T,Y) \in \mathcal{A} \setminus \tilde{\mathcal{A}}] \left(1 - \mathbb{P}_P((X,T,Y) \in \tilde{\mathcal{A}} \mid X \in \mathcal{A})\right) \tag{2.35}$$

$$\geq 2M \left(2\mathbb{P}_P((X,T,Y) \in \tilde{\mathcal{A}} \mid X \in \mathcal{A}) - 1\right) \tag{2.36}$$

As $\mathbb{E}_P[\tilde{Y}|(X,T,Y) \in \mathcal{A} \setminus \tilde{\mathcal{A}}] \geq -2M$. Putting this together implies:

$$2M - \hat{\omega}_o(\mathbf{x}) \geq 2M \left(2\mathbb{P}_P((X,T,Y) \in \tilde{\mathcal{A}} \mid X \in \mathcal{A}) - 1\right) \tag{2.37}$$

$$\implies \mathbb{P}_P((X,T,Y) \in \tilde{\mathcal{A}} \mid X \in \mathcal{A}) \leq 1 - \frac{\hat{w}_o(\mathbf{x})}{4M} \tag{2.38}$$

$$\implies \mathbb{P}_P((X,T,Y) \in \tilde{\mathcal{A}}) \leq 1 - \frac{\hat{w}_o(\mathbf{x})}{4M} \tag{2.39}$$

Which completes the proof. □

Theorem 15. *Fix a sensitivity model and let $\mathcal{D}$ be a dataset sampled from $P^{(n)}$ where $P \in \mathcal{E}_{M,\pi}$. Let $\left[\underline{C}\left(\mathcal{D}\right), \bar{C}\left(\mathcal{D}\right)\right]$ be a confidence interval for $\Gamma(P)$ in that it satisfies the following coverage requirement:*

$$\inf_{P \in \mathcal{E}_{0,M}} \mathbb{P}_{\mathcal{D} \sim P^{(n)}} \left(\Gamma(P) \in C\left(\mathcal{D}_n\right)\right) \geq 1 - \alpha \tag{2.40}$$

Then $\bar{C}\left(\mathcal{D}\right) = \Gamma_1$ with probability $1 - \alpha$. That is, there are no non trivial upper bounds on $\Gamma(P)$.

Proof. This follows from the fact that for any $\gamma \in [\Gamma_0, \Gamma_1]$ we have that $\mathcal{Q}_{M,\pi}(\underline{f}_\gamma, \bar{f}_\gamma)$ is dense in $\mathcal{P}_{M,\pi}(\underline{f}_\gamma, \bar{f}_\gamma)$. Therefore, for any distribution $P \in \mathcal{E}_{0,M}$ is arbitrarily close in total variation to a distribution whose true sensativity parameter is arbitrarily high. Therefore, if we are to satisfy the coverage requirment uniformly over all distributions we must have $\bar{C}\left(\mathcal{D}\right) = \Gamma_1$ with probability $1 - \alpha$. □

2.C Gaussian Process

Proposition 16. *There exists a function $\phi : \mathcal{X} \to \mathbb{R}$ such that the IPW pseudo-outcome can be written as:*

$$\tilde{Y} = \tau(X) + \left(\frac{T - \pi(X)}{\pi(X)\left(1 - \pi(X)\right)}\right)\phi(X) + \tilde{\epsilon}$$

Where $\mathbb{E}[\tilde{\epsilon} \mid X, T] = 0$ and $\tilde{\epsilon} \mid T, X$ is Gaussian if the original errors on Y are.

Proof. Suppose we have:

$$Y = \mu(X,T) + \epsilon \tag{2.41}$$

Where $\mu(X,T) = \mathbb{E}[Y \mid X,T]$ so that $\mathbb{E}[\epsilon \mid X,T] = 0$. Now the error of the pseudo-outcome from $\tau(X)$ is:

$$\frac{T - e(X)}{e(X)\left(1 - e(X)\right)} Y - \left(\mu(X,1) - \mu(X,0)\right) = \frac{\left(T - e(X)\right)\left(\mu(X,T) + \epsilon\right)}{e(X)\left(1 - e(X)\right)} - \left(\mu(X,1) - \mu(X,0)\right) \tag{2.42}$$

$$= \begin{cases} \frac{(\mu(X,1)+\epsilon)}{e(X)} - \left(\mu(X,1) - \mu(X,0)\right) & \text{if } T = 1 \\ -\frac{(\mu(X,0)+\epsilon)}{1-e(X)} - \left(\mu(X,1) - \mu(X,0)\right) & \text{if } T = 0 \end{cases} \tag{2.43}$$

$$= \begin{cases} \frac{1}{e(X)}\left(\left(1 - e(X)\right)\mu(X,1) + e(X)\mu(X,0) + \epsilon\right) & \text{if } T = 1 \\ \frac{-1}{1-e(X)}\left(e(X)\mu(X,0) + \left(1 - e(X)\right)\mu(X,1) - \epsilon\right) & \text{if } T = 0 \end{cases} \tag{2.44}$$

If we let $\phi(X) = \left(1 - e(X)\right)\mu(X,1) + e(X)\mu(X,0)$ then we have that (2.44) is equal to:

$$= \frac{T - e(X)}{e(X)\left(1 - e(X)\right)} \left(\phi(X) + (-1)^{T+1}\,\epsilon\right) \tag{2.45}$$

Now if we let $\tilde{\epsilon} = (-1)^{T+1}\,\epsilon \frac{T - e(X)}{e(X)(1-e(X))}$ we can see that we now have $\mathbb{E}\left[\tilde{\epsilon} \mid X,T\right] = 0$. Moreover, since the distribution of $\tilde{\epsilon}$ given X,T is just a constant scaled version of ϵ we have that $\tilde{\epsilon} \mid X,T$ is Gaussian if and only if $\epsilon \mid X,T$ is. □

2.C.1 Closed Form Posterior Expressions

Now, for the closed form expressions have as follows, were the training dataset is $\{\mathbf{x}_i, t_i, \tilde{y}_i - \hat{\omega}_o(\mathbf{x}_i)\}_{i=1}^{n_e}$:

$$\mathbf{M}_N = \left(\left(\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)\right)_{t_i,t_j}\right)_{i,j} \tag{2.46}$$

$$\mathbf{\Sigma}_N = \operatorname{diag}\left(\left(\sigma^2_{t_i}\right)_i\right) \tag{2.47}$$

$$\mathbf{y}_N = \left(\tilde{y}_i - \hat{\omega}_o(\mathbf{x}_i)\right)_i \tag{2.48}$$

$$\boldsymbol{k}_N(\mathbf{x}) = \left(k(\mathbf{x}_i, \mathbf{x})\right)_i \tag{2.49}$$

$$\tilde{\Delta}_{\mathcal{D}_e}(\mathbf{x}) = \boldsymbol{k}_N(\mathbf{x})^\top (\mathbf{M}_N + \mathbf{\Sigma}_N)^{-1} \mathbf{y}_N \tag{2.50}$$

$$k_{\mathcal{D}_e}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \boldsymbol{k}_N(\mathbf{x})^\top (\mathbf{M}_N + \mathbf{\Sigma}_N)^{-1} \boldsymbol{k}_N(\mathbf{x}') \tag{2.51}$$

2.C.2 Closed Form Bounds

Theorem 17. *Let the posterior for $\Delta(\cdot)$ from the GP model defined in Section 2.4.1 be given pointwise by $\mathcal{N}(\tilde{\Delta}(\mathbf{x}), \sigma^2(\mathbf{x}))$ where $\tilde{\Delta}, \sigma : \mathcal{X}_o \to \mathbb{R}$. Then, under assumption 2, for fixed $\delta \in (0,1), \tau \in \mathbb{R}^+$ we have:*

$$P(|\Delta(\mathbf{x}) - \tilde{\Delta}(\mathbf{x})| \leq B(\mathbf{x}) \ \ \forall \mathbf{x} \in \mathcal{X}_o) > 1 - \delta$$

$$B(\mathbf{x}) = \sqrt{2 \log\left(\frac{M(\tau, \mathcal{X}_o)}{\delta}\right)} \sigma(\mathbf{x}) + \gamma(\tau, \mathbf{X}_e, L_k, L_\Delta)$$

Where $M(\tau, \mathcal{X}_o)$ is the τ covering number of $\mathcal{X}_o$, defined as the minimum number of spherical balls of radius τ needed to cover $\mathcal{X}_o$, and $\gamma(\tau, \mathbf{X}_e, L_k, L_\Delta)$ is defined in Appendix 2.C.

Proof. This follows from theorem 3.1 in Lederer et al. [133] which we reproduce here:

Theorem (Lederer et al. [133])**.** *Consider a zero mean Gaussian process defined through the continuous covariance kernel $k(\cdot,\cdot)$ with Lipschitz constant L_k on the compact set $\mathbb{X}$. Furthermore, consider a continuous unknown function $f : \mathbb{X} \to \mathbb{R}$ with Lipschitz constant L_f and $N \in \mathbb{N}$ observations y_i satisfying:*

Assumption 3. *The unknown function $f(\cdot)$ is a sample from a Gaussian process $\mathcal{GP}(0, k(\boldsymbol{x}, \boldsymbol{x}'))$ and observations $y = f(\boldsymbol{x}) + \epsilon$ are perturbed by zero mean i.i.d. Gaussian noise ϵ with variance σ_n^2.*

Then, the posterior mean function $\nu_N(\cdot)$ and standard deviation $\sigma_N(\cdot)$ of a Gaussian process conditioned on the training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ are continuous with Lipschitz constant L_{ν_N} and modulus of continuity $\omega_{\sigma_N}(\cdot)$ on $\mathbb{X}$ such that:

$$L_{\nu_N} \leq L_k \sqrt{N} \left\| \left(k(\boldsymbol{X}_N, \boldsymbol{X}_N) + \sigma_n^2 \boldsymbol{I}_N \right)^{-1} \boldsymbol{y}_N \right\|$$

$$\omega_{\sigma_N}(\tau) \leq \sqrt{2\tau L_k \left(1 + N \left\| \left(k(\boldsymbol{X}_N, \boldsymbol{X}_N) + \sigma_n^2 \boldsymbol{I}_N \right)^{-1} \right\| \max_{\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}} k(\boldsymbol{x}, \boldsymbol{x}') \right)}$$

Moreover, pick $\delta \in (0,1), \tau \in \mathbb{R}_+$ and set

$$\beta(\tau) = 2\log\left(\frac{M(\tau, \mathbb{X})}{\delta}\right)$$
$$\gamma(\tau) = (L_{\nu_N} + L_f)\,\tau + \sqrt{\beta(\tau)}\omega_{\sigma_N}(\tau)$$

Then, it holds that

$$P\left(|f(\boldsymbol{x}) - \nu_N(\boldsymbol{x})| \leq \sqrt{\beta(\tau)}\sigma_N(\boldsymbol{x}) + \gamma(\tau), \forall \boldsymbol{x} \in \mathbb{X}\right) \geq 1 - \delta$$

In our case, the bounds on the Lipschitz constants and modulus of continuity are different. However by the same argument as Lederer et al. [133] we have:

$$L_{\tilde{\Delta}_{\mathcal{D}_e}} \leq L_k\sqrt{N}\|(\mathbf{M}_N + \mathbf{\Sigma}_N)^{-1}\mathbf{y}_N\| \tag{2.52}$$
$$\omega_{\sigma_N}(\tau) \leq \sqrt{2\tau L_k\left(1 + N\,\|(\mathbf{M}_N + \mathbf{\Sigma}_N)^{-1}\|\max_{\boldsymbol{x},\boldsymbol{x}'\in\mathbb{X}} k\,(\boldsymbol{x}, \boldsymbol{x}')\right)} \tag{2.53}$$

Putting this in to the above Theorem of Lederer et al. [133] that, under assumption 2, for fixed $\delta \in (0, 1), \tau \in \mathbb{R}^+$ we have

$$P(|\Delta(\mathbf{x}) - \tilde{\Delta}(\mathbf{x})| \leq B(\mathbf{x})\ \ \forall \mathbf{x} \in \mathcal{X}_o) > 1 - \delta$$
$$\begin{aligned} B(\mathbf{x}) = &\sqrt{2\log\left(\frac{M\,(\tau, \mathcal{X}_o)}{\delta}\right)}\sigma(\mathbf{x}) \\ &+ \left(L_k\sqrt{N}\|(\mathbf{M}_N + \mathbf{\Sigma}_N)^{-1}\mathbf{y}_N\| + L_f\right)\tau \\ &+ \sqrt{\beta(\tau)}\sqrt{2\tau L_k\left(1 + N\,\|(\mathbf{M}_N + \mathbf{\Sigma}_N)^{-1}\|\max_{\boldsymbol{x},\boldsymbol{x}'\in\mathbb{X}} k\,(\boldsymbol{x}, \boldsymbol{x}')\right)} \end{aligned}$$

Where we can see the claimed convergence property in τ. □

2.C.3 LCM Kernel and Causal Multitask Kernel of Alaa and Van Der Schaar [4]

In this the work of Alaa and Van Der Schaar [4], CATE is modelled using a multitask Gaussian process [28]. Multitask Gaussian Processes use a GP in vector-valued Reproducing Kernel Hilbert Space (vv-RKHS) to share information between tasks [7]. In Alaa and Van Der Schaar [4], learning the conditional outcome function for each treatment is seen as a separate task, so we jointly model:

$$Y|\mathbf{x}, t \sim \mathcal{N}(0, f_t(\mathbf{x}), \sigma_t^2) \tag{2.54}$$

Where each f_t is a Gaussian Process. The kernel $\tilde{\mathbf{K}}_\eta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{2\times 2}$ is now a symmetric positive semi-definite matrix-valued function, with hyper-parameters η. In the case of Alaa and Van Der Schaar [4] they use a *linear model of coregionalization*[4], giving the kernel as:

$$\tilde{\mathbf{K}}(\mathbf{x}, \mathbf{x}') = \mathbf{A}_0 k(\mathbf{x}, \mathbf{x}') + \mathbf{A}_1 l(\mathbf{x}, \mathbf{x}') \tag{2.55}$$

Where $\mathbf{A}_t$ is given by:

$$\tilde{\mathbf{A}}_0 = \begin{bmatrix} \theta_{00}^2 & \rho_0 \\ \rho_0 & \theta_{01}^2 \end{bmatrix}, \tilde{\mathbf{A}}_1 = \begin{bmatrix} \theta_{10}^2 & \rho_1 \\ \rho_1 & \theta_{11}^2 \end{bmatrix}. \tag{2.56}$$

And $\mathbf{A}_t$ are now free hyperparameters to learn.

This is equivalent to the LCM kernel that we make use of for our experiments, however we regress onto pseudo-outcomes as opposed to observed outcomes. This is equivalent to scaling the task t is scaled by $\frac{t-\pi(\mathbf{x}))}{\pi(\mathbf{x})(1-\pi(\mathbf{x}))}$ and leaving all hyper-parameters free to learn. As scaled Gaussian processes are still Gaussian processes this is same as using the kernel:

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \mathbf{M}\mathbf{A}_0 k(\mathbf{x}, \mathbf{x}') + \mathbf{M}\mathbf{A}_1 l(\mathbf{x}, \mathbf{x}') \tag{2.57}$$

Where:

$$\mathbf{M} = \left(\left(\frac{t - \pi(\mathbf{x}))}{\pi(\mathbf{x})(1-\pi(\mathbf{x}))} \right) \left(\frac{t' - \pi(\mathbf{x}))}{\pi(\mathbf{x})(1-\pi(\mathbf{x}))} \right) \right)_{t,t'=0,1} \tag{2.58}$$

This demonstrates an equivalence between the Pseudo-outcome based LCM method we use for the experiments and the methods of [4].

2.D Experiment Details

2.D.1 Model Tuning Details

For each of the models we regress from $\mathbf{x}, \mathbf{t}$ onto $\tilde{y} - \hat{w}_o(\mathbf{x})$ where $\hat{w}_o(\mathbf{x})$ is our estimate of $w_o(\mathbf{x})$. We do so as:

$$\tilde{Y} - \hat{w}_o(X) \sim \mathcal{N}(f_t(X), \sigma_t^2) \tag{2.59}$$

With specific model details as follows:

[4] See Alvarez et al. [7] for more details.

1. **Standard or Naive GP** Taking $f_0 = f_1 = f$ and $\sigma_0^2 = \sigma_1^2 = \sigma_2^2$. We then model f directly using a $\text{GP}(0, k_\theta)$ where k_θ is a kernel with hyper parameters θ. Described throughout as standard or naive GP. The hyperparmeters are given by θ, σ.

2. **LCM GP** Modelling f_t using a multitask GP. This corresponds to using the vector valued kernel:

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \mathbf{A}_0 k_\theta(\mathbf{x}, \mathbf{x}') + \mathbf{A}_1 l_\eta(\mathbf{x}, \mathbf{x}') \tag{2.60}$$

   Where $\mathbf{A}_t$ is given by:

$$\mathbf{A}_0 = \begin{bmatrix} \theta_{00}^2 & \rho_0 \\ \rho_0 & \theta_{01}^2 \end{bmatrix}, \mathbf{A}_1 = \begin{bmatrix} \theta_{10}^2 & \rho_1 \\ \rho_1 & \theta_{11}^2 \end{bmatrix}. \tag{2.61}$$

   CATE differences are then formed as the weighted average between both treatments. So modelled as:

$$\Delta(\mathbf{x}) = \sum_{t=0}^{1} f_t(\mathbf{x}) \tag{2.62}$$

   Where f_t is the prediction for task t. For this method the hyper-parameters to learn are $\theta, \eta, \mathbf{A}_0, \mathbf{A}_1, \sigma_0, \sigma_1$.

3. **Our Approach**. We use a multitask Gaussian process given by:

$$\mathbf{K} = \mathbf{a}\mathbf{a}^\top k_\theta + \mathbf{b}\mathbf{b}^\top l_\eta \tag{2.63}$$

$$\mathbf{a} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \tag{2.64}$$

$$\mathbf{b} = \begin{bmatrix} \frac{-1}{1-\pi(X)} & \frac{1}{\pi(X)} & 0 \end{bmatrix} \tag{2.65}$$

   Where the first task models f_0, the second f_1, and the final is the CATE gap, $\Delta(\mathbf{x})$. Using the decomposition:

$$\tilde{Y} = \tau(X) + \left(\frac{T - \pi(X)}{\pi(X)\left(1 - \pi(X)\right)} \right) \phi(X) + \tilde{\epsilon}$$

   This is equivalent to modelling $\tau(X) \sim \text{GP}(0, k_\theta)$ and $\phi(X) \sim \text{GP}(0, l_\eta)$. The hyper-parameters for this method are $\eta, \theta, \sigma_1, \sigma_0$.

2.D.2 Simulation Details

For the first experiment, we simulate data as follows:

$$X|E=o \sim \mathcal{U}([-3,3]^d), \;\; X|E=e \sim \mathcal{U}([-1,1]^d) \tag{2.66}$$

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{2}\sum_{j=1}^{1} \beta_{i,j}^{\top}(t^j \odot \mathbf{x}^i) + \epsilon \tag{2.67}$$

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{2}\sum_{j=1}^{1} \beta_{i,j}^{\top}(t^j \odot \mathbf{x}^i) + \mathbf{f}_t(\mathbf{x}) + \epsilon \tag{2.68}$$

$$\epsilon \sim \mathcal{N}(0,\sigma_0), \mathbf{f}_t \sim GP(0, k_{\theta_0}) \tag{2.69}$$

Where $\beta_{i,j} = 1$ for each i, j and $\sigma_0 = 0.5$

2.D.3 IHDP details

We simulate covariates following a similar approach to Hussain et al. [106]. For the observational dataset we uniformly sample from the IHDP covariate distribution. For the experimental covariate distribution we sample with weights:

$$w_i = 0.8^{\mathbb{1}\{\text{mother is smoker}\}+\mathbb{1}\{\text{is male}\}} \tag{2.70}$$

So that the experiment dataset is significantly more likely to include male babies whose mothers are smokers. For the experimental dataset treatment is simulated as $\text{Ber}(p)$. The outcome is then simulated as:

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{1}\sum_{j=1}^{1} \beta_{i,j}^{\top}(t^j \odot \mathbf{x}^i) + \epsilon \tag{2.71}$$

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{1}\sum_{j=1}^{1} \beta_{i,j}^{\top}(t^j \odot \mathbf{x}^i) + \mathbf{f}_t(\mathbf{x}) + \epsilon \tag{2.72}$$

$$\epsilon \sim \mathcal{N}(0,\sigma_0), \mathbf{f}_t \sim GP(0, k_{\theta_0}) \tag{2.73}$$

Where, $\beta_{i,j} = Z_i N_j$ where $Z_i \sim \text{Ber}(0.3)$ and $N_j \sim \mathcal{N}(0,1)$ and $\sigma_0 = 0.5$.

2.E Additional Results

2.E.1 Simulated Experiment Additional Results

	MSE			Coverage			Interval Width		
Dim X	Ours	Standard	LCM	Ours	Standard	LCM	Ours	Standard	LCM
5	0.301 ± 0.006	1.21 ± 0.06	0.352 ± 0.007	0.935 ± 0.003	0.858 ± 0.008	0.932 ± 0.003	1.97 ± 0.02	3.22 ± 0.02	2.15 ± 0.02
10	$1.77 + \pm 0.0157$	2.05 ± 0.02	1.91 ± 0.02	$0.785 + \pm 0.005$	0.796 ± 0.007	0.303 ± 0.021	3.31 ± 0.03	3.65 ± 0.04	1.09 ± 0.08
25	2.15 ± 0.02	2.27 ± 0.03	2.02 ± 0.01	0.779 ± 0.004	0.754 ± 0.008	0.128 ± 0.012	3.60 ± 0.02	3.50 ± 0.04	0.463 ± 0.045

Table 2.E.1: In distribution results for $n_{\text{exp}} = 1000$ across dimension, average across 200 runs with 95% confidence interval.

	MSE			Coverage			Interval Width		
Dim X	Ours	Standard	LCM	Ours	Standard	LCM	Ours	Standard	LCM
5	2.10 ± 0.05	2.17 ± 0.06	2.31 ± 0.07	0.825 ± 0.010	0.817 ± 0.010	0.997 ± 0.01	3.91 ± 0.00	3.92 ± 0.00	9.65 ± 0.20
10	2.01 ± 0.02	2.05 ± 0.02	2.38 ± 0.02	0.832 ± 0.002	0.829 ± 0.003	0.720 ± 0.041	3.91 ± 0.02	3.91 ± 0.02	3.85 ± 0.15
25	2.01 ± 0.01	2.03 ± 0.01	2.04 ± 0.00	0.832 ± 0.001	0.831 ± 0.002	0.239 ± 0.030	3.92 ± 0.00	3.92 ± 0.00	0.910 ± 0.14

Table 2.E.2: Out of distribution results for $n_{\text{exp}} = 1000$ across dimension, average across 200 runs with 95% confidence interval.

	MSE			Coverage			Interval Width		
Dim X	Ours	Standard	LCM	Ours	Standard	LCM	Ours	Standard	LCM
5	0.155 ± 0.002	0.808 ± 0.06	0.290 ± 0.010	0.950 ± 0.002	0.897 ± 0.006	0.892 ± 0.005	1.48 ± 0.01	2.86 ± 0.08	1.76 ± 0.03
10	$1.49 + \pm 0.01$	1.94 ± 0.02	1.69 ± 0.02	$0.807 + \pm 0.003$	0.812 ± 0.006	0.481 ± 0.019	3.17 ± 0.02	3.67 ± 0.04	1.70 ± 0.07
25	2.08 ± 0.01	2.19 ± 0.02	2.01 ± 0.02	0.804 ± 0.002	0.775 ± 0.005	0.104 ± 0.008	3.742 ± 0.01	3.587 ± 0.03	0.373 ± 0.03

Table 2.E.3: In distribution results for $n_{\text{exp}} = 2500$ across dimension, average across 200 runs with 95% confidence interval.

	MSE			Coverage			Interval Width		
Dim X	Ours	Standard	LCM	Ours	Standard	LCM	Ours	Standard	LCM
5	2.13 ± 0.11	2.21 ± 0.12	2.76 ± 0.27	0.819 ± 0.011	0.810 ± 0.013	0.999 ± 0.001	3.92 ± 0.00	3.92 ± 0.00	11.42 ± 0.32
10	2.01 ± 0.02	2.03 ± 0.02	2.33 ± 0.02	0.833 ± 0.002	0.831 ± 0.003	0.931 ± 0.011	3.92 ± 0.00	3.92 ± 0.00	5.80 ± 0.18
25	1.99 ± 0.01	2.01 ± 0.02	2.01 ± 0.00	0.834 ± 0.001	0.832 ± 0.002	0.216 ± 0.011	3.92 ± 0.00	3.92 ± 0.00	0.794 ± 0.18

Table 2.E.4: Out of distribution results for $n_{\text{exp}} = 2500$ across dimension, average across 200 runs with 95% confidence interval.

2.E.2 IHDP Experiment Additional Results

Model	**MSE**	**Coverage**	**Interval Width**
Ours	$\mathbf{1.19 \pm 0.04}$	$\mathbf{0.795 \pm 0.008}$	$\mathbf{2.572 \pm 0.02}$
Naive GP	2.43 ± 0.13	0.752 ± 0.010	3.42 ± 0.03
LCM	1.57 ± 0.05	0.752 ± 0.010	3.21 ± 0.05

Table 2.E.5: In of distribution results for the IHDP setting described in Section 2.5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.93 \pm 0.04}$	$\mathbf{0.812 \pm 0.008}$	$\mathbf{3.65 \pm 0.02}$
Naive GP	2.27 ± 0.13	0.797 ± 0.010	3.81 ± 0.03
LCM	2.50 ± 0.05	0.961 ± 0.010	8.03 ± 0.05

Table 2.E.6: Out of distribution results for the IHDP setting described in Section 2.5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.63 \pm 0.04}$	$\mathbf{0.643 \pm 0.008}$	$\mathbf{2.27 \pm 0.02}$
Naive GP	2.59 ± 0.13	0.732 ± 0.010	3.48 ± 0.03
LCM	2.04 ± 0.05	0.931 ± 0.010	6.33 ± 0.05

Table 2.E.7: Out of distribution results for the IHDP setting described in Section 2.5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{2.07 \pm 0.04}$	$\mathbf{0.778 \pm 0.008}$	$\mathbf{3.58 \pm 0.02}$
Naive GP	2.37 ± 0.13	0.789 ± 0.010	3.82 ± 0.03
LCM	2.58 ± 0.05	0.931 ± 0.010	6.33 ± 0.05

Table 2.E.8: Out of distribution results for the IHDP setting described in Section 2.5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

2.E.3 Robustness Results

We now repeat the experiment for the IHDP dataset but we add squared terms to the simulation as follows:

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{1}\sum_{j=1}^{1} \beta_{i,j}^{\top}(t^j \odot \mathbf{x}^j) + \gamma_{i,j}^{\top}(t^j \odot \mathbf{x}^j)^2 + \epsilon \tag{2.74}$$

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{1}\sum_{j=1}^{1} \beta_{i,j}^{\top}(t^j \odot \mathbf{x}^j) + \gamma_{i,j}^{\top}(t^j \odot \mathbf{x}^j)^2 + \mathbf{f}_t(\mathbf{x}) + \epsilon \tag{2.75}$$

$$\epsilon \sim \mathcal{N}(0, \sigma_0), \mathbf{f}_t \sim GP(0, k_{\theta_0}) \tag{2.76}$$

We still fit a linear model for $\omega_o(\mathbf{x})$ which ensures that $\Delta(\mathbf{x})$ is not a GP.

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.10 \pm 0.03}$	$\mathbf{0.824 \pm 0.03}$	$\mathbf{2.63 \pm 0.02}$
Naive GP	2.13 ± 0.10	0.761 ± 0.01	3.39 ± 0.04
LCM	1.34 ± 0.04	0.832 ± 0.01	2.96 ± 0.04

Table 2.E.9: In of distribution results for the IHDP setting described in Section 2.5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.91 \pm 0.03}$	$\mathbf{0.821 \pm 0.003}$	$\mathbf{3.66 \pm 0.05}$
Naive GP	2.15 ± 0.04	0.805 ± 0.004	3.80 ± 0.014
LCM	2.22 ± 0.09	0.832 ± 0.01	7.76 ± 0.27

Table 2.E.10: In of distribution results for the IHDP setting described in Section 2.5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

2.F Uniform Error Bounds

Finally, we repeat the experiment in Section 2.E.3 but with the uniform error bounds.

n_e	500	1000	5000	10000
Whole Function Coverage	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Interval width in Distribution	21.3 ± 0.12	14.0 ± 0.01	5.36 ± 0.01	3.76 ± 0.03
Interval width out of distribution	31.7 ± 0.04	30.4 ± 0.04	28.9 ± 0.05	28.6 ± 0.1

Table 2.F.1: Uniform error bounds averaged over 100 runs. We vary the sample size to show that the in distribution bounds decrease in width as n_e increases.

3

Is Merging Worth It? Securely Evaluating the Information Gain for Causal Dataset Acquisition

Abstract

Merging datasets across institutions is a lengthy and costly procedure, especially when it involves private information. Data hosts may therefore want to prospectively gauge which datasets are most beneficial to merge with, without revealing sensitive information. For causal estimation this is particularly challenging as the value of a merge depends not only on reduction in epistemic uncertainty but also on improvement in overlap. To address this challenge, we introduce the first *cryptographically secure* information-theoretic approach for quantifying the value of a merge in the context of heterogeneous treatment effect estimation. We do this by evaluating the *Expected Information Gain* (EIG) using multi-party computation to ensure that no raw data is revealed. We further demonstrate that our approach can be combined with differential privacy (DP) to meet arbitrary privacy requirements whilst preserving more accurate computation compared to DP alone. To the best of our knowledge, this work presents the first privacy-preserving method for dataset acquisition tailored to causal estimation. Code is publicly available: https://github.com/LucileTerminassian/causal_prospective_merge.

3.1 Introduction

As the demand for data-driven decision making grows, the question of how to optimally collect data for a given task becomes increasingly important. Data fusion [36], which integrates pre-existing data from various sources, is a popular method to increase sample size, reduce sampling variability, and enhance statistical power and robustness [134, 60]. However, merging datasets is often a time-consuming and resource-intensive task. This is especially true in sensitive domains such as healthcare, where concerns surrounding privacy, downstream applications, and data security mean long ethical approval procedures are required before undergoing a merge [165, 150, 68]. Consequently, practitioners crucially need methods to determine the value of a potential merge in advance, whilst also complying with privacy requirements [1].

In this work, we focus on data fusion in the context of heterogeneous treatment effect estimation. As a concrete example of this problem, consider a hospital that wishes to assess the impact of a medical intervention on its patients. Upon finding that its own data is insufficient to get accurate estimates, the hospital plans to select one of K possible candidate hospitals for a potential data merge.

Given the costs involved, the hospital would like to identify *in advance* which potential dataset would provide the most information upon merging, *whilst* complying with privacy regulations. We propose a solution for such types of problems, allowing for scenarios where patient outcomes at the candidate hospitals to be unobserved or simply masked.

We quantify the value of a merge in a principled, information theoretic way by applying techniques from Bayesian experimental design [169, 139, 37]. Our solution shares similarities with standard Bayesian dataset acquisition [145, 119], however in causal contexts data serve an additional, distinct purpose. Specifically acquired data should not only enhance our understanding of the outcome function, but also assist in combating the selection bias that is inherent to causal effect estimation [100] by balancing treatment. Put differently, whilst generic data fusion aims at reducing the epistemic uncertainty from incomplete knowledge of the outcome, causal estimation also seeks to improve treatment overlap [184, 52], i.e. reducing the epistemic uncertainty for counterfactual outcomes. To resolve this, we make use of favourable parameterisations available in popular Bayesian causal inference methods [91, 4], which allows us to prioritise

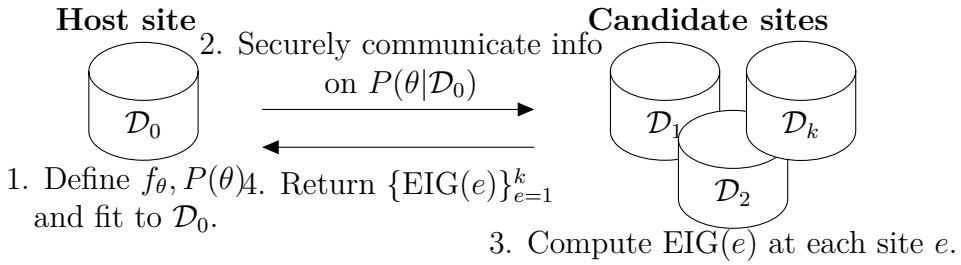


Figure 3.1.1: Flow chart depicting our method. In step 1 the *host site* chooses a parameterised model class for the conditional outcome, given by f_θ and sets the prior $P(\theta)$. They then perform a local Bayesian update, and communicate information on the posterior $P(\theta|\mathcal{D}_0)$ to the *candidate site* using secure multi party computation (MPC). The candidate applies MPC once more to privately calculate the Expected Information Gain (EIG) and communicate it back to host. This allows the host to select the best merge out of the potential candidates.

information gain in the parameters relevant for the causal problem, rather than gaining information in irrelevant parts of the conditional outcome.

To ensure privacy, we employ Secure Multi-Party Computation [226, 69, 121]. This cryptographic protocol enables multiple parties to jointly compute the output of a function without revealing any of their own private inputs. A classic example involves determining the wealthiest person in a group without anyone revealing their personal net worth. In our context it allows the different candidate sites to compute their expected information gains relative to the initial sites' data, without exposing the contents of their datasets. This ensures that the noise required for privacy guarantees can be added to the final statistic as opposed to the raw data at each site.

Our contributions can be summarised as follows:

- We propose information-theoretic methods to measure the value of a data merge in the context of heterogeneous treatment effects estimation. Our primary contribution is a novel approach that specifically targets the reduction of entropy in parameters that directly influence the conditional average treatment effect (CATE). We also present a standard approach based on expected entropy reduction in all parameters of a conditional outcome model.
- We demonstrate how both of the approaches can be used with three popular CATE estimators; Bayesian Polynomial Regression [86], Causal Multitask Gaussian Processes [4], and Bayesian Causal Forests [91]. We derive closed form expressions

for the expected entropy reduction in the first two models and give a Monte Carlo estimator for the other.

- We provide a privacy protocol for our methods based on multi-party computation [MPC; 226]. This ensures that statistic can be computed without any party revealing their raw data. Therefore, differential privacy [DP; 63] guarantees can be achieved by noising the *final* computed statistic, rather than to the original raw data, ensuring less loss of accuracy.
- We experimentally validate our methodology across a range of synthetic and semi-synthetic tasks, demonstrating strong agreement between our prospective rankings and the true rankings obtained after performing the merge. Moreover, we find that our proposed methodology to target CATE parameters improves over traditional Bayesian data selection and a number of other baselines. Finally, we show that, for the same level of privacy guarantees, our MPC protocol chained with DP performs better than applying DP to the raw inputs in the linear case.

3.2 Problem Statement, Assumptions & Notation

Notation The random variables X, T, and Y represent the covariates, treatment, and outcomes, with domains $\mathcal{X}$, $\{0,1\}$, and $\mathcal{Y}$, respectively; $\mathbf{x}$, t, and y denote realisations of these variables. We let $\mathcal{D}_e$ be the dataset comprised of elements $\{\mathbf{x}_i, t_i, y_i\}_{i=1}^{n_e}$, drawn i.i.d. from a distribution $P_e(\mathbf{x}, t, y)$, where $e \in \{0, \cdots, K\}$ indexes the datasets. Vectors of observations in $\mathcal{D}_e$ are denoted in bold, i.e. $\mathbf{y}_e = (y_i)_{i=1}^{n_e}$, $\mathbf{t}_e = (t_i)_{i=1}^{n_e}$, and $\mathbf{X}_e = (\mathbf{x}_i)_{i=1}^{n_e}$ refers to the data matrix. We use potential outcomes framework [182], so that $Y(t)$ represents the outcome resulting from an intervention setting $T = t$.

Assumptions and objective Throughout we focus on estimating the *Conditional Average Treatment Effect* (CATE), given by:

$$\tau(\mathbf{x}) = \mathbb{E}\left[Y(1) - Y(0) | X = \mathbf{x}\right]$$

To estimate CATE, we begin with an initial dataset, $\mathcal{D}_0$, referred to as the *host*. Our goal is to accurately estimate CATE with respect to the distribution of this dataset, $P_0(\mathbf{x})$. Specifically, we focus on minimising the Precision in Estimation of Heterogeneous Effects [PEHE; 144] given by $\epsilon_{\text{PEHE}}(f) = \int_{P_0(\mathbf{x})} \left(\hat{\tau}_f(\mathbf{x}) - \tau(\mathbf{x})\right)^2 d\mathbf{x}$, where $\hat{\tau}_f$ is the CATE estimate arising the outcome model f.

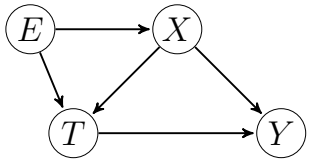


Figure 3.2.1: Assumed DAG

We consider a set of potential datasets for merging, $\mathcal{D}_e$, referred to as the *candidate* sites. In these candidate datasets, we assume that the outcomes are unmeasured or masked, and so denote them by $\mathcal{D}_e = \{\mathbf{x}_i^e, t_i^e, Y_i^e\}_{i=1}^{n_e}$ to show the randomness in Y_i^e. The goal is to prospectively identify which of the candidate datasets $\mathcal{D}_e$, would reduce the uncertainty over CATE if we were to measure or unmask the Y_i^e's and merge with the host dataset $\mathcal{D}_0$.

We assume that datasets are generated according to the Directed Acyclic Graph (DAG) in Figure 3.2.1.[1] This implies that $P(y|\mathbf{x}, t, e)$ is fixed across environments, but both covariate distributions and treatment allocation schemes are free to vary i.e. $P(\mathbf{x}, t|e)$ can depend on e. We also assume positivity over the whole population, so that $\forall\ \mathbf{x}, 0 < P(T = 1|X = \mathbf{x}) < 1$, but allow for violations at the site level. Adding the consistency assumption (i.e. $Y(t) = Y$ when $T = t$) to positivity and the causal structure in Figure 3.2.1 (which implies no hidden confounders) we have that CATE is identifiable [162, 175], constant across environments e, and given by:

$$\begin{aligned}\tau(\mathbf{x}) &= \mathbb{E}\left[Y|X = \mathbf{x}, T = 1\right] - \mathbb{E}\left[Y|X = \mathbf{x}, T = 0\right] \\ &= \mathbb{E}\left[Y|X = \mathbf{x}, T = 1, E = e\right] \\ &\quad - \mathbb{E}\left[Y|X = \mathbf{x}, T = 0, E = e\right]\end{aligned}$$

3.3 Method

We take a Bayesian approach to modelling the CATE. Specifically, we define $f_\theta : \mathcal{X} \times \{0, 1\} \to \mathcal{Y}$ to be a function representing the expected outcome conditional on covariates and treatment, i.e., $f_\theta(\mathbf{x}, t)$ seeks to approximate $\mathbb{E}\left[Y|X = \mathbf{x}, T = t\right]$. We assign a prior distribution $P(\theta)$, to the parameters θ and denote the conditional likelihood of the outcome given parameters, covariates, and treatment as $P(Y|\theta, \mathbf{x}, t)$,

[1] We make us of the SWIG framework to combine causal graphical models with potential outcomes. More details can be found in Richardson and Robins [175].

which is chosen appropriately based on the type of outcome being modelled. For example, we can select a normal likelihood for continuous outcomes, or a Bernoulli likelihood for binary ones. The data-generating process can be written as

$$\theta \sim P(\theta), \quad Y|f_\theta(\mathbf{x}, t) \sim P(Y|\theta, (\mathbf{x}, t)).$$

Throughout this paper we focus on continuous y and use a normal likelihood with fixed variance σ^2, i.e. $P(Y|\mathbf{x}, t, \theta) = \mathcal{N}(Y; f_\theta(\mathbf{x}, t), \sigma^2)$. CATE is then estimated via the current posterior mean, so that if we have conditioned on data $\mathcal{D}$ our estimate is $\hat{\tau}(\mathbf{x}) = \mathbb{E}_{\theta \sim p(\theta|\mathcal{D})}[f_\theta(\mathbf{x}, 1) - f_\theta(\mathbf{x}, 0)]$.

3.3.1 Quantifying Data Merge Utility through Expected Information Gain

In order to quantify the value of a data merge, we draw inspiration from Bayesian experimental design [BED; 37, 169]. BED applies information theory to provide a measure of what performing a particular experiment would tell us about a parameter of interest, relative to our current beliefs about the parameter value. In our context, the 'design' of the experiment corresponds to choosing a dataset $\mathcal{D}_e$, and the outcome is observing $\{Y_i^e\}_{i=1}^{n_e}$. The value of an experiment is quantified via the expected information gain [EIG; 139], which measures the expected reduction in uncertainty in the parameters, as measured by Shannon entropy, when moving from the post host posterior, $P(\theta|\mathcal{D}_0)$ to the post merge posterior, $P(\theta|\mathcal{D}_0, \mathcal{D}_e)$:

$$\text{EIG}_{\theta|\mathcal{D}_0}(e) = \mathbb{E}[H[P(\theta|\mathcal{D}_0)] - H[P(\theta|\mathcal{D}_0, \mathcal{D}_e)]],$$

where the expectation is over $P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) = \mathbb{E}_{P(\theta|\mathcal{D}_0)}[P(\mathbf{y}_e|\theta, \mathbf{X}_e, \mathbf{t}_e)]$—the Bayesian marginal distribution of $\mathbf{y}_e$. The EIG is equivalent to the mutual information between parameters and outcomes under the design and can be equivalently written as:

$$\text{EIG}_{\theta|\mathcal{D}_0}(e) = \mathbb{E}\left[\log \frac{P(\mathbf{y}_e|\theta, \mathbf{X}_e, \mathbf{t}_e)}{P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}\right], \tag{3.1}$$

where the expectation is over $P(\mathbf{y}_e, \theta|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)$. This form known as Bayesian Active Learning by Disagreement (BALD) in the active learning literature [102]. For certain classes of models, such as polynomial regression or Gaussian processes, the EIG is

available in closed form [189]. In other cases if the likelihood function isn't analytically available, we can approximate it using nested Monte Carlo [NMC; 168, 83] as follows:

$$\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathcal{D}_0}(e) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{P(\mathbf{y}_e^{(i)}|\theta^{(i)}, \mathbf{X}_e, \mathbf{t}_e)}{\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}, \tag{3.2}$$

where $\theta^{(i)}, \mathbf{y}_e^{(i)} \sim P(\theta|\mathcal{D}_0)P(\mathbf{y}_e|\theta^{(i)}, \mathbf{X}_e, \mathbf{t}_e)$, and further M_1 samples $\theta'_j \sim P(\theta|\mathcal{D}_0)$ for the denominator

$$\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) = \frac{1}{M_1}\sum_{j=1}^{M_1} P\left(\mathbf{y}_e^{(i)}|\theta'_j, \mathbf{X}_e, \mathbf{t}_e\right). \tag{3.3}$$

Note that since we assume a normal likelihood, we can construct a Rao-Blackwellised estimator by analytically computing the entropy of the likelihood in Eq. 3.1 to give:

$$\begin{aligned}\widehat{\mathrm{EIG}}^{\mathrm{RB}}_{\theta|\mathcal{D}_0}(e) = &-\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{1}{M_1}\sum_{j=1}^{M_1} P(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \theta'_j)\right) \\ &- \frac{n_e}{2}(1 + \log(2\pi\sigma^2)).\end{aligned}$$

Nested estimators are biased but consistent, converging to the true EIG at a rate $\mathcal{O}((N^{-1} + cM_1^{-2})^{\frac{1}{2}})$ [168]. A detailed algorithm for both estimators is given in Appendix 3.A.1.

These estimators allow us gauge the value of a merge before observing outcomes $\{Y_i^e\}_{i=1}^{n_e}$ in the dataset $\mathcal{D}_e$. However, whilst this approach provides us with information on the parameters for the conditional outcome, it may not necessarily lead to improved CATE predictions. This is because $\mathrm{EIG}_{\theta|\mathcal{D}_0}$ encourages *uniform entropy reduction* across all dimensions of θ, not just the ones relevant for CATE estimation. For instance, a dataset with untreated individuals only would provide information about the conditional outcome, but less for CATE, which requires viewing treated individuals as well. This motivates our improved methodology, which we present in the next section.

3.3.2 EIG Targeting CATE Parameters

Many causal inference models have additional parameter structure that allows us to target CATE estimation more directly. Specifically, the parameter set, θ can often be split as $\theta = \theta_{\mathrm{c}} \cup \theta_{\mathrm{nc}}$ where θ_{c} parameterises the CATE model and θ_{nc} is a set of nuisance parameters. For example, in Bayesian Causal Forests [91] the model is parameterised

as $f_\theta(\mathbf{x}, t) = \mu_{\theta_{\text{nc}}}(\mathbf{x}) + t\tau_{\theta_{\text{c}}}(\mathbf{x})$, where $\mu_{\theta_{\text{nc}}}$ and $\tau_{\theta_{\text{c}}}(\mathbf{x})$ jointly model the conditional outcome, and $\tau_{\theta_{\text{c}}}(\mathbf{x})$ directly models CATE. Leveraging such a parameterisation, we can prioritise uncertainty reduction in θ_c, and therefore CATE, by maximising

$$\text{EIG}_{\theta_{\text{c}}|\mathcal{D}_0}(e) = \mathbb{E}\left[\log \frac{P(\mathbf{y}_e|\theta_{\text{c}}, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}{P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}\right], \tag{3.4}$$

where expectation is over $P(\mathbf{y}_e, \theta_{\text{c}}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)$. Here:

$$P(\mathbf{y}_e|\theta_c, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) = \mathbb{E}_{P(\theta_{nc}|\theta_c, \mathcal{D}_0)}[P(\mathbf{y}_e|\theta, \mathbf{X}_e, \mathbf{t}_e)], \tag{3.5}$$

is the Bayesian distribution of the outcomes conditional on the CATE-related parameters only, and is generally not available in closed form. To deal with this intractability, we suggest approximating both the numerator and denominator empirically, yielding the following estimator:

$$\widehat{\text{EIG}}^{\text{NMC}}_{\theta_{\text{c}}|\mathcal{D}_0}(e) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\widehat{P}(\mathbf{y}_e^{(i)}|\theta_{\text{c}}^{(i)}, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}{\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}, \tag{3.6}$$

where $\theta_c^{(i)}, \mathbf{y}_e^{(i)} \sim P(\theta_{\text{c}}|\mathcal{D}_0)P(\mathbf{y}_e|\theta_{\text{c}}, \mathbf{X}_e, \mathbf{t}_e)$, the denominator $\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)$ is as in Eq. 3.3, and use further M_2 samples $\theta_{nc}^{(ik)} \sim P(\theta_{nc}^{(ik)}|\theta_{\text{c}}^{(i)}, \mathcal{D}_0)$ for the numerator:

$$\widehat{P}(\mathbf{y}_e^{(i)}|\theta_{\text{c}}^{(i)}, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) = \frac{1}{M_2}\sum_{k=1}^{M_2} P\left(\mathbf{y}_e^{(i)}|\theta_{nc}^{(ik)} \cup \theta_c^{(i)}, \mathbf{X}_e, \mathbf{t}_e\right)$$

This ensures that we prioritise a gain in information in the part of the model directly responsible for CATE. We again give an algorithm in Appendix 3.A.1.

3.3.3 Procedure and Model Classes

We now apply both procedures to three popular Bayesian causal inference methods: Bayesian Polynomial Regression, Bayesian Causal Forests [91], and Causal Multi-task Gaussian Processes [4]. We describe the standard predictive method as well as the parameter split used to target CATE.

Bayesian Polynomial Regression Due to its ubiquity across numerous fields, we first apply our method to Bayesian Polynomial Regression model. This involves specifying an initial polynomial transformation[2] $\phi : \mathcal{X} \times \mathcal{T} \to \mathbb{R}^p$, a mean μ_0, and precision matrix Λ_0 to parameterise the prior as:

$$f_\theta(\mathbf{x}, t) = \phi(\mathbf{x}, t)^\top \theta, \quad \theta \sim \mathcal{N}(\mu_0, \sigma^2\Lambda_0^{-1}). \tag{3.7}$$

[2]This could be an arbitrary non-linear function but we focus on polynomial transformations.

ϕ allows for higher order, or interaction terms. We further assume $\phi(\mathbf{x}, t)$ and θ can be split as:

$$\phi(\mathbf{x}, t) = \begin{bmatrix} \phi_{\text{nc}}(\mathbf{x}) & t\phi_{\text{c}}(\mathbf{x}) \end{bmatrix}^\top, \quad \theta = \begin{bmatrix} \theta_{\text{nc}} & \theta_{\text{c}} \end{bmatrix}^\top$$

This covers a broad range of Bayesian polynomial regressions, including the selection used in Gelman et al. [86]. In Appendix 3.B.2.1 we show that both EIGs can be computed in closed form:

Proposition 18. *For the Bayesian Polynomial Regression model defined in Eq. 3.7 we have:*

$$\begin{aligned} \text{EIG}_{\theta|\mathcal{D}_0}(e) &= \log\det\left(\Phi_e^\top\Phi_e + \Phi_0^\top\Phi_0 + \Lambda_0\right) + C \\ \text{EIG}_{\theta_{\text{c}}|\mathcal{D}_0}(e) &= \log\det(\Phi_{c,e}^\top\Phi_{c,e} + \Phi_{c,0}^\top\Phi_{c,0} + \Lambda_0^{[\text{c,c}]}) + C', \end{aligned}$$

where $\Phi_e = \phi(\mathbf{X}_e, \mathbf{t}_e)$, $\Phi_0 = \phi(\mathbf{X}_0, \mathbf{t}_0)$, $\Phi_{c,e} = \mathbf{t}_e \odot \phi_{\text{c}}(\mathbf{X}_e)$, $\Phi_{c,0} = \mathbf{t}_0 \odot \phi_{\text{c}}(\mathbf{X}_0)$, *with* ϕ, ϕ_c *applied row-wise and* $\odot$ *denoting element-wise multiplication;* C, C' *are constant in* e.

Bayesian Causal Forest Bayesian Causal Forests [BCF; 91] are one of the most popular causal inference methods, building upon Bayesian Additive Regression Trees [BART; 44], which are themselves a mainstay in observational causal inference [98]. The BCF model can be expressed as $f_\theta(\mathbf{x}, t) = \mu_{\theta_{\text{nc}}}(\mathbf{x}) + t\tau_{\theta_{\text{c}}}(\mathbf{x})$, where $\mu_{\theta_{\text{nc}}}$ and $\tau_{\theta_{\text{c}}}(\mathbf{x})$ are independent BART models; further details and alternative parameterisations can be found in Appendix 3.B.1. As the posterior is only available via sampling, we estimate both EIGs using NMC as given in Eq. 3.2 and Eq. 3.6.

Causal Multi-task Gaussian Processes Causal multi-task Gaussian Processes [4] use a vector-valued GP [7] to jointly model the conditional outcomes, allowing information sharing between them:

$$\mathbf{f} = \begin{bmatrix} f(\mathbf{x}, 0) & f(\mathbf{x}, 1) \end{bmatrix}^\top, \text{ where } \mathbf{f} \sim \mathcal{GP}(0, \mathbf{K}) \tag{3.8}$$

for a vector-valued kernel $\mathbf{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{2\times 2}$. The outcomes are then modelled by evaluating the relevant portion of the GP. Under this setting CATE is given by $\tilde{\tau} = \mathbf{ef}$ for $\mathbf{e} = \begin{bmatrix} -1 & 1 \end{bmatrix}$, meaning that $\tilde{\tau}$ is also a GP given by $\tilde{\tau} \sim \mathcal{GP}(0, \mathbf{e}^\top\mathbf{Ke})$. The advantage of causal multi-task GPs is that they allow us to get a closed form posterior for τ without having to observe any samples from $Y(1) - Y(0)$.

As GPs are inherently non-parametric they do not fit directly into the framework laid out in subsection 3.3.1 and subsection 3.3.2. This is not a problem for the predictive case as we can replace θ with $\mathbf{f}$ in Eq. 3.1 and get closed form expressions [102]. However, for the causal case this creates challenges as we cannot directly evaluate the expressions in Eq. 3.4 with $\tilde{\tau}$ in the place of θ_c. To resolve this we instead focus on

entropy reduction in CATE predictions on the host dataset. We denote this by $\tilde{\tau}(\mathbf{X}_0)$, where $\mathbf{X}_0$ is the host data matrix. The information gains e denote these by $\mathrm{EIG}_{\mathbf{f}|\mathcal{D}_0}(e)$ and $\mathrm{EIG}_{\tilde{\tau}(\mathbf{X}_0)|\mathcal{D}_0}(e)$ respectively. As the following proposition shows, both of these are now available in closed form:

Proposition 19. *Let $n_e^{(t)}$ be the number of subjects receiving treatment t in dataset e. For the causal multi-task GP model, defined in Eq. 3.8 we have*

$$\begin{aligned}\mathrm{EIG}_{\mathbf{f}|\mathcal{D}_0} &= \frac{1}{2}\log\det\left(\Sigma_1\right) \\ &\quad - n_e^{(0)}\log(\sigma_0) - n_e^{(1)}\log(\sigma_1) \\ \mathrm{EIG}_{\tilde{\tau}(\mathbf{X}_0)|\mathcal{D}_0}(e) &= \frac{1}{2}\log\left(\det(\Sigma_1)\det(\Sigma_2)\right) \\ &\quad - \frac{1}{2}\log\left(\det(\Sigma)\right),\end{aligned}$$

where $\Sigma_1, \Sigma_2, \Sigma$ and the proof are given Appendix 3.B.3.

3.4 Privacy

For privacy we use *Multi-Party Computation* [MPC; 69]. First introduced by Yao [226], MPC focuses on a setting where m separate parties wish to compute the value of a function $f(x_1, \cdots, x_m)$ where the i^{th} party inputs x_i and wishes to keep this private. To resolve this, MPC involves the specification of a protocol of message passing between parties which if followed would lead to the computation of f. In this work, we focus on the *semi-honest* setting, in which all parties follow the specified protocol, but some *corrupt* parties will try to learn as much about their peers inputs in the process. The goal is to devise a protocol which will preserve the privacy of the non-corrupt party's inputs, up to a given computational budget by the adversary. In our setting this means that any collection of corrupt sites are unable to learn anything about the other sites data during the EIG calculation, so any noise needed for privacy guarantees can be added to the final statistic.

For implementing multi-party computation, we employ the open source library CrypTen [121]. CrypTen builds upon PyTorch [161] allowing for standard tensor operations to be performed in an MPC protocol. For arithmetic operations on floating-point values this is achieved as follows: A float, x_F, is multiplied by some large scaling factor $B = 2^L$ and rounded to the nearest integer $\lfloor x_F \rceil$, where L is the number of precision bits. The integer $\lfloor x_F \rceil$ is then associated with its equivalence class $x \in \mathbb{Z}/Q\mathbb{Z}$ where $\mathbb{Z}/Q\mathbb{Z}$ is a ring of Q elements. The value x can then be shared across all m parties using Shamir secret sharing [192], in which each party gets access to a share of x is given by $[x]_i \in \mathbb{Z}/Q\mathbb{Z}$ which is generated such that the sum of all shares recovers the original value, so $x = \sum_{i=1}^{m} [x]_i \mod Q$. At any point all parties can combine their

shares to decode the output as $x_F \approx x/B$. We let $[x] = \{[x]_i\}_{i=1}^m$ denote the set of all shares corresponding to the secret value x.

Arithmetic operations building on addition are performed locally, so that for two secret values $[x], [y]$ each party performs $[z]_i = [x]_i + [y]_i$ and the result, z is obtained by all parties summing their share. M ultiplication is implemented using Beaver triples [20], logarithms are approximated using householder iterations [103], and reciprocals use Newton-Raphson. We implement log-determinants using Cholesky LDL decompositions, which are preferred to standard Cholesky factorisations as they avoid the use of square roots, which would require additional approximation in Crypten. This is possible as we only compute the log determinant of positive semi-definite matrices. This provides all the operations necessary to implement the above EIG calculations in a private manner using MPC.

When returning EIG statistics to the host, we add a small amount of noise to prevent information leakage. If we only need to output the best site, we use the exponential protocol [63] ensuring minimal information leakage. Finally, we note that the discretisation required to represent a float in the ring $\mathbb{Z}/Q\mathbb{Z}$ involves some degree of precision loss. Nevertheless, as we empirically demonstrate in the next section, this leads to minimal depreciation in performance compared to differential privacy.

3.5 Experiments & Results

We experimentally validate our approach in a setting where the host has to rank a number of candidate datasets based on the estimated gain from merging. We use selection of synthetic and semi-synthetic benchmark datasets as this allows us to use the known CATE to get a ground truth ranking. For each model, we do this by ranking datasets on the true PEHE on $P_0(\mathbf{x})$ of the relevant model trained on the merged dataset $\mathcal{D}_0 \cup \mathcal{D}_e$. This is then compared against implied EIG rankings. Throughout, we tune any model hyper-parameters on the host site when measuring the information gain as well as re-tuning parameters on each merged dataset when getting the ground truth ranking.

To generate the datasets, we begin with an initial, large dataset $\mathcal{D}$ from which we subsample the host and candidate sites. We do this by choosing a selection function, $S_e(\mathbf{x}, t)$, and using it to subsample a dataset of size n_e, where $S_e(\mathbf{x}_i, t_i)$ is the probability of subsampling point i for dataset $\mathcal{D}_e$. Varying the selection functions across sites ensures heterogeneity amongst datasets whilst complying with the independences given by the causal structure in Figure 3.2.1. We begin by providing an illustrative example on synthetic data before evaluating on the causal benchmarks, specifically: Lalonde [130] and Infant Health and Development Program [IHDP; 144]. Further details and results can be found in Appendix 3.C and 3.D, respectively.

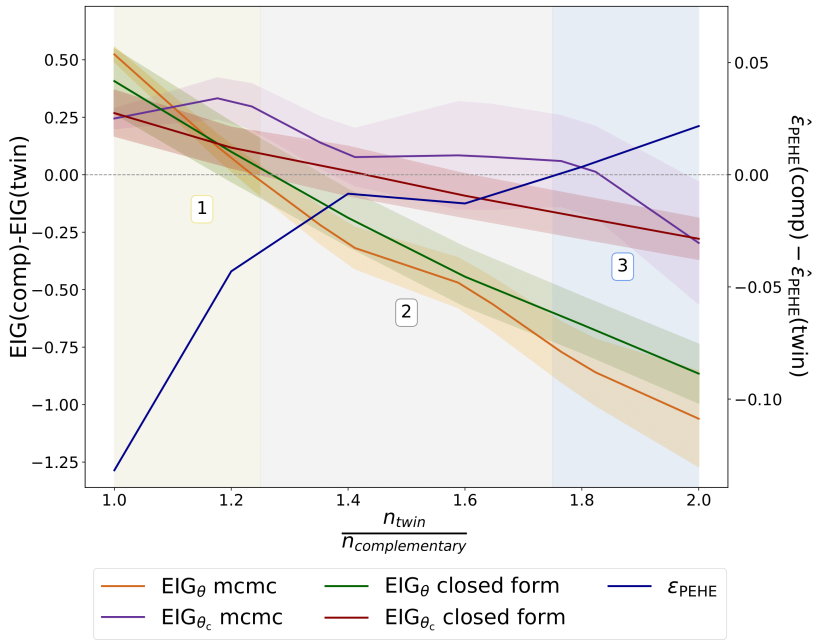


Figure 3.5.1: Difference in post host EIG and $\hat{\epsilon}_{\text{PEHE}}$ for a linear CATE model trained on $\mathcal{D}_0 \cup \mathcal{D}_{comp}$ and $\mathcal{D}_0 \cup \mathcal{D}_{twin}$ for increasing $\frac{n_{twin}}{n_{comp}}$ and fixed n_{host}=n_{comp}=100. PEHE is evaluated on hold out data from the host distribution. Lines show mean ±1 s.d. across 50 seeds. The three regions show different datasets preferences. In region 2, $\text{EIG}_{\theta|\mathcal{D}_0}$ incorrectly favours $\mathcal{D}_{twin}$, whilst $\text{EIG}_{\theta_c|\mathcal{D}_0}$ correctly selects $\mathcal{D}_{comp}$ for the causal task.

3.5.1 Illustrative experiment

Concept To illustrate the difference between our two methods—the standard approach based on the full parameter set, EIG_θ (Eq. 3.1), and the CATE-targeting one, EIG_{θ_c} (Eq. 3.4)—we start with a simple example where the host must choose between two candidate sites. We design these sites as follows: a *complementary* site, representing the "ideal" merge for causal purposes, and a *twin* site, containing information similar to the host. To create these, we first simulate an initial large dataset $\mathcal{D}$ as if it were a randomised controlled trial with an equal probability of treatment and subsample the host dataset, $\mathcal{D}_0$, using a selection function $S_0(\mathbf{x}, t)$. The data of the complementary site, $\mathcal{D}_{comp}$, is subsampled using $1 - S_0(\mathbf{x}, t)$ as a selection function, whilst the twin site uses $S_0(\mathbf{x}, t)$. This ensures the twin dataset mirrors the host's distribution and the complementary causally "complements" it. Assuming equal sizes, merging $\mathcal{D}_0$ with $\mathcal{D}_{comp}$ would recreate the initial randomised trial $\mathcal{D}$, as the variable e acts as a collider for $\mathbf{x}$ and t, removing their conditional dependency (see causal DAG in Appendix 3.C.1). Therefore, $\mathcal{D}_{comp}$ represents an ideal merge as it balances treatment allocation, whereas $\mathcal{D}_{twin}$ covers similar regions of the data space to those in the host, $\mathcal{D}_0$, potentially amplifying pre-existing biases and imbalances.

Model	Objective	$\rho(\uparrow)$	p@1 ($\uparrow$)	p@3 ($\uparrow$)	p@5 ($\uparrow$)
Polynomial	$\text{EIG}_{\theta_{c\|\mathcal{D}_0}}$	$\mathbf{0.70 \pm 0.08}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.04}$	$\mathbf{0.78 \pm 0.04}$
	$\text{EIG}_{\theta\|\mathcal{D}_0}$	$\mathbf{0.68 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.06}$	$\mathbf{0.76 \pm 0.04}$
	Best baseline	0.40 ± 0.11	0.40 ± 0.15	0.60 ± 0.15	0.66 ± 0.04
Causal GP	$\text{EIG}_{\tilde{\tau}(X_0)\|\mathcal{D}_0}$	$\mathbf{0.49 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.50 \pm 0.08}$	$\mathbf{0.62 \pm 0.06}$
	$\text{EIG}_{\mathbf{f}\|\mathcal{D}_0}$	0.33 ± 0.06	0.30 ± 0.15	0.43 ± 0.05	$\mathbf{0.60 \pm 0.04}$
	Best baseline	0.31 ± 0.12	0.10 ± 0.20	0.20 ± 0.07	0.46 ± 0.05
Bayesian CF	$\text{EIG}_{\theta_{c\|\mathcal{D}_0}}$	$\mathbf{0.54 \pm 0.10}$	$\mathbf{0.60 \pm 0.15}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$
	$\text{EIG}_{\theta\|\mathcal{D}_0}$	0.36 ± 0.10	0.30 ± 0.14	0.50 ± 0.07	$\mathbf{0.66 \pm 0.05}$
	Best baseline	0.45 ± 0.11	$\mathbf{0.60 \pm 0.14}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$

Table 3.5.1: Ranking experiment for the IHDP dataset, measured by Spearman ρ and precision at k (p@k). We include the best performing baseline method, which is different for different models.

For the illustrative experiment, we vary the ratio of sample sizes, $\frac{n_{twin}}{n_{comp}}$, in order to compare which dataset is chosen by $\text{EIG}_{\theta|\mathcal{D}_0}$ and the causally targeted $\text{EIG}_{\theta_c|\mathcal{D}_0}$ for linear regression. The aim is to demonstrate that $\text{EIG}_{\theta|\mathcal{D}_0}$ will prefer $\mathcal{D}_{twin}$ at points where $\mathcal{D}_{comp}$ dataset is still the preferable dataset for CATE estimation. Indeed, for large values of the ratio $\frac{n_{twin}}{n_{comp}}$, $\mathcal{D}_{twin}$ provides significant information about the conditional outcome, but not in the regions that are most relevant for causal estimation. On the other hand, EIG_{θ_c} should continue to select $\mathcal{D}_{twin}$ whilst it remains preferable for CATE estimation. We simulate $\mathbf{x} \in \mathbb{R}^3$ where x_1 is Bernoulli and other covariates are normal. The true outcome is sampled from a normal linear model, and the selection functions are logistic regressions. We also include sample based estimates for each EIG to demonstrate how they differ from their closed form counterparts. Experimental details provided in Appendix 3.C.

Results Figure 3.5.1 shows the results of the experiment divided into three regions. In region one, both methods choose the complementary dataset over the twin dataset which is consistent with ground truth ranking given by the PEHE upon merging. In region two, $\text{EIG}_{\theta|\mathcal{D}_0}$ chooses $\mathcal{D}_{twin}$ whilst $\text{EIG}_{\theta_{c|\mathcal{D}_0}}$ opts for $\mathcal{D}_{comp}$. Here, the complementary dataset is still the optimal in terms of CATE, but $\text{EIG}_{\theta|\mathcal{D}_0}$ preferences $\mathcal{D}_{twin}$ as it leads to a greater entropy reduction in the full set of parameters. This result shows that by focusing on the causal parameters alone, $\text{EIG}_{\theta_{c|\mathcal{D}_0}}$ is able to make the correct decision in selecting $\mathcal{D}_{comp}$. In the final region we see all lines have crossed the x axis, showing that all methods agree with the ground truth in choosing $\mathcal{D}_{twin}$. Finally, we note the MCMC estimates agree with the closed form counterparts, with increased variability due to sampling.

Method	MSE (↓)	$\boldsymbol{\rho}$(↑)
MPC Linear	$\mathbf{(3.80 \pm 0.04) \times 10^{-6}}$	$\mathbf{0.797 \pm 0.06}$
DP Linear	9.80 ± 0.30	0.06 ± 0.11

Table 3.5.2: Multi-Party Computation Results for EIG_{θ_c}.

3.5.2 Ranking experiment

Concept For our main experiment we validate our framework in a setting where the host must choose between many potential candidates, each with different distributions. To do so we begin with a standard causal inference benchmark dataset, $\mathcal{D}$, and form the host and candidates datasets using subsampling functions, $S_e(\mathbf{x}, t)$ as detailed above where subsampling function is a logistic regression with random parameters. This ensures that each site has a different covariate distribution. We apply the three methods described in Section subsection 3.3.3 to estimate both the two expected information gains for each candidate site, $\mathcal{D}_e$. Ultimately, like before, we compare the implied rankings with the ground truth ranking given by the PEHE. Full details are provided in Appendix 3.C.

Baselines We provide a number of simple comparison methods as baselines for our task. Specifically, (a) ranking by sample size, (b) ranking by similarity of covariate distribution measured by a multivariate Gaussian fit to the host, and (c) ranking by dissimilarity of treatment allocation measured by the error of a propensity model fit on the host. We compare using Spearman rho(ρ) and precision at k (p@k).

Results Table 3.5.1 shows the results of our experiment on the IHDP dataset [144] where the host has to rank the best datasets out of 10 candidates. We report the average performance of the rankings across 20 repeated experiments, and according to Spearman ρ [198] and precision at k. We include the best baseline by Spearman ρ performance. Additional metrics and baseline performance can be found in Appendix 3.D. These results demonstrate that across all models, our EIG_{θ_c} based approach, focusing on causal parameters, outperforms the standard approach which seeks expected entropy reduction in all parameters. Further, our method works significantly better than all baselines.

3.5.3 Multi-Party Computation Experiments

Finally, we experimentally demonstrate the performance of our cryptographic protocol. We repeat a similar experiment as in subsection 3.5.2 on the IHDP dataset, this time choosing between twenty datasets. Results are given using a linear model as this allows us to compare our MPC based approach against differential privacy [DP; 63, see Appendix 3.A.2 for definition]. We use $\epsilon = 100$ and Laplace noising [65], following

existing work on DP linear regression [23, 12]. In order to ensure a fair comparison, we noise the final statistics from the MPC computation with an appropriate amount of Laplace noise. This comes from the sensitivity of the EIG statistic which we derive in Appendix 3.A.3. Table 3.5.2 shows both the MSE induced by the privacy method and the Spearman ρ against the noise-free ranking. Results show that MPC vastly outperforms differential privacy, both in terms of MSE and ranking, with DP producing a near random ranking, despite the relaxed noising.

3.6 Related Work

Federated Learning Via Multi-Party Computation Our setting bares large with federated learning [135], a distributed machine learning approach that enables multiple parties to collaboratively train a shared model while keeping their raw data decentralised and private. Our goal differs from this as we focus not on learning a model across sites, but deciding which sites to pool. The application of multiparty computation within federated learning is a growing area [137, 157, 114], with much of the work focusing on how to learn predictive models in a secure cross site fashion. Most similar to our work is Muazu et al. [156], who develop a similar federated approach to data fusion focusing on healthcare, however they do not take a causal approach. To the best of our knowledge, there are no existing applications of multi-party computation within causal inference.

Federated/Private Causal Estimation There are, however, federated approaches to estimating causal effects. In this area, a majority of works partition the loss function into multiple components, with each component corresponding to a specific data source [212, 140, 213]. However, modelling complex, non-linear relationships remains challenging [6]. Many of these algorithms come without privacy guarantees, with the exception of Niu et al. [159], who add DP guarantees to various popular CATE estimation techniques. The latter approach however contrasts with ours as the use of sample splitting reduces data efficiency.

Causal Bayesian Active Learning Bayesian Active Learning by Disagreement [BALD; 102] is a framework for strategic training data acquisition, focusing on regions of high uncertainty. Our method based on maximising $\text{EIG}_{\theta|\mathcal{D}_0}$ (Eq. 3.1) can be viewed as applying BALD after an initial update to an entire datasets rather than individual datapoints. Most similar to our work is CausalBALD [108], which applies an active learning approach to CATE estimation. However, the acquisition function cannot easily be extended to datasets without ignoring the correlation in information provided by different points. Most applications of active learning in causality focus on causal discovery and intervention selection [205, 95, 10].

3.7 Discussion

Limitations Due to the cost of computing high dimensional determinants and performing multiple rounds of conditional sampling, our method can be computationally costly in high dimensions. Moreover, the cost of multi-party computation will also increase as higher order approximations are required to retain accuracy. However, both of these remain negligible compared to the data engineering and expenses of data fusion [31, 110]. Finally, whilst our method offers a secure and principled way to prospectively quantify the value of dataset merges, the amount of information needed to justify such expenses may vary depending on application. It might, therefore, be beneficial to consider introducing a problem-specific threshold to determine when a proposed merger is worthwhile.

Conclusion We introduce an information-theoretic, cryptographically secure framework for evaluating the utility of potential data merges for causal estimation. To the best of our knowledge, this is the first work addressing this relevant challenge. Through empirical evaluation, we demonstrate that our framework can reliably rank datasets according to their ability to improve CATE estimation. We show that entropy reduction in the CATE parameters alone gives an improvement when compared to a more standard approach to Bayesian dataset acquisition. Finally, we demonstrate that our cryptographic procedure can be applied in conjunction with DP, resulting in lower loss of accuracy, compared to applying DP alone.

Acknowledgements

We would like to thank Andrew Yiu, Amartya Sanyal, Shahine Bouhabid, Xi Lin, Patrick Hough, and Sahra Ghalebikesabi for their comments. JF gratefully acknowledges funding from the EPSRC. LTM is supported by EPSRC through the Modern Statistics and Statistical Machine Learning (StatML) CDT programme, grant no. EP/S023151/1.

Appendix

3.A Mathematical Details

3.A.1 Algorithms for Computing EIG

Algorithm 1 Algorithm for $\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathcal{D}_0}(e)$

Input: Data Matrix $\mathbf{X}_e$, Treatment Vector $\mathbf{t}_e$, Variance σ^2
Output: $\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathcal{D}_0}(e)$
Require: $l = NM_1$ for some $n, M_1 > 0$
  $S \leftarrow 0$
  Sample $\{\theta_i\}_{i=1}^{l} \sim P(\theta \mid \mathcal{D}_0)$ for $l = NM_1$
  Split $\{\theta_i\}_{i=1}^{l}$ as $\{(\theta_i), (\theta_{i,j})_{j=1}^{M_1}\}_{i=1}^{N}$
  for $i \in \{1, \cdots, N\}$ **do**
    Sample $\mathbf{y}_e^{(i)} \sim P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \theta_i)$
    $S \leftarrow S + \log \frac{P(\mathbf{y}_e^{(i)}|\theta_i, \mathbf{X}_e, \mathbf{t}_e)}{\frac{1}{M_1}\sum_{j=1}^{M_1} P(\mathbf{y}_e^{(i)}|\theta_{i,j}, \mathbf{X}_e, \mathbf{t}_e)}$
  end for
  $\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathcal{D}_0}(e) \leftarrow \frac{1}{N} S$
  return $\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathcal{D}_0}(e)$

Algorithm 2 Algorithm for $\widehat{\text{EIG}}^{\text{RB}}_{\theta|\mathcal{D}_0}(e)$

Input:$\{\theta_i\}_{i=1}^{l} \sim P(\theta \mid \mathcal{D}_0)$, Data Matrix $\mathbf{X}_e$, Treatment Vector $\mathbf{t}_e$, Variance σ^2
Output: $\widehat{\text{EIG}}^{\text{RB}}_{\theta|\mathcal{D}_0}(e)$

$S \leftarrow 0$
Sample $\{\theta_i\}_{i=1}^{l} \sim P(\theta \mid \mathcal{D}_0)$ for $l = NM_1$
Split $\{\theta_i\}_{i=1}^{l}$ as $\{(\theta_i), (\theta_{i,j})_{j=1}^{M_1}\}_{i=1}^{N}$
for $i \in \{1, \cdots, N\}$ **do**
  Sample $\mathbf{y}_e^{(i)} \sim P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \theta_i)$
  $S \leftarrow S - \log \frac{1}{M_1} \sum_{j=1}^{M_1} P(\mathbf{y}_e^{(i)}|\theta_{i,j}, \mathbf{X}_e, \mathbf{t}_e)$
end for
$\widehat{\text{EIG}}^{\text{RB}}_{\theta|\mathcal{D}_0}(e) \leftarrow \frac{1}{N} S$
return $\widehat{\text{EIG}}^{\text{NMC}}_{\theta|\mathcal{D}_0}(e)$

Algorithm 3 Algorithm for $\widehat{\text{EIG}}_{\theta_c|\mathcal{D}_0}(e)$

Input Data Matrix $\mathbf{X}_e$, Treatment Vector $\mathbf{t}_e$, Variance σ^2
Output: $\widehat{\text{EIG}}_{\theta_c|\mathcal{D}_0}(e)$

$S \leftarrow 0$
Sample $\{\theta_i\}_{i=1}^{N} \sim P(\theta \mid \mathcal{D}_0)$
for $i \in \{1, \cdots, N\}$ **do**
  Sample $\mathbf{y}_e \sim P(\mathbf{y}_e|\theta, \mathbf{X}_e, \mathbf{t}_e)$
  Sample $\{\theta'_j\}_{j=1}^{M_1} \sim P(\theta \mid \mathcal{D}_0)$
  Sample $\{(\theta_{\text{nc}})^{(ik)}\}_{k=1}^{M_2} \sim P(\theta_{\text{nc}} \mid (\theta_{\text{c}})_i, \mathcal{D}_0)$
  $\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) \leftarrow \frac{1}{M_1} \sum_{j=1}^{M_1} P\left(\mathbf{y}_e^{(i)}|\theta'_j, \mathbf{X}_e, \mathbf{t}_e\right)$
  $\widehat{P}(\mathbf{y}_e^{(i)}|\theta_{\text{c}}^{(i)}, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0) \leftarrow \frac{1}{M_2} \sum_{k=1}^{M_2} P\left(\mathbf{y}_e^{(i)}|\theta_{nc}^{(ik)} \cup \theta_c^{(i)}, \mathbf{X}_e, \mathbf{t}_e\right)$
  $S \leftarrow S + \log\left(\frac{\widehat{P}(\mathbf{y}_e^{(i)}|\theta_{\text{c}}^{(i)}, \mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}{\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathcal{D}_0)}\right)$
end for
return $\frac{1}{N} S$

3.A.2 Differential Privacy Definition

Definition 15. *We say a randomised algorithm, $\mathcal{A}$ satisfies ϵ differential privacy if for any input dataset $\mathcal{D}$ and dataset $\mathcal{D}'$ differing by a single entry, we have*

$$P(\mathcal{A}(\mathcal{D}) \in \mathcal{O}) \leq \exp(\epsilon) P(\mathcal{A}(\mathcal{D}') \in \mathcal{O}).$$

3.A.3 Sensitivity of Linear Statistic

We derive the sensitivity of the linear EIG statistic in order to give a fair comparison with naive differential privacy in subsection 3.5.3.

Proposition 20. *Let:*

$$f(\mathbf{X}_e) = \log\det(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{X}_0^\top \mathbf{X}_0 + \Lambda_0) \tag{3.9}$$

If $\Lambda_0 = cI$ *and* $\|\mathbf{x}\|_\infty \leq M$ *for all* $\mathbf{x} \sim P_e(\mathbf{x})$ *and* e*. Then we have that:*

$$\Delta_f = \max_{\mathbf{X}'_e, \mathbf{X}_e} |f(\mathbf{X}'_e) - f(\mathbf{X}_e)| \leq \frac{Md}{\sqrt{c}} \tag{3.10}$$

Where $\mathbf{X}'_e, \mathbf{X}_e$ *differ in at most one row. This implies* $f(\mathbf{X}_e) + Z$ *for* $Z \sim \text{Laplace}(\frac{Md}{\epsilon\sqrt{c}})$ *is a* ϵ *differentially private release of* $f(\mathbf{X}_e)$*.*

Proof. To prove this we will use the fact that if $\max_{\|\mathbf{X}\| \leq M} \|Df(\mathbf{X})\|_F = L$ we have that $|f(\mathbf{X}'_e) - f(\mathbf{X}_e)| \leq L \|\mathbf{X}'_e - \mathbf{X}_e\|_F$ for all $\mathbf{X}'_e, \mathbf{X}_e$ with norm bounded by M where $\|\cdot\|_F$ is the Frobenious norm. Write $\mathbf{E} = \mathbf{X}_0^\top \mathbf{X}_0 + \Lambda_0$, via the chain rule we have that $Df(\mathbf{X}) = \left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \mathbf{X}_e^\top$ Now:

$$\left\| \left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \mathbf{X}_e^\top \right\|_F = \sqrt{\text{tr}\left(\left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \mathbf{X}_e^\top \mathbf{X}_e \left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \right)} \tag{3.11}$$

$$= \sqrt{\text{tr}\left(\left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} - \left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \mathbf{E} \left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \right)} \tag{3.12}$$

$$\leq \sqrt{\text{tr}\left(\left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \right)} \leq \sqrt{\frac{d}{c}} \tag{3.13}$$

We have used the fact that $\left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1} \mathbf{E} \left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1}$ is positive semi definite and so has positive trace, and that the eigenvalues of $\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}$ are bounded below by c so the eigenvalues of $\left(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\right)^{-1}$ are bounded above by $\frac{1}{c}$. Finally for neighbouring datasets we can change at most d entries by M so $\|\mathbf{X}'_e - \mathbf{X}_e\|_F \leq \sqrt{d}M$ □

3.B Model Details

Throughout all methods will aim to model the outcome as some function plus an error, so:

$$Y_i = f(\mathbf{x}_i, t_i) + \epsilon_i \tag{3.14}$$

3.B.1 Models via Sampling

Bayesian Additive Regression Trees (BART) BART models [44] f as the sum of L piecewise constant binary regression trees, so we have:

$$f(\mathbf{x}, t) = \sum_{l=1}^{L} g_i(\mathbf{x}, t, T_l, \mathbf{m}_l) \tag{3.15}$$

where T_l is a regression tree given by a partition $(\mathcal{A}_1, \cdots, \mathcal{A}_{\mathcal{B}(l)})$ of $\mathcal{X} \times \mathcal{T}$ and the set of leaf parameter values $\mathbf{m}_l = (m_{l1}, \cdots, m_{l\mathcal{B}(l)})$ so that:

$$g_l(\mathbf{x}, t) = m_j \text{ if } \mathbf{x}, t \in \mathcal{A}_j \tag{3.16}$$

The mean parameters are given with independent normal parameters $m_{lj} \sim \mathcal{N}(0, \sigma_m)$. Over trees, the prior is such that the probability of a node having children at depth d is given by:

$$\alpha(1+d)^{-\theta} \text{ for } \alpha \in (0,1), \theta \in [0, \infty) \tag{3.17}$$

The original BART model explores this space using Metropolis-Hastings Markov chain Monte Carlo, but we make use of XBART [96] for accelerated posterior sampling.

Bayesian Causal Forest Bayesian Causal Forests [BCF; 91] build upon BART models utilising specific parameterisations for causal inference tasks. The two parameterisations suggested in Hahn et al. [91] are firstly:

$$f(\mathbf{x}, t) = \mu(\mathbf{x}) + t\tau(\mathbf{x}) \tag{3.18}$$

Where μ, τ are independent BART models. To draw specific attention to their parameters will write them as $\mu_{\theta_{\text{nc}}}, \tau_{\theta_{\text{c}}}$ noting that $\tau_{\theta_{\text{c}}}$ is a model for CATE. Hahn et al. [91] note that this parameterisation is not invariant to which treatment is assigned as positive or negative, leading them to propose the following invariant parameterisation:

$$f_\theta(\mathbf{x}, t) = \tilde{\mu}_{\tilde{\theta}_{\text{nc}}}(\mathbf{x}) + b_t \tilde{\tau}_{\tilde{\theta}_{\text{c}}}(\mathbf{x}) \tag{3.19}$$

Where $b_t \sim \mathcal{N}(0, \frac{1}{2})$. Under this parameterisation a CATE estimate is given by $(b_1 - b_0)\tilde{\tau}_{\tilde{\theta}_{\text{c}}}(\mathbf{x})$. When sampling we make use of the accelerated BCF approach [127] which builds upon XBART and uses the following slightly modified model:

$$f_\theta(\mathbf{x}, t) = a\tilde{\mu}_{\tilde{\theta}_{\text{nc}}}(\mathbf{x}) + b_t \tilde{\tau}_{\tilde{\theta}_{\text{c}}}(\mathbf{x}) \tag{3.20}$$

Where $a \sim \mathcal{N}(0, 1)$.

We define the set θ_{c} to be any parameters affiliated with the τ model, including b_t for the invariant parameterisation. In order to sample $P(\theta_{\text{nc}} | \theta_{\text{c}}, \mathcal{D}_0)$ we refit θ_{nc} parameters

on the dataset $\mathcal{D}_0$ to the residuals resulting from subtracting the τ portion of the model. So this refitting μ is as follows for the standard parametrisation:

$$Y - t\tau_{\theta_c}(\mathbf{x}) = \mu_{\theta_{\text{nc}}}(\mathbf{x}) \tag{3.21}$$

Or for the accelerated BCF approach [127]:

$$Y - b_t\tilde{\tau}_{\tilde{\theta}_c}(\mathbf{x}) = a\tilde{\mu}_{\tilde{\theta}_{\text{nc}}}(\mathbf{x}) \tag{3.22}$$

Where any parameters on the left hand side are fixed.

3.B.2 Closed Form Models

In this section we give details of models for which the EIG is available in closed form. We provide details of the models as well as proofs for the expressions.

3.B.2.1 Bayesian Polynomial Regression Derivations

In this we derive the results for Bayesian polynomial regression. We have modelled our data as:

$$y \sim \mathcal{N}(\phi(\mathbf{x}, t)^\top \theta, \sigma^2), \quad \theta \sim \mathcal{N}(\mu_0, \sigma^2\Lambda_0^{-1}) \tag{3.23}$$

In this context, the posterior is available in closed form as:

$$\theta|\mathcal{D}_0 \sim \mathcal{N}(\mu_0, \sigma^2\tilde{\Lambda}_0^{-1}) \tag{3.24}$$

$$\tilde{\Lambda}_0 = \left(\phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \phi\left(\mathbf{X}_0, \mathbf{t}_0\right) + \Lambda_0^{-1}\right) \tag{3.25}$$

$$\mu_0 = \left(\Lambda_0\right)^{-1}\left(\phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \phi\left(\mathbf{X}_0, \mathbf{t}_0\right)\hat{\theta}_0 + \Lambda_0\mu_0\right) \tag{3.26}$$

$$\hat{\theta}_0 = \left(\phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \phi\left(\mathbf{X}_0, \mathbf{t}_0\right)\right)^{-1} \phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \mathbf{y}_0 \tag{3.27}$$

Expected Information Gain Over all parameters We use the fact that $\text{EIG}_{\theta|\mathcal{D}_0}$ can be written as:

$$\text{EIG}_{\theta|\mathcal{D}_0}(e) = \mathbb{E}_{P(\mathbf{y}_e|\mathbf{X}_e,\mathbf{t}_e,\mathcal{D}_0)}\left[H[P(\theta|\mathcal{D}_0)] - H[P(\theta|\mathcal{D}_0, \mathcal{D}_e)]\right],$$

As the posterior over θ is Gaussian we can directly evaluate these expressions using the closed form entropy for a Gaussian distribution as:

$$H[P(\theta|\mathcal{D}_0)] = \frac{n_e}{2}\left(1 + \log(2\pi)\right) + \frac{1}{2}\left(\log\det\left(\tilde{\Lambda}_0^{-1}\right)\right)$$

The distribution $\theta|\mathcal{D}_0, \mathcal{D}_e)$ can be obtained as above, where we now use $\tilde{\Lambda}_0$ as the prior precision matrix before updating on $\mathcal{D}_0$. This gives:

$$H[P(\theta|\mathcal{D}_0)] = \frac{n_e}{2}\left(1 + \log(2\pi)\right) + \frac{1}{2}\left(\log\det\left(\left(\phi\left(\mathbf{X}_e, \mathbf{t}_e\right)^\top \phi\left(\mathbf{X}_e, \mathbf{t}_e\right) + \tilde{\Lambda}_0\right)^{-1}\right)\right)$$

Using the above form for $\tilde{\Lambda}_0$ and collecting all constants gives the expression presented in the text.

$\text{EIG}_{\theta_c|\mathcal{D}_0}$: **Expected Information Gain Over all parameters** This follows as above but using the fact that the block form of the matrices to allow us write the covariance precision matrix for the post host posterior over θ_c as follows:

$$\left(\mathbf{t}_0 \odot \phi_{\text{c}}(\mathbf{X}_0)\right)^\top \left(\mathbf{t}_0 \odot \phi_{\text{c}}(\mathbf{X}_0)\right) + (\Lambda_0)_{[\text{c,c}]}$$

Where $(\Lambda_0)_{[\text{c,c}]}$ corresponds to block of Λ_0 with entries after $[c, c]$ in the row and column.

3.B.3 Causal Multitask Gaussian Processes [4]

In this work, CATE is modelled using a multitask Gaussian process [28]. Multitask Gaussian Processes use a GP in vector-valued Reproducing Kernel Hilbert Space (vv-RKHS) to share information between tasks [7]. In Alaa and Van Der Schaar [4], learning the conditional outcome function for each treatment is seen as a separate task, so we jointly model:

$$Y|\mathbf{x}, t \sim \mathcal{N}(0, f_t(\mathbf{x}), \sigma_t^2) \tag{3.28}$$

Where each f_t is a Gaussian Process. The kernel $\mathbf{K}_\eta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{2\times 2}$ is now a symmetric positive semi-definite matrix-valued function, with hyper-parameters η. In the case of Alaa and Van Der Schaar [4] they use a *linear model of coregionalization*[3], giving the kernel as:

$$\mathbf{K}_\eta(\mathbf{x}, \mathbf{x}') = \mathbf{A}_0 k_0(\mathbf{x}, \mathbf{x}') + \mathbf{A}_1 k_1(\mathbf{x}, \mathbf{x}') \tag{3.29}$$

Where k_t is the RBF kernel, given by:

$$k_t(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2}\left(\mathbf{x} - \mathbf{x}'\right)^\top \mathbf{R}_t^{-1}\left(\mathbf{x} - \mathbf{x}'\right)) \tag{3.30}$$

Where $\mathbf{R}_t^{-1} = \text{diag}\left(\ell_{1,t}^2, \cdots, \ell_{d,t}^2\right)$ and $\ell_{j,t}$ is the length-scale parameter for the treatment $T = t$ in the j^{th} coordinate of $\mathbf{x}$. $\mathbf{A}_t$ is given by:

$$\mathbf{A}_0 = \begin{bmatrix} \theta_{00}^2 & \rho_0 \\ \rho_0 & \theta_{01}^2 \end{bmatrix}, \mathbf{A}_1 = \begin{bmatrix} \theta_{10}^2 & \rho_1 \\ \rho_1 & \theta_{11}^2 \end{bmatrix}. \tag{3.31}$$

Where $\theta_i j$ and ρ_i determine the variances and covariances of the shared tasks f_t. So we have that the full set of hyper-parameters $\eta = (\theta_0, \theta_1, \mathbf{R}_0, \mathbf{R}_1, \mathbf{A}_0, \mathbf{A}_1)$. Once all these hyper-parameters have been learnt we have that the covariance between different function evaluations, $f_t(\mathbf{x}), f_{t'}(\mathbf{x}')$, is given the t, t' coordinate of $\mathbf{K}_\eta(\mathbf{x}, \mathbf{x}')$. So:

$$\text{cov}(f_t(\mathbf{x}), f_{t'}(\mathbf{x}')) = \mathbf{K}_\eta(\mathbf{x}, \mathbf{x}')_{[t,t']} \tag{3.32}$$

[3] See Alvarez et al. [7] for more details.

Now, if we let $K((\mathbf{x},t),(\mathbf{x}',t') = \mathbf{K}_\eta(\mathbf{x},\mathbf{x}')_{[t,t']}$ then we can obtain the posterior kernel in a similar way to the standard case. Precisely if we the training data be given by:

$$\tilde{\mathbf{X}} = \left[\{\mathbf{x}_i\}_{T_i=0}, \{\mathbf{x}_i\}_{T_i=1}\right]^T, \tag{3.33}$$

$$\tilde{\mathbf{Y}} = \left[\left\{y_i^{(T_i)}\right\}_{T_i=0}, \left\{y_i^{(T_i)}\right\}_{t_i=1}\right]^T, \tag{3.34}$$

$$\mathbf{\Sigma} = \text{diag}\left(\sigma_0^2\mathbf{I}_{n-n_1}, \sigma_1^2\mathbf{I}_{n_1}\right) \tag{3.35}$$

$$n_1 = \sum_i W_i, \tag{3.36}$$

$$\mathbf{K}_\eta(x) = \left(\mathbf{K}_\eta\left(x, X_i\right)_i .\right) \tag{3.37}$$

Then we have that the posterior multitask GP has mean and posterior kernel given by:

$$m^{\text{post}}(\mathbf{x}) = \mathbf{K}_\eta^T(\mathbf{x})\left(\mathbf{K}_\eta(\mathbf{X},\mathbf{X}) + \Sigma\right)^{-1}\tilde{\mathbf{Y}} \tag{3.38}$$

$$\mathbf{K}_{\eta^*}^{\text{post}}(\mathbf{x},\mathbf{x}') = \mathbf{K}_{\eta^*}(\mathbf{x},\mathbf{x}') - \mathbf{K}_{\eta^*}(\mathbf{x})\left(\mathbf{K}_\eta(\mathbf{X},\mathbf{X}) + \Sigma\right)^{-1}\mathbf{K}_{\eta^*}^T(\mathbf{x}') \tag{3.39}$$

$$\tag{3.40}$$

This leads to the following posterior over CATE, where $\mathbf{e} = [-1,1]^\top$:

$$\tilde{\tau}(x) \sim \mathcal{N}(m^{\text{post}}(\mathbf{x})^\top\mathbf{e}, \mathbf{e}^\top\mathbf{K}_{\eta^*}^{\text{post}}(\mathbf{x},\mathbf{x}')\mathbf{e}) \tag{3.41}$$

3.B.3.1 Expected Information Gain

First, let $\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(0)}$ and $\mathbf{y}_e^{(1)}, \mathbf{y}_e^{(0)}$ be the covariance and outcomes for environment e that is treated and untreated respectively. To avoid confusion with treatment we will use $\mathbf{X}_{e^*}$ to refer to the host environment for this derivation. We will also use $\mathbf{K}_{|\mathcal{D}_0}$ to refer to the posterior kernel. Now to derive the Expected Information Gain in closed form we need the distribution of the following vector:

$$\begin{bmatrix} \mathbf{y}_e^{(1)} \\ \mathbf{y}_e^{(0)} \\ \tilde{\tau}(\mathbf{X}_{e^*}) \end{bmatrix} | \mathbf{X}_e, \mathcal{D}_0 \sim \mathcal{N}\left(\mathbf{m}, \Sigma\right) \tag{3.42}$$

Where we have that:

$$\Sigma_1 = \begin{bmatrix} \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_e^{(0)},\mathbf{X}_e^{(0)}) + \sigma_0^2 I_{n_e^0} & \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_e^{(1)},\mathbf{X}_e^{(0)})) \\ \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_e^{(0)},\mathbf{X}_e^{(1)}) & \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_e^{(1)},\mathbf{X}_e^{(1)}) + \sigma_1^2 I_{n_e^1} \end{bmatrix} \tag{3.43}$$

$$\Sigma_2 = \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(1)},\mathbf{X}_{e^*}^{(1)}) + \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(0)},\mathbf{X}_{e^*}^{(0)}) - 2\mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(1)},\mathbf{X}_{e^*}^{(0)}) \tag{3.44}$$

$$\Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_2 \end{bmatrix} \text{ where } \Sigma_{12} = \begin{bmatrix} \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(1)},\mathbf{X}_e^{(0)}) - \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(0)},\mathbf{X}_e^{(0)}) \\ \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(1)},\mathbf{X}_e^{(1)}) - \mathbf{K}_{|\mathcal{D}_0}(\mathbf{X}_{e^*}^{(0},\mathbf{X}_e^{(1)}) \end{bmatrix} \tag{3.45}$$

Now from the covariance matrix we can derive the standard Expected Information Gain $\text{EIG}_{\theta|\mathcal{D}_0}$ and CATE-specific $\text{EIG}_{\theta_{c|\mathcal{D}_0}}$. Throughout we will use $\mathbf{K}$ to the posterior kernel irrespective if it has been fit or not.

Expected Information Gain over the conditional outcome parameters For the $\text{EIG}_{\mathbf{f}}$ we use the BALD form:

$$H(\mathbf{y}_e|\mathbf{X}_e, \mathcal{D}_0) - H(\mathbf{y}|\mathbf{X}_e, \mathbf{f}, \mathcal{D}_0) \tag{3.46}$$

Using the standard form of entropy for a Gaussian distribution we can read $H(\mathbf{y}_e|\mathbf{X}_e, \mathcal{D}_0)$ off of the covariance matrix above as:

$$H(\mathbf{y}_e|\mathbf{X}_e, \mathcal{D}_0) = \frac{n_e}{2}\left(1 + \log(2\pi)\right) + \frac{1}{2}\left(\log|(\Sigma_1|)\right) \tag{3.47}$$

$$\text{where } \Sigma_1 = \begin{bmatrix} \mathbf{K}_\eta(\mathbf{X}_e^{(0)}, \mathbf{X}_e^{(0)}) + \sigma_0^2 I_{n_e^0} & \mathbf{K}_\eta(\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(0)})) \\ \mathbf{K}_\eta(\mathbf{X}_e^{(0)}, \mathbf{X}_e^{(1)}) & \mathbf{K}_\eta(\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(1)}) + \sigma_1^2 I_{n_e^1} \end{bmatrix} \tag{3.48}$$

And $H(\mathbf{y}_e|\mathbf{f}, \mathbf{X}_e, \mathcal{D}_0)$ being:

$$H(\mathbf{y}_e|\mathbf{f}, \mathbf{X}_e, \mathcal{D}_0) = \frac{n_e}{2}\left(1 + \log(2\pi)\right) + \frac{1}{2}\left(n_e^{(0)}\log(\sigma_0^2) + n_e^{(1)}\log(\sigma_1^2)\right) \tag{3.49}$$

This gives the Expected Information Gain as:

$$\mathcal{I}_{\mathbf{f}}(e) = \frac{1}{2}\log|(\Sigma_1|) - \frac{1}{2}\left(n_e^{(0)}\log(\sigma_0^2) + n_e^{(1)}\log(\sigma_1^2)\right) \tag{3.50}$$

Expected Information Gain on the CATE parameters We now target an Expected Information Gain on the CATE parameters on our host dataset, so $\tilde{\tau}(\mathbf{X}_0)$. By using the fact that the Expected Information Gain can be written as the mutual information between $\tilde{\tau}(X_0)$ and the observed outcomes in dataset e, we use the closed form mutual information for Gaussian's to directly write this as:

$$\mathcal{I}_{\tilde{\tau}(\mathbf{x}_0)}(e) = \frac{1}{2}\log\left(\frac{|\Sigma_1||\Sigma_2|}{|\Sigma|}\right) \tag{3.51}$$

$$\text{where } \Sigma_2 = \mathbf{K}_\eta(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_{e^*}^{(1)}) + \mathbf{K}_\eta(\mathbf{X}_{e^*}^{(0)}, \mathbf{X}_{e^*}^{(0)}) - 2\mathbf{K}_\eta(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_{e^*}^{(0)}) \tag{3.52}$$

3.C Experimental Details

General experimental settings and hyperparameters All standard deviations and precisions were taken equal to 1. Throughout experiments, priors were taken as zero-valued vector.

In the **illustrative experiment**, 400 samples were used for computing outer expectations, and 800 samples for inner expectations. Here, we consider $X = (X_0, X_1, X_2) \in \mathbb{R}^3$. We use the sampling selection function $P_{\text{host}}(x, t) = \text{sigmoid}(1 + 2 \times x_0 - x_1 + 2 \times t) + \epsilon$ and outcome model $y = 1 + x_0 - x_1 + x_2 + 5 \times t + 2 \times x_0 + 2 \times x_0 - 4 \times x_2 + \epsilon$ with $X_0 \sim \mathcal{B}(12, 3)$, $X_1 \sim \mathcal{N}(4, 1)$, $X_2 \sim \mathcal{B}(1, 7)$ and $\epsilon \sim \mathcal{N}(0, 1)$.

(a) Before merging: causal structure in D_{host}, or D_{twin} or D_{comp} *taken separately.* E acts as a collider and thus creates a dependency between X and T

(b) After merging: causal structure in $D_{host} \cup D_{comp}$

Figure 3.C.1: Causal structure for the illustrative experiment.

In both **ranking experiments**, the selection functions are randomly generated. We first generate a binary vector of the size of the dimension of X to define the subset of covariates that would be impact selection. We then generate two other random vectors, one for the multiplicative coefficients for each selected covariate, and another to define a power for each term in the sum. Ultimately, the probability of selection is taken as the sigmoid of this randomly generated polynomial.

In the **ranking experiment with IHDP**, 10 candidates were generated with a sample size ranging from 300 to 500. The host sample size was equal to 400. The experiment was across 20 seeds. We kept a minimum of 50 subjects in each treatment group. The hold out test dataset had a sample size of 2000. For the linear model, X, T and $X \times T$ were included. For the Gaussian Process model, a maximum of 1000 iterations was set. In CBF, both the predictive and conditional models were used with a maximum depth of 250, and a shrinkage $\alpha = 0.95$.

In the **ranking experiment with Lalonde**, 15 candidates were generated with a sample size ranging from 200 to 400. The host sample size was equal to 600. The experiment was across 20 seeds. We kept a minimum of 50 subjects in each treatment group. The hold out test dataset had a sample size of 2000. For the linear model, X, T and $X \times T$ were included. For the Gaussian Process model, a maximum of 1000 iterations was set. In CBF, both the predictive and conditional models were used with a maximum depth of 200, and a shrinkage $\alpha = 0.9$.

Datasets We describe the two datasets used in our experiments, with high-level summary given in Table 3.C.1.

Table 3.C.1: Description of the datasets: Lalonde [130] and IHDP [144].

	ihdp	**lalonde**
Number of samples	747	16,177
Number of features	24	8

The **Infant Health and Development Program, or IHDP** is a randomized controlled study designed to assess how home visits by specialist doctors impact the cognitive test scores of premature infants. Initially, the dataset serves as a benchmark for evaluating treatment effect estimation algorithms, as described in [98]. This evaluation introduces selection bias by excluding non-random subsets of treated individuals to construct an observational dataset, with outcomes derived from the original covariates and treatments.

The **Lalonde** originates from the National Supported Work Demonstration used by Dehejia and Wahba [58] to evaluate propensity score matching methods. The data consists of demographic variables (age, race, academic background, and previous real earnings), as well as a treatment indicator. The outcome is the real earnings in the year 1978.

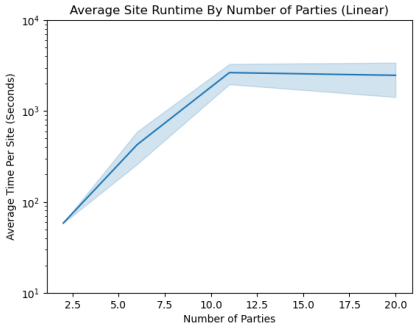
Compute times Approximate compute times for the ranking experiment on causal benchmark datasets are given in Table 3.C.2. Experiments were performed on an Apple M3 chip with a 12-core CPU and 18 GB of RAM.

Table 3.C.2: Approximate compute times.

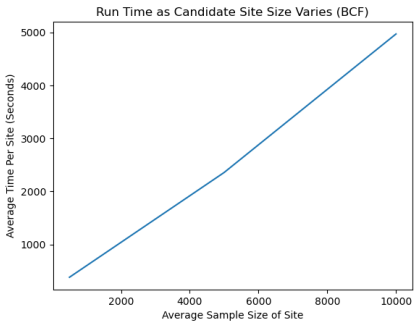
	ihdp	**lalonde**
Polynomial	$<$ 1 min	$<$ 1 min
Causal GP	3 mins	3 mins
BART	14 mins	9 mins

3.D Further experimental results

For completeness, we include the performance of all baselines on the IHDP dataset in Table 3.D.1 and the Lalonde in Table 3.D.2.



(a) In this we repeat the multi-party computation experiments from subsection 3.5.3 varying the number of sites. Each site is treated as a separate party in the multi-party computation and the average runtime per site is reported. Runtime recorded on an M3 macbook. Averaged over 5 runs.



(b) Runtime of Bayesian Causal Forest as the candidate sample size increases. We run the algorithm for the task of selecting between 5 sites with average size 500, 5000, and 10,000.

Table 3.D.1: IHDP dataset ranking experiment results with 10 candidate datasets

Model	Objective	$\boldsymbol{\rho}(\uparrow)$	**p@1** $(\uparrow)$	**p@3** $(\uparrow)$	**p@5** $(\uparrow)$
Polynomial	$\text{EIG}_{\theta_{c}\mid\mathcal{D}_0}$	$\mathbf{0.70 \pm 0.08}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.04}$	$\mathbf{0.78 \pm 0.04}$
	$\text{EIG}_{\theta\mid\mathcal{D}_0}$	$\mathbf{0.68 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.06}$	$\mathbf{0.76 \pm 0.04}$
	PropScore Error	0.40 ± 0.11	0.40 ± 0.15	0.60 ± 0.15	0.66 ± 0.04
	Sample Size	0.34 ± 0.08	0.10 ± 0.08	0.27 ± 0.04	0.50 ± 0.06
	CovDist	0.03 ± 0.07	0.10 ± 0.08	0.23 ± 0.06	0.48 ± 0.04
Causal GP	$\text{EIG}_{\tilde{\tau}(X_0)\mid\mathcal{D}_0}$	$\mathbf{0.49 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.50 \pm 0.08}$	$\mathbf{0.62 \pm 0.06}$
	$\text{EIG}_{\mathbf{f}\mid\mathcal{D}_0}$	0.33 ± 0.06	0.30 ± 0.15	0.43 ± 0.05	$\mathbf{0.60 \pm 0.04}$
	Sample Size	0.31 ± 0.12	0.10 ± 0.20	0.20 ± 0.07	0.46 ± 0.05
	PropScore Error	0.21 ± 0.09	0.10 ± 0.20	0.30 ± 0.07	0.43 ± 0.05
	CovDist	0.03 ± 0.06	0.10 ± 0.20	0.160 ± 0.04	0.46 ± 0.05
Bayesian CF	$\text{EIG}_{\theta_{c}\mid\mathcal{D}_0}$	$\mathbf{0.54 \pm 0.10}$	$\mathbf{0.60 \pm 0.15}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$
	$\text{EIG}_{\theta\mid\mathcal{D}_0}$	0.36 ± 0.10	0.30 ± 0.14	0.50 ± 0.07	$\mathbf{0.66 \pm 0.05}$
	PropScore Error	0.45 ± 0.11	$\mathbf{0.60 \pm 0.14}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$
	Sample Size	0.16 ± 0.09	0.20 ± 0.11	0.26 ± 0.06	0.52 ± 0.05
	CovDist	0.07 ± 0.09	0.00 ± 0.00	$0.26 \pm 0.06a$	0.46 ± 0.05

3.E Related work: further details

On Causal Federated Learning Federated learning is a distributed machine learning approach that enables multiple parties to collaboratively train a shared model while keeping their raw data decentralised and private. Various federated learning approaches have been proposed, including federated stochastic gradient descent [193],

Table 3.D.2: Lalonde dataset ranking experiment results with 15 candidate datasets

Model	Objective	$\boldsymbol{\rho}(\uparrow)$	**p@1** $(\uparrow)$	**p@3** $(\uparrow)$	**p@5** $(\uparrow)$
Polynomial	$\text{EIG}_{\theta_{c\|\mathcal{D}_0}}$	$\mathbf{0.47 \pm 0.05}$	$\mathbf{0.40 \pm 0.11}$	$\mathbf{0.60 \pm 0.05}$	$\mathbf{0.79 \pm 0.04}$
	$\text{EIG}_{\theta\|\mathcal{D}_0}$	$\mathbf{0.43 \pm 0.05}$	$\mathbf{0.35 \pm 0.13}$	0.48 ± 0.04	0.53 ± 0.03
	PropScore Error	0.19 ± 0.10	0.20 ± 0.17	0.32 ± 0.06	0.49 ± 0.05
	Sample Size	0.24 ± 0.10	0.25 ± 0.08	0.38 ± 0.08	0.58 ± 0.06
	CovDist	0.20 ± 0.04	0.25 ± 0.07	0.38 ± 0.06	0.48 ± 0.07
Causal GP	$\text{EIG}_{\tilde{\tau}(X_0)\|\mathcal{D}_0}$	$\mathbf{0.42 \pm 0.07}$	$\mathbf{0.5 \pm 0.12}$	$\mathbf{0.55 \pm 0.05}$	$\mathbf{0.72 \pm 0.03}$
	$\text{EIG}_{\mathbf{f}\|\mathcal{D}_0}$	$\mathbf{0.41 \pm 0.04}$	$\mathbf{0.4 \pm 0.1}$	0.43 ± 0.04	0.58 ± 0.07
	PropScore Error	0.19 ± 0.05	0.21 ± 0.15	0.32 ± 0.06	0.43 ± 0.07
	Sample Size	0.13 ± 0.07	0.15 ± 0.08	0.31 ± 0.09	0.53 ± 0.06
	CovDist	0.22 ± 0.04	0.25 ± 0.07	0.36 ± 0.08	0.48 ± 0.04
Bayesian CF	$\text{EIG}_{\theta_{c\|\mathcal{D}_0}}$	$\mathbf{0.44 \pm 0.05}$	$\mathbf{0.45 \pm 0.08}$	$\mathbf{0.55 \pm 0.05}$	$\mathbf{0.78 \pm 0.04}$
	$\text{EIG}_{\theta\|\mathcal{D}_0}$	$\mathbf{0.39 \pm 0.05}$	$\mathbf{0.45 \pm 0.06}$	0.43 ± 0.04	0.52 ± 0.03
	PropScore Error	0.22 ± 0.06	0.2 ± 0.07	0.32 ± 0.06	0.47 ± 0.07
	Sample Size	0.18 ± 0.07	0.2 ± 0.06	0.35 ± 0.09	0.41 ± 0.06
	CovDist	0.34 ± 0.03	0.3 ± 0.07	0.42 ± 0.08	0.61 ± 0.04

federated averaging [148], and more recently, methods for joint learning of deep neural network models [186, 215]. However, these algorithms do not inherently support causal inference, as the dissimilar distributions across different data sources may lead to biased causal effect estimation. To date, limited research has been conducted on the federated estimation of causal effects, highlighting the need for further exploration in this area. Due to the scope of our work, in the following paragraphs, we will focus on presenting Federated Learning methods for CATE estimation, where covariate distribution and treatment allocation are not assumed to be identical across datasets.

Several methods propose disentangling the loss function to facilitate federated learning. Vo et al. [212] propose CausalRFF, an adaptive kernel approach for causal inference that utilises Random Fourier Features to partition the loss function into multiple components, with each component corresponding to a specific data source. However, CausalRFF approach lacks strong privacy guarantees to prevent data recovery, and modelling complex non-linear relationships remains challenging [6]. Liu et al. [140] introduce a Bayesian method where parameters refer to a local disentangled loss and are updated cross-silo using server aggregation. Similarly, Vo et al. [213] divide the loss function into site-specific functions, and specify a variational posterior distribution for each local loss. Instead of tackling the loss function, Almodóvar et al. [6]) introduce a method based on disentanglement of latent factors into instrumental, confounding, and risk factors, which are then used for treatment effect estimation. They apply federated averaging on a neural network-based generative causal inference model. Ultimately,

FedCov [203] is a parametric method for federated adjustment of covariate distributions between sites, where sample weights are derived from a propensity-like model. In all the aforementioned methods, the accuracy of causal estimation is reduced due to the constraints of federated learning. Conversely, our approach does not alter the causal estimation step, thereby maintaining optimal estimation accuracy. The framework we propose focuses on federated learning of the Expected Information Gain that would be obtained by merging with a dataset. While some Federated Causal Learning methods [212, 213] provide uncertainty bounds, which could potentially be used to decide which dataset to merge with by comparing the uncertainty in these bounds, the provided bounds apply to the federated estimate and not the causal estimate potentially obtained after merging. Ultimately, none of these methods provide strong privacy guarantees, such as differential privacy (DP), which would ensure that raw data cannot be recovered from the model parametrisation or summary statistics. Moreover, all these methods use the outcome values for training their federated model, and outcome values tend to be more sensitive in nature.

On Causal Differential Privacy Contrasting with previous approaches, Niu et al. [159] introduce a meta-algorithm that adds differential privacy (DP) guarantees to various popular CATE estimation frameworks, addressing the privacy concerns mentioned earlier. However, their method relies on multiple sample splitting, where separate subsets of the data are used for estimating the propensity score and the joint response model. This approach allows for parallel composition, a property of differential privacy. In contrast, our work prioritises data efficiency, and aims to utilise the entire dataset for CATE estimation without the need for sample splitting.

On Bayesian Experimental Design Bayesian Active Learning by Disagreement (BALD) [102] is a method designed to strategically acquire training data by focusing on regions of high uncertainty. BALD introduces an acquisition function rooted in information theory, which guides the data acquisition process. When reducing entropy towards all parameters in subsection 3.3.1, we introduce a new setting for BALD where dataset are considered as data points. In the CausalBALD [108] approach, the acquisition function is altered to specifically target areas where the distributions of different treatment groups overlap, thereby maximizing sample efficiency for learning personalised treatment effects. CausalBALD is also made for the acquisition of individual data points. However, contrarily to BALD, CausalBALD's acquisition function cannot provide a scalar measure if we compute it for a dataset (i.e. a matrix $\{\mathbf{x}_i, t_i\}_{i=1}^{n_e}$) instead of data points (i.e. a vector $\mathbf{x}_i, t_i$ for a given i). To apply CausalBALD in our setting, one would need to approximate the higher-order interaction terms between all combinations of data points within each dataset, thus making the computation intractable.

4

Selection, Ignorability and Challenges with Causal Fairness

Abstract

In this paper we look at popular fairness methods that use causal counterfactuals. These methods capture the intuitive notion that a prediction is fair if it coincides with the prediction that would have been made if someone's race, gender or religion were counterfactually different. In order to achieve this, we must have causal models that are able to capture what someone would be like if we were to counterfactually change these traits. However, we argue that any model that can do this must lie outside the particularly well behaved class that is commonly considered in the fairness literature. This is because in fairness settings, models in this class entail a particularly strong causal assumption, normally only seen in a randomised controlled trial. We argue that in general this is unlikely to hold. Furthermore, we show in many cases it can be explicitly rejected due to the fact that samples are selected from a wider population. We show this creates difficulties for counterfactual fairness as well as for the application of more general causal fairness methods.

4.1 Introduction

Recently there has been a large body of work on the problem of fair machine learning. This has stemmed from concerns that training data often contains human and societal biases that can be replicated by machine learning models, causing unfair treatment to certain groups on the basis of protected attributes such as race, gender and disabilities. This has given rise to a large variety of statistical fairness definitions such as demographic parity [79], equality of opportunity [93], fairness through awareness [64] and many more [208]. Following this there have been many different approaches to achieve these definitions, such as variational inference [143], adversarial learning [228] and optimal transport [43].

Following results showing many statistical fairness definitions are mutually incompatible [120, 167], and so can not be simultaneously satisfied apart from in trivial scenarios, new definitions were proposed based on causality [117, 229, 128, 158, 41]. This work argues that causal definitions using interventions and counterfactuals capture a more intuitive and correct understanding of what it means for an algorithm to be fair, and that only by understanding the causal relationships in our data can we hope to satisfy fairness [42, 142].

In this paper we focus on the most popular causal fairness definitions, which use causal counterfactuals [128, 158, 41]. Causal counterfactuals aim to answer questions of the form "what would have happened to Y had X been different, given we hold anything that doesn't depend on X constant?". In fairness settings the counterfactuals are based on what would have happened had the value of a sensitive attribute been different, given we hold all other background conditions constant. Counterfactual Fairness [128] says our predictions are fair for an individual if they align with those in a counterfactual world in which their sensitive attribute had been different. For example, a prediction of the probability that a woman defaults on her loan is fair if it coincides with the prediction they would receive if they had they been counterfactually born a man, given everything else is held constant.

To achieve this requires a causal model to be fitted to data. This model allows us to compute approximate counterfactuals which our model is fair in relation to. Therefore, it is critically important that the class of causal models we search over contains at least one model with the correct counterfactuals. The most common class in causal literature is the class of models with independent noise. Informally, this assumes that our factual data and the counterfactuals can be described by a set of deterministic equations with the addition of random noise that is not correlated with anything else.

In this paper we challenge this in a fairness context. Our argument rests on the fact that if you assume this, the approximate counterfactuals generated by these models have properties you would only expect to see in a randomised controlled trial, and so seem implausible. Moreover, data in fairness problems is usually selected in some

way. So, we show that if an independent model could fit the general population the faithfulness assumption inherent to graphical models suggests that none could fit the selected population. Hence, the noise variables effectively must be dependent.

We argue this creates problems for achieving counterfactual fairness and for causal fairness more generally. This is because correctly fitting models with dependent noise is considerably more challenging as we do not know the correct nature of the dependency and cannot tell without more data. It also means that the modelling assumptions required for many methods from the field of causality do not hold. We give an explicit example of this in the case of path-specifc fairness.

4.1.1 Paper Outline

In Section 4.3 we introduce the ignorabillity assumption which is key to our argument. We discuss it's relevance to a randomised controlled trial, how it arises from common modelling assumptions in the fairness literature and why it is unlikely to hold in practice. Following this in Section 4.4 we lay out an explicit causal argument against ignorability. This leads to conditions for when we can assume an independent noise model and a constraint which can show no independent noise model fits. Finally in Section 4.5 we discuss the difficulties this raises for causal fairness.

4.2 Preliminaries

4.2.1 Notation and Definitions

4.2.1.1 Causal Definitions

Following Pearl [162] and Peters et al. [164] a *Structural Causal Model* (SCM) $\mathcal{M} = \langle U, V, F, P(U) \rangle$ consists of:

- U, a set of *noise variables* or latent background variables; these are factors not caused by any variable in the set V of *observable variables*.
- F, a set of *structural equations* $\{f_1, \ldots, f_n\}$, one for each $V_i \in V$, such that $V_j = f_j(pa_j, U_j)$, $pa_j \subseteq V \setminus \{V_j\}$ and $U_j \subseteq U$ where pa_j is notation for the parents of V_i. This notation comes from the fact that the model gives rise to a causal graph which we assume to be a DAG.
- A *probability distribution* $P(U)$ over the latent variables U.

We may model the distribution of a set Z following an intervention on a subset of the other variables $W \subseteq V \setminus Z$, by replacing the structural equation for each $W_i \in W$ by the fixed value $W_i = w_i$. We use the potential outcome notation, so $Z(w)$ is a random variable that has distribution of Z after we have intervened to set $W = w$. Further the SCM allows for the computation of *structural counterfactuals*. That is, for an

individual with background variables $U = u$, the structural counterfactual for Z given $W = w$ is denoted by $Z(w, u)$ and is the unique solution for Z given $U = u$ and by replacing the equations for W with the fixed value $W = w$. We often omit the u when it is clear from context and just write $Z(w)$.

Given our probability distribution, $P(u)$, we can infer the distributions over our structural counterfactuals given evidence. That is, we can compute $P(Z(w) = z \mid E = e)$ by finding the posterior distribution for U given $E = e$, substituting $W = w$ in our equations, and then using our posterior distribution $P(U \mid E = e)$ to give the probability in question. The evidence, E, could be something counterfactual; for example, we might have observed $Z = z', W = w'$ and then want to compute the probability that $Z = z$ in a counterfactual world in which W is fixed to the value w.

We refer to these as structural counterfactuals in order to emphasise that they are counterfactuals from a structural causal model. This does not mean they are truly counterfactuals and generally we can only give them this interpretation when the causal model is suitably motivated by our beliefs about the true state of the world. Whenever we make the assumption that there is a true causal model that generates our data we will denote it by $\mathcal{M}^*$. Whenever a true causal model $\mathcal{M}^*$ is assumed, we will use $V^*(a)$ to denote the counterfactual of V according to this causal model.

The key focus of this paper is whether or not it is valid to assume that the noise variables, U, are jointly independent in fairness settings. This assumption is commonplace in the wider causal literature [164] and many key results rely on it, the most obvious being that d-separation implies conditional independence.

4.2.1.2 Counterfactual Fairness

Now we introduce counterfactual fairness [128]. First in the general fairness setup we suppose we have access to a dataset $\Delta = \{a^n, x^n, y^n\}_{n=1}^N$ of individuals where a^n indicates the *sensitive attributes*, x^n is a list of covariates and y^n is some outcome of interest we wish to predict. We want to form a predictor $\widehat{Y}$ which is not discriminatory on the basis of our sensitive attributes, this is known as being fair'. In order to do this we need some definition of fairness. For counterfactual fairness we assume that there is some true causal model of the world $\mathcal{M}^*$ and that relative to this model a predictor $\widehat{Y}$ is *counterfactually fair* if for all contexts $X = x, A = a$ we have

$$P\left(\widehat{Y}(a) = y \mid X = x, A = a\right) = P\left(\widehat{Y}(a') = y \mid X = x, A = a\right) \tag{4.1}$$

for all values y and a'.

In order to achieve counterfactual fairness, we fit some causal model $\mathcal{M}$ aiming to approximate $\mathcal{M}^*$. We then either use noise variables arising from $\mathcal{M}$ and covariates that are not causally dependent on A as inputs to $\widehat{Y}$, or use $\mathcal{M}$ to generate structural counterfactuals and use regularisation to enforce that the predictor outputs the same

value on the potential outcomes as in their observed values [185]. Kusner et al. [128] emphasise that the causal model $\mathcal{M}$ we fit must be suitably causally motivated for us to expect any predictor formed in this way to be counterfactually fair, relative to the true causal model $\mathcal{M}^*$. Therefore we would hope there exists some causal model in the space we search over that has the correct counterfactuals. That is, the structural counterfactuals align with these 'true counterfactuals' from $\mathcal{M}^*$.

4.2.2 Introducing the Law School Example

We use one of the main examples from Kusner et al. [128] throughout to explain our ideas. They aim to form a predictor for US law school admissions, which should be fair with respect to the sensitive attributes race and sex. To train the predictor we have data from people who previously attended law school. We have their college GPA and LSAT scores, as well as their sensitive attributes race and sex. From here we aim to impute their first year average grade (FYA). We would then use this to predict if a new applicant would succeed at law school and so if they should be admitted or not. Kusner et al. [128] assume the observed variables follow the causal DAG in Figure 4.2.1; whilst they use different causal models to estimate the noise variables, this structure over the observed variables remains constant.

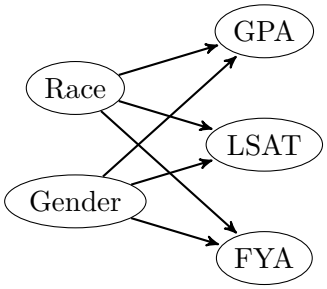
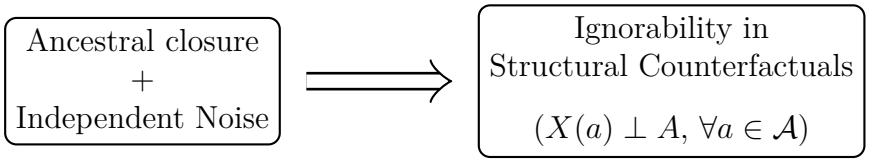


Figure 4.2.1: A causal DAG for the Law School example

4.2.3 Ancestral Closure of the Sensitive Attributes

When drawing these causal graphs to describe our data, a common assumption is that the set of sensitive attributes is *ancestrally closed*. By this we mean it has no observable cause or unobserved cause that is shared with another variable. This makes sense in many scenarios; for example, nothing could be said to cause someone's gender or many disabilities. In the case of counterfactual fairness ancestral closure is mentioned as an explicit requirement on the set of sensitive attributes [128]. This is because otherwise we could discriminate on the basis of things that cause our sensitive attribute. Kusner et al. [128] give the example that we could discriminate on the basis of mother's race if only race was sensitive and we did not enforce ancestral closure. Therefore given the fact that ancestral closure can be seen as a requirement and is also assumed in most DAGs we can find in the literature [128, 185, 118, 158, 41], we make this assumption throughout.



4.3 Ignorability

As stated, our main focus will be on whether it is reasonable to assume that the noise variables, U, are mutually independent in fairness settings. The difficulty comes from the fact that independent noise and ancestral closure together imply that our structural counterfactuals satisfy *ignorability*[1]. This assumption is more commonly seen in the context of randomised controlled trials and it is as follows:

$$X(a) \perp A, \qquad \forall a \in \mathcal{A}. \tag{4.2}$$

This implies that for all a, a':

$$(X \mid \{A = a\}) \stackrel{d}{=} (X(a) \mid \{A = a\}) \stackrel{d}{=} (X(a) \mid \{A = a'\}). \tag{4.3}$$

Imagined in a randomised control trial where A is now our treatment and X is the measured covariates, this says if we want to know how the untreated group would look had we counterfactually given them the treatment then we only need to look at what happened in the treated group. That is, due to the randomisation of the treatment, those individuals in the untreated group would (on average) look like individuals in treated group, had we counterfactually chosen to treat them instead. This is what allows us to estimate the effect of a treatment in a randomised controlled trial by the difference in means between the treated and untreated groups. Ignorability is therefore a very strong assumption in general, the fact that know we it is satisfied in randomised controlled trials is what makes them the 'gold standard' of causal inference.

In fairness settings where there is no such motivation from randomisation of 'treatment', ignorability seems like a much stronger assumption. In the context of the law school example with counterfactuals relating to sex, this would mean if the group of males that applied were counterfactually born female, they would look like the group of applying females.

This implies that for every male that applied, if they had counterfactually been born female then they would have attended college to get a GPA, taken the LSAT, and that their GPA and LSAT grades would be indistinguishable from the grades attained in the real world by the women who applied to law school.

It seems natural to be concerned that this will not hold. We might feel our society unfairly pushes women away from considering a career in law, and so believe if some men

[1] Also known as exchangeability.

who applied for law school had been born female instead they would have been deterred from taking the LSAT, for example. If we imagined the same scenario in the 1950s, when there was poor access to higher education for many women, we would almost certainly expect that most men who attend college would not have attended college if they had been born female and therefore would not have a GPA. If our structural counterfactuals are not at all similar to what we would expect a true counterfactual to be like in this extreme case, it is hard to see how can we have confidence that they resemble true counterfactuals when applied to other fairness scenarios.

Furthermore, we can imagine scenarios when treating counterfactuals like this could be intuitively very unfair. As an example, again we look at law school admissions, but now our sensitive attribute A is the presence of a particular disability with severe adverse effects; for instance, it might mean that, on average, sufferers can only work or study for half the amount of time per day than someone who does not have this disability. If an individual were able to attend college to get a GPA, take the LSAT, and perform well enough in both of these to apply to law school *despite* having this disability, it is reasonable to assume that they are an exceptional candidate and would have performed exceptionally well relative to all candidates had they been born without any disability. However, their structural counterfactuals formed as above would look like an average applicant born without the disability. Therefore a predictor which is counterfactually fair relative to these structural counterfactuals would simply treat this candidate as an average applicant without the disability. This does not capture an intuitive notion of fairness in this scenario. Further it does not align with what we imagine counterfactual fairness as doing. The structural counterfactuals are failing to correcting for the difficulties of having this sensitive attribute, which is one of the main appeals and claims of counterfactual fairness.

We note that almost all models we found in the literature on fairness using causal counterfactuals assumes a causal model that is both ancestrally closed and has independent noise variables, either explicitly [128, 185, 118] or implicitly for identification results [158, 41]. We now give more detailed analysis of when this may seem to be a reasonable or unreasonable assumption.

4.4 Selection

4.4.1 Theoretical Results

In order to formalise the issues raised in the previous section we cast it as a problem due to selection from a wider population. Key to our analysis is that in this population we are comfortable with the assumption that the 'true' counterfactuals satisfy ignorability. Therefore we focus our analysis on birth sex and take the wider population to be the general population of, for example, a country. Now ignorability does not seem such a strong assumption as birth sex is random and there is no selection whatsoever, so

we can loosely imagine this as a large randomised trial[2]. This allows us to make the following assumption about the nature of the data generating process:

Assumption 4. *There is some true causal model $\mathcal{M}^*$ that generates our covariates, X, for the entire population. In $\mathcal{M}^*$ the noise variables are independent and in the causal DAG following from $\mathcal{M}^*$ the sensitive attribute set is ancestrally closed. Further, we allow the domain of X to be expanded so that we write $X_j = \emptyset$ if an individual does not possess the jth covariate.*

We denote the structural counterfactuals arising from $\mathcal{M}^*$ by $X^*(a)$ and call these the *true counterfactuals*. Due to the form of $\mathcal{M}^*$ the counterfactuals will satisfy $X^*(a) \perp A$; however as we stated, this is a more comfortable assumption in the general population. It is important to note we do not consider race here, as the assumption of ignorability or behaving like a randomised control trial seems much more unreasonable. We discuss this further in Appendix 4.A. However it should be said that if whenever it is assumed that an SCM with independent noise in which race is ancestrally closed can fit the counterfactuals we make the same assumption of ignorability and this is still unlikely to hold.

Continuing with the example of the law school predictor our covariates X are GPA and LSAT and we use GPA $= \emptyset$ to indicate when an individual in the general population has not completed college and so lacks a GPA. As above we focus our analysis on birth sex, and assume A can only take two values a, a'; we do this because it is consistent with the measurements used in many of the datasets we will look at.

We use a binary variable S to indicate if an individual lies in the dataset we have access to. For example, in the law school example $S = 1$ if an individual applied to law school. Now to try to achieve counterfactual fairness with our dataset we would be fitting a causal model $\mathcal{M}$ on those with $S = 1$. As discussed in Section 4.3 many models in the causal fairness literature fall into the following class:

Definition 16. *Let $\mathbb{M}_{S=1}$ be the set of causal models that fit the data, in which all noise variables are jointly independent and further, give rise to a DAG in which the set of sensitive attributes is ancestrally closed.*

The question is: when does there exist a model $\mathcal{M} \in \mathbb{M}_{S=1}$ such that the structural counterfactuals of $\mathcal{M}$ align with the true counterfactuals from $\mathcal{M}^*$? The following gives a characterisation of this:

Proposition 21. *There exists an $\mathcal{M} \in \mathbb{M}_{S=1}$ such that the structural counterfactuals from $\mathcal{M}$ align with the true counterfactuals from $\mathcal{M}^*$ if and only if we have:*

$$X^*(a) \perp A \mid S = 1 \ , \ \forall a \in A. \tag{4.4}$$

[2]We note that even still this is an approximation and at most there would be near ignorability. We take this assumption exactly simply to allow us to make some formal analysis. However rejecting this assumption in general supports our argument as it shows in no population should a practitioner be happy with ignorability.

That is, if we satisfy *ignorability under selection*. Furthermore, we can state this in terms of a constraint on selection which resembles a scaled version of counterfactual fairness:

Proposition 22. *We have ignorability under selection if and only if selection satisfies the following:*

$$\frac{P(S(a) = 1 \mid X = x, A = a)}{P(S = 1 \mid A = a)} = \frac{P(S(a') = 1 \mid X = x, A = a)}{P(S = 1 \mid A = a')}$$

for all a, a' *and* x *such that* $P(X = x \mid A = a) > 0$.

Therefore if either of these conditions are violated we should not expect any model in this class to capture the correct counterfactuals. This is clearly a problem if the justification of our causal fairness method relies on it being able to—in principal—capture the correct counterfactuals.

4.4.2 When will Ignorability Under Selection Hold?

We now look further at when we can expect ignorability under selection to hold and therefore when we can fit or assume a model in $\mathbb{M}_{S=1}$. In order to do so we first introduce the definition of faithfulness:

Definition 17. *A distribution* $P(V)$ *is* faithful *with respect to some graph* $\mathcal{G}$ *if whenever we have* $A \perp B \mid C$ *for sets of variables* A, B, C *then* A *is d-separated from* B *by* C *in* $\mathcal{G}$*. That is any conditional independences are implied by d-separation.*

Faithfulness is the converse to the statement that d-separation implies conditional independence and is commonplace in the causal inference literature. A violation of faithfulness entails an exact balancing out of causal effects so that they leave no probabilistic trace and the assumption is often justified by arguing that this is unlikely to occur in practice. Moreover, there are theoretical results showing that for certain families of distributions such as discrete or Gaussian, violations will occur on a set of measure zero with respect to any continuous measures over parameters [149]. However many argue that this should not be taken as a blanket assumption and that sometimes particular causal effects can occur precisely to balance out other ones, such as in for example biological systems or policy decisions [101, 8].

Faithfulness is relevant in as if we consider the very general graphical model in Figure 4.4.1 to describe our scenario we can see by conditioning on S we open up the paths $A \to X \leftarrow X^*(a)$ and $A \to S \leftarrow X \leftarrow X^*(a)$. Therefore if these paths are present and faithfulness holds we will have $X^*(a) \not\perp A \mid S = 1$ and so the counterfactuals cannot be captured by a model in $\mathbb{M}_{S=1}$[3]. Therefore if it is assumed that the underlying data

[3]Technically we are referring to faithfulness in a twin network which does not hold in general due to the deterministic relations implied by consistency. This is discussed more in [196]. However as we discuss in appendix 4.D this does not create an issue for this graph.

generating mechanism follows a model in $\mathbb{M}_{S=1}$, either for particular properties of causal DAGs or to approximate counterfactuals, justification should be given for one of the following, ordered by the strength of the assumption:

1. There is no selection from the general population; a census would satisfy this condition, for example.
2. There is selection from the general population, but we are without a direct path from either A or $X^*(a)$ to S. This could be the case if we randomly sample individuals from the general population.
3. The selection is such that a direct paths from A and $X^*(a)$ to S are present. However, this dependence occurs in such a way as to violate faithfulness. In practice it is hard to see how to make a clear argument for this in fairness contexts as it would amount to saying that selection occurs in some suitable 'fair' way given by Proposition 22.

Now applying this to the running example of law school prediction we can see the law school applicants are a subset of the general population, so we violate 1. Further this selection is not random, and in general the likelihood of application would depend on GPA and LSAT so we also violate 2. Therefore in order to suppose a model in $\mathbb{M}_{S=1}$ that can fit the counterfactuals we have to make an argument for 3, but there is no obvious reason to suggest the data distribution would violate faithfulness. Therefore, there is no reason to believe structural counterfactuals from any model in $\mathbb{M}_{S=1}$ could capture the counterfactuals. Thus, by using these steps we have given a clear causal argument supporting the concerns we raised about the counterfactuals for the law school example in Section 4.3.

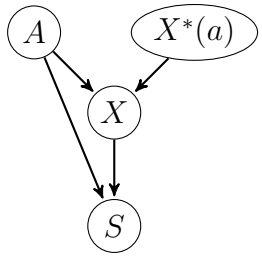


Figure 4.4.1: A causal DAG for selection

These steps can be applied to any dataset and if we plan on fitting a model in $\mathbb{M}_{S=1}$, justification for at least one of these points should be given. Unfortunately, in many fairness settings it is hard to imagine being able to argue for any of these, and so we run into the difficulties discussed at the end of the previous section. That is, we cannot generate both our dataset and the counterfactuals by models usually assumed in causal inference. Further, fitting any model to approximate the counterfactuals becomes significantly harder in practice.

4.4.3 Explicit Violation in certain cases

The scaled counterfactual fairness condition leads to a constraint that can be tested on a dataset to explicitly verify that no model in $\mathbb{M}_{S=1}$ can correctly capture the counterfactuals. It is worth noting that if this constraint is not clearly violated, that is

Table 4.4.1: Constraint for popular causal fairness datasets

Does there exist an x such that	Adult	Law school	German Credit
$P(S = 1 \mid X = x, A = \text{Female}) >$	0.475	0.753	0.421

not good evidence that a model in $\mathbb{M}_{S=1}$ does fit the counterfactuals, and instead the list of conditions for ignorability under selection in Section 4.4.2 should be referred to. The aim is instead to show beyond doubt that no such model fits. The constraint is as follows:

Corollary 23. *If there exist x, a, a' such that $P(X = x \mid A = a) > 0$ with*

$$P(S = 1 \mid X = x, A = a) > \frac{P(A = a \mid S = 1)P(A = a')}{P(A = a' \mid S = 1)P(A = a)}$$

then there exists no model in $\mathbb{M}_{S=1}$ that has the correct counterfactuals.

Note this places no restriction on the form of $\mathcal{M}^*$ apart from Assumption 4.

We now apply this method to some datasets used in the causal fairness literature with sex as our sensitive attribute. The results are shown in Table 4.4.1. The bound is computed using census data to estimate the probability of individuals having a given sex in the general population.

We use the Adult dataset as an example to demonstrate the usefulness of this result; the observations in this dataset are taken from census data with certain deterministic constraints on the covariates. For example, the total number of hours worked has to be positive and the yearly earnings must be more than 100 dollars. Anyone satisfying these who is part of the census is guaranteed to get selected, and so for women with x satisfying this $P(S = 1 \mid X = x, A = \text{Female}) = 1 > 0.477$. Therefore this violates the constraint and so there is no causal model in $\mathbb{M}_{S=1}$ that can accurately capture the true counterfactuals, regardless of the true causal model $\mathcal{M}^*$ that describes the world. A similar argument to the above can be applied whenever there is a deterministic rule for selection from the general population. Namely if $P(S = 1 \mid A = a) < P(S = 1 \mid A = a')$ and we can find a set of values for the covariates $X = x$ such that $P(S = 1 \mid X = x, A = a) = 1$, then the constraint is violated.

4.5 Challenges for Causal Fairness

In this section we discuss the challenges created when no model in $\mathbb{M}_{S=1}$ fits. We first look at the issues this causes to the general application of causal fairness methods and then to specific difficulties this creates for counterfactual fairness and path-specific counterfactual fairness.

4.5.1 Difficulties when no model in $\mathbb{M}_{S=1}$ fits

If no model in $M_{S=1}$ fits this does not mean that there exists no causal model which captures the true counterfactuals. However it does mean that in any correct causal model, the distribution of the noise variables will depend on A. This creates the following two challenges for the application of causal fairness methods.

Firstly the models that lie outside of $\mathbb{M}_{S=1}$ are not well behaved enough to guarantee many properties and identification results that are normally assumed in causal inference. The most obvious is that d-separation in the graph will not generally imply a conditional independence in the distribution. This is a problem as many key results in causality rely on d-separation, for example identification results, the do-calculus, and the adjustment criteria. Therefore if we find no model in $\mathbb{M}_{S=1}$ fits we should be cautious about applying results from the wider causal inference literature in fairness problems without clearly justifying that we still satisfy the required assumptions.

Secondly this makes the identification of counterfactuals strictly harder as we lose any way to connect them to real world observed variables. Under the assumption that some model in $\mathbb{M}_{S=1}$ fits we could identify the distribution of the group level counterfactuals without knowing the true model. That is, we know $P(X^*(a) \mid A = a', S = 1) = P(X \mid A = a, S = 1)$ using the assumption of ignorability under selection. Therefore we only need a way to find $P(X^*(a) \mid A = a, X = x', S = 1)$ in order to satisfy individual level counterfactual fairness. However if no model in $\mathbb{M}_{S=1}$ fits we are now in a strictly more challenging setting, where we cannot even identify the distribution of the group level counterfactuals. This is because the noise variables in our model are dependent on our sensitive attribute and the structure of the dependency is not clear without further assumptions or data. In the selection context this corresponds to the fact the distribution of those noise variable in the whole population is not identifiable from the data [18].

This creates problems for fairness based on causal counterfactuals as we now need to introduce some dependencies between the sensitive attributes and any noise variables we include in our model. However this cannot be done arbitrarily as it is unclear how a particular dependency would affect the accuracy of your counterfactuals and therefore the fairness of the model. Thus, making any arbitrary changes to the model without any clear idea of its affect on the fairness would be high contentious. Furthermore making principled changes would require more assumptions, data or both.

4.5.2 Do Structural Counterfactuals from models in $\mathbb{M}_{S=1}$ have a causal interpretation?

We now ask, can we give the structural counterfactuals from a model in $\mathbb{M}_{S=1}$ a separate causal interpretation? Maybe as some other kind of counterfactual? We will argue not, and look at two possible interpretations. In the first the structural

counterfactuals represent the counterfactual given that in the world where an individual is born with a different attribute they would have made it into the selected set. In the second it captures how an individual would appear at the time of selection if they counterfactually had a different sensitive attribute.

The first is in general difficult as we cannot identify who, if anyone, would have been selected in the real world and the counterfactual one in which they were born with a different sensitive attribute. This is related to work on causal inference in the presence of competing effects. Stensrud et al. [202] explain competing events using the example of a 3 year medical trial and note that people in both the treatment and control group may die of other related effects before the trial is completed. As a result it is hard to come up with an appropriate counterfactual contrast for treatment effects as we cannot pinpoint who, if anyone, would have survived if they were in both the treatment and the control arm. Therefore the identification and definition of any counterfactual contrasts relies on strong untestable assumptions about a group of people who survive in both arms. In the same way it is hard to come up with the correct counterfactual contrast for fairness here as we cannot tell who, if anyone, would have made it to our selected set regardless of the value of their sensitive attribute at birth. This makes it challenging to take this interpretation and further to asses if we have achieved it.

In the second case, if we are trying to look at how an individual would appear if at the time of selection they had had a different sensitive attribute. The difficulty here is whether this is what we are aiming for: why should we propagate a causal effect through covariates that occur pre-selection? Again using the law school example, we could imagine saying we want our counterfactually fair predictor to align with the one in which an individual had a different sex in the moment of application. This seems to align with the intuition our predictor is fair if a female applicant will get in if the same applicant would be accepted if they were male. However if this is our aim, then it makes little sense to propagate causal effects through GPA and LSAT, as these occur before application. Instead the correct counterfactual would be obtained by simply flipping the value of our sensitive attribute, as Wachter et al. [214] advocate for.

Therefore we argue that if ignorability is violated in our dataset violated then it is hard to believe that the structural counterfactuals from any model in $\mathbb{M}_{S=1}$ can be given a correct causal interpretation. This violation could either be due to a general rejection of ignorability (as in the case of race) or via the selection arguments in Section 4.4.

4.5.3 Counterfactual Fairness and Demographic Parity

In this section we consider what fairness guarantees are obtained when we aim to achieve counterfactual fairness with a model in $\mathbb{M}_{S=1}$ when none fits. In the previous section we argue that this will give no causal guarantees. Therefore the only fairness properties will be statistical. To this avail we provide the following:

Proposition 24. *Let $\widehat{Y}$ be a predictor that is counterfactually fair according to some causal model $\mathcal{M} \in \mathbb{M}_{S=1}$. Then $\widehat{Y}$ satisfies demographic parity on the data it fits. That is $\widehat{Y} \perp A \mid S = 1$.*

This also leads to the following corollary which gives a more general relationship between counterfactual fairness and demographic parity when we make Assumption 4.
Corollary 25. *Let $\mathcal{F}$ be the set of predictors, possibly with random noise, which are counterfactually fair according to the true model $\mathcal{M}^*$ under Assumption 4. Then we have that all $\widehat{Y} \in \mathcal{F}$ will satisfy demographic parity if and only if we have ignorability under selection.*

In other words, if we do not have the conditions for ignorability under selection then truly counterfactually fair predictors need not necessarily satisfy demographic parity.

Therefore we argue that in many fairness settings, applying a model from $\mathbb{M}_{S=1}$ with the hope of achieving counterfactual fairness will only give us demographic parity, with no extra causal interpretation.

4.5.4 Path Specific Fairness

Recently, new causal fairness variants have been proposed that rely on the use of path specific effects [158, 41]. These involve labelling pathways from our sensitive attribute to outcome of interest as fair or unfair, and controlling for the effect along the unfair pathways in our predictor. We give a brief example of these methods using the classic UC Berkeley Gender discrimination case. In this dataset we find that there is a correlation between acceptance rate to Berkeley and Gender, which seems unfair. However after stratifying by department this correlation disappears. Therefore the argument is the original correlation is due to the fact that the women who apply, apply for more competitive departments.

In Chiappa and Isaac [42] they represent this by the causal DAG in Figure 4.5.1 with covariates gender (A), department (D), qualifications (Q) and acceptance to Berkeley (Y). They argue that gender has two potential causal effects on outcome, a direct effect and an effect through department. They label the direct effect as unfair and the indirect effect as fair. Now according to path specific fairness the outcome is unfair if it takes in any causal effect in through unfair pathways, therefore by the same argument we have that a fair predictor will only take information about the sensitive attribute through fair pathways.

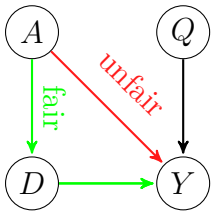


Figure 4.5.1: Causal DAG for the Berkeley Gender Discrimination

4.5.4.1 Identifiability of Path Specific Effects

In order to be able to apply these path specific fairness methods we need to be able to identify the underlying path specific effects. In the Berkeley data, for example, we need to be able to identify the distribution of $Y(D(a), a')$, denoting the effect when A is set to a along the pathway leading to D and a' on the direct pathway to Y. In general results for the identification of path specific effects rely on the correct causal model having independent noise. Without this assumption we do not necessarily have identifiability of path specific effects [194]. However if we assume an ancestrally closed sensitive attribute set, this is just $\mathbb{M}_{S=1}$. Therefore, the conditions in Section 4.4.2 can also allow us to reason about when we can apply path specific fairness variants with such DAGs. Furthermore we can apply the results in Section 4.4.3 to the experiments on the Adult dataset in both Chiappa [41] and Nabi and Shpitser [158]. As the constraint in Corollary 23 is violated, we do not have the necessary conditions to guarantee identifiability of the path specific effects these experiments rely on.

4.6 Conclusion

In this paper, we have argued that more care should be taken as to what type of SCM is assumed in order to achieve fairness through the use of structural counterfactuals. We show that often any SCM with independent noise cannot capture the correct counterfactuals and further that the structural counterfactuals from these models cannot be given a clear causal interpretation. Finally we have show this creates issues for a variety of causal fairness methods due to model fitting and an inability to apply results from the wider causal inference literature.

4.7 Acknowledgments

We would like to Siu Lun Chau, Jean-Francois Ton and the reviewers for their helpful comments that have greatly improved this paper. We also thank Thomas Richardson for his comments on faithfulness within twin networks. Jake Fawkes gratefully acknowledges funding from the EPSRC.

Appendix

4.A Race and Ignorability

The analysis in this paper using selection did not include race and this is because we do not feel it correct to make the assumption of ignorability at any stage when race is our sensitive attribute. Therefore Assumption 4 would be misguided.

This is because race is correlated with many covariates and it is unclear in general what we are trying to counterfactually correct for. This point relates more to recent philosophical work on counterfactual fairness [115, 104, 125] as well as work on race in causal studies [190]. This is beyond the scope of this paper as we have raised challenges for counterfactual fairness methods from within the causal framework, as opposed to the work referenced which raises problems with the use of the causal framework in this setting. However we give a brief example to highlight why we did not include race.

In America and many other Western nations, the race a child is born to is correlated with many crucial demographic features; these include the level of education in their family, socioeconomic status and the neighbourhood of their birth. This makes any comparisons to randomised control trials seem far fetched as these features are likely to affect almost all outcomes we measure in later life. Should our counterfactuals correct for this or not? This relates to what philosophical definition one uses of race, the point made in 'Race as a bundle of sticks' [190]. The perspective of a racial constructivist, the most popular view in the social sciences, says that racial categories are not a biological fact but they are a social reality and they are inextricably tied with historic 'differences in resources, opportunities, and well-being' [227]. Therefore maybe our counterfactuals should correct for this. However taking this point of view it is not clear how we should interpret any counterfactual or if the causal framework fairness for race makes sense as Kohler-Hausmann [125], Hu and Kohler-Hausmann [104] and Kasirzadeh and Smart [115] point out.

If one instead takes an essentialist point of view (this is largely unpopular in the social sciences but it is often implicitly assumed in causal studies) then potentially not, but then we clearly will not satisfy ignorability as we have features that are not caused by race at birth but are correlated with it. [128] mention possibly including as parents' race as a causal ancestor of race in our DAG and also having this as a

protected attribute. However, we again run into the same problems as parents' race will be correlated with the same features but one generation back. Therefore when, if ever, would we be happy to say our data could be described by a causal graph with ancestrally closed sensitive attribute set and independent noise variables? This tracking of features back generations in an effort to counterfactually correct for them once again relates more to racial constructivist perspectives, since we are struggling to separate race as something to correct for the discriminatory effects for from the historic context that created it.

We also note these perspectives also apply to gender and other sensitive attributes. However as we mention at the start of the appendix we have raised challenges for counterfactual fairness methods from within the causal framework as opposed to potential problems with the framework.

4.B Proof of Proposition 21

Proof. First for any model in $\mathbb{M}_{S=1}$ due to the ancestral closure and the fact $U \perp A$ we must have for all the potential outcomes $X(a)$ that is generates:

$$X(a) \perp A \ , \forall a \in A$$

Hence if we have for some a, $X^*(a) \not\perp A \mid S = 1$ then no model in $\mathbb{M}_{S=1}$ can generate these $X^*(a)$.

Now if we have $X^*(a) \perp A \mid S = 1 \ , \forall a \in A$ then we construct a causal model $\bar{\mathcal{M}} \in \mathbb{M}_{S=1}$ with the correct counterfactuals by taking our U to be $\{X^*(a), \forall a \in A\}$ and simply $X = \sum \mathbb{I}(A = a)X(a)$. This clearly generates our data for $S = 1$, we have $U \perp A$ as $X(a) \perp A \mid S = 1 \ , \forall a \in A$. Finally this causal model trivially has the same counterfactuals as $\mathcal{M}^*$ □

4.C Proof of Proposition 22

Proof. First we note the independence $X^*(a) \perp A \mid S = 1$ is equivalent to saying for all a, a' with $P(S = 1 \mid A = a) > 0$ and $(S = 1 \mid A = a') > 0$:

$$P(X^*(a) = x \mid A = a, S = 1) = P(X^*(a) = x \mid A = a', S = 1).$$

Applying Bayes rule gives:

$$\frac{P(S = 1 \mid X^*(a) = x, A = a)P(X^*(a) = x \mid A = a)}{P(S = 1 \mid A = a)} =$$
$$\frac{P(S = 1 \mid X^*(a) = x, A = a')P(X^*(a) = x \mid A = a')}{P(S = 1 \mid A = a')}.$$

Now we use the fact that $X^*(a) \perp A$ in the population, so we have $P(X^*(a) = x \mid A = a) = P(X^*(a) = x \mid A = a')$; hence since we have $P(X^*(a) = x) = P(X = x \mid A = a) > 0$ we can cancel these to give:

$$\frac{P(S = 1 \mid X^*(a) = x, A = a)}{P(S = 1 \mid A = a)} = \frac{P(S = 1 \mid X^*(a) = x, A = a')}{P(S = 1 \mid A = a')}.$$

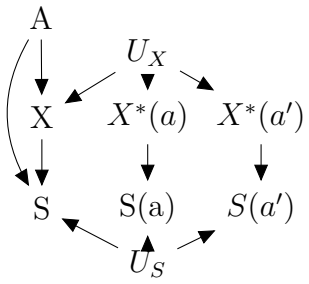
All that remains to show is that:

$$P(S = 1 \mid X^*(a) = x, A = a') = P(S(a') = 1 \mid X = x, A = a).$$

We have:

$$\begin{aligned} P(S = 1 \mid X^*(a) = x, A = a') &= P(S(a') = 1 \mid X^*(a) = x, A = a') && (4.5) \\ &= P(S(a') = 1 \mid X^*(a) = x, A = a) && (4.6) \\ &= P(S(a') = 1 \mid X = x, A = a), && (4.7) \end{aligned}$$

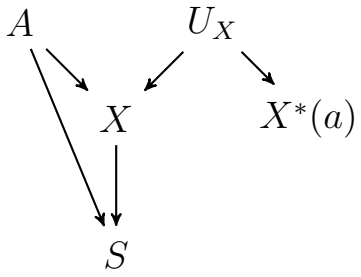
where (4.5) and (4.7) follow from the consistency property, and (4.6) follows from the fact that $S(a') \perp A \mid X^*(a)$. This can be read off the following 'triplet' network which is Markovian under our Assumption 4.



Hence under assumption one the required equation is equivalent to ignorability under selection. □

4.D Faithfulness in the Twin Network

We note that when talking about the independence $X^*(a) \perp A \mid S = 1$ in the graph in figure 4.4.1 we are actually referring to a conditional independence in the following twin network:



However a potential problem with this is that d-separation in twin-networks is not complete. Therefore it is possible to have an open path between two vertices with no models in the class able to give the corresponding dependence. This problem arises due to the deterministic relationships within twin networks.

This concern is relevant as it could be possible that there is no model in the class with $X^*(a) \not\perp A \mid S = 1$ despite the fact that conditioning on S opens up a path between A and $X^*(a)$. However we can show that for this particular twin network this is not a problem. TO do so, consider the following example of a model violating this consider the following where all variables are binary:

$$A \sim Ber(\frac{1}{2}) \tag{4.8}$$

$$U_X \sim Ber(\frac{1}{2}) \tag{4.9}$$

$$X = A \oplus U_X \tag{4.10}$$

$$S = X \tag{4.11}$$

Now under this model the dependency $X(a) \perp A \mid S = 1$ corresponds to $X(a) \perp A \mid X = 1$. However given $X = 1$ we know $X(a) = 1$ if $A = a$ and $X(a) = 0$ if $A = \neg a$. Therefore $X(a) = \mathbb{1}(A = a)$ given $X = 1$ and so $X(a) \not\perp A \mid S = 1$ for this model. As this constraint is violated for one discrete model it must be violated on all but a measure zero subset of discrete models. Therefore as one model exists in the class which violates the independence we must have this for almost all discrete models. Therefore giving faithfulness level guarantees.

4.E Proof of Corollary 23

Proof. By rearranging the requirement given in Proposition 22, we have that if $X^*(a) \perp A \mid S = 1$ then for all x with $P(X = x \mid A = a) > 0$:

$$\begin{aligned} P(S = 1 \mid X = x, A = a) &= P(S(a) = 1 \mid X = x, A = a) \frac{P(S = 1 \mid A = a)}{P(S = 1 \mid A = a')} \\ &= P(S(a) = 1 \mid X = x, A = a) \frac{P(A = a \mid S = 1) P(A = a')}{P(A = a' \mid S = 1) P(A = a)} \qquad (4.12) \\ &\leq \frac{P(A = a \mid S = 1) P(A = a')}{P(A = a' \mid S = 1) P(A = a)} \qquad (4.13) \end{aligned}$$

where (4.12) follows from applying Bayes' rule and (4.13) uses that a probability is bounded above by 1. Hence if there exists an x which violates this bound we must have $X^*(a) \not\perp A \mid S = 1$ and so by Lemma 21 no model in $\mathcal{M}_{S=1}$ captures the true counterfactuals. □

4.F Detailing the Calculations and Datasets

In this appendix we detail all the datasets and calculations from Table 4.4.1. All population data for this section is from the World Bank [220].

4.F.1 Adult Dataset

The Adult Dataset [124] contains data on 48 842 individuals taken from the 1994 US census database. The dataset contains 16 attributes for each individual with an aim to predict if an individuals income is greater than $50,000. The dataset was formed by taking all individuals in the US census database with these 16 attributes recorded and then removing based on certain attributes to get clean records. For example all individuals who were logged as not working any hours were removed.

In the this dataset the gender distribution is 67% male and 33% female. The World Bank estimates that in 1994, 49.1% of the US population were male and 50.9% were female. Plugging this in gives:

$$\begin{aligned} P(S = 1 \mid X = x, A = \text{Female}) &> \frac{P(A = \text{Female} \mid S = 1) P(A = \text{Male})}{P(A = \text{Male} \mid S = 1) P(A = \text{Female})} \\ &= \frac{0.33 \times 0.491}{0.67 \times 0.509} \\ &= 0.475. \end{aligned}$$

4.F.2 German Credit Dataset

The German Credit Dataset [99] describes has the financial details of 1000 bank customers applying for a loan. The task is to predict from a list of 20 covariates if someone is a good or bad credit risk. This dataset is from the year 1994.

In the German credit dataset the gender is 31% female and 69% male. The World Bank estimates that in 1994, 51.5% of the German population were female and 48.4% were male. This gives:

$$\begin{aligned} P(S = 1 \mid X = x, A = \text{Female}) &> \frac{P(A = \text{Female} \mid S = 1)P(A = \text{Male})}{P(A = \text{Male} \mid S = 1)P(A = \text{Female})} \\ &= \frac{0.31 \times 0.484}{0.69 \times 0.516} \\ &= 0.421. \end{aligned}$$

4.F.3 Law School Dataset

The law Sshool dataset [219] is as described in Section 4.2.2. The dataset was collected in 1998.

Again using World Bank estimates we have that in 1998 the US population is 49.7% female and 50.3% male. In the law school dataset the gender distribution is 43.8% female and 56.2% male. This gives:

$$\begin{aligned} P(S = 1 \mid X = x, A = \text{Female}) &> \frac{P(A = \text{Female} \mid S = 1)P(A = \text{Male})}{P(A = \text{Male} \mid S = 1)P(A = \text{Female})} \\ &= \frac{0.438 \times 0.503}{0.562 \times 0.497} \\ &= 0.789. \end{aligned}$$

4.G Proof of Lemma 24

Proof. As noted previously we have $X(a) \perp A$ for the counterfactuals generated by causal models satisfying the assumptions on $\mathcal{M}$ in this Lemma. Hence as $\widehat{Y}$ is a

function of X (possibly also with some independent noise) we have $\widehat{Y}(a) \perp A$.

$$
\begin{aligned}
P(\widehat{Y} \mid A = a') &= P(\widehat{Y}(a') \mid A = a') && (4.14)\\
&= P(\widehat{Y}(a') \mid A = a) && (4.15)\\
&= \mathbb{E}_{P(X|A=a)}\left(P(\widehat{Y}(a') \mid X, A = a)\right) && (4.16)\\
&= \mathbb{E}_{P(X|A=a)}\left(P(\widehat{Y}(a) \mid X, A = a)\right) && (4.17)\\
&= P(\widehat{Y}(a) \mid A = a)\\
&= P(\widehat{Y} \mid A = a)
\end{aligned}
$$

Where (4.14) follows from consistency, (4.15) follows from $\widehat{Y}(a') \perp A$, (4.16) uses the law of total expectation and (4.17) uses the definition of counterfactual fairness. Hence $\widehat{Y} \perp A$. □

4.H Proof of Corollary 25

Proof. We take predictors to mean any function, since counterfactual fairness places no restriction on what value the predictors take.

First if we satisfy the counterfactual outcome independence under selection then we have by Lemma 21 that we have a model in $\mathcal{M}_{S=1}$ with the correct counterfactuals and $\widehat{Y}$ is clearly counterfactually fair relative to this as it has the same counterfactuals as the true model. Therefore by Lemma 24 $\widehat{Y}$ is independent of A on the dataset, so when $S = 1$.

Now for the converse we show that functions f counterfactually fair according to $\mathcal{M}^*$ will not in general satisfy $f(X, A) \perp A \mid S = 1$ by finding a specific function which violates this.

First as $X^*(a) \not\perp A \mid S = 1$ for some a we have some coefficient X_1 such that $X_1^*(a) \not\perp A \mid S = 1$. Now let f_1 be a function such that for inputs $X = x$ and $A = a'$, $f_1(x, a')$ will be a random draw from the true posterior for $X^*(a)$ arising from $\mathcal{M}^*$, that is $P(X_1^*(a) \mid X = x, A = a')$.

Now, clearly this function will be counterfactually fair according to the true model. However we have $f_1(X, A) \not\perp A \mid S = 1$. This is because given $A = a'$ a random draw from $f_1(X, A) \mid \{S = 1\}$ will be a random draw from $X_1^*(a) \mid \{A = a', S = 1\}$. As we know $X_1^*(a) \not\perp A \mid S = 1$ we conclude $f_1(X, A) \not\perp A \mid S = 1$ and so we are done. □

5

The Fragility of Fairness: Causal Sensitivity Analysis for Fair Machine Learning

Abstract

Fairness metrics are a core tool in the fair machine learning literature (FairML), used to determine that ML models are, in some sense, "fair." Real-world data, however, are typically plagued by various measurement biases and other violated assumptions, which can render fairness assessments meaningless. We adapt tools from causal sensitivity analysis to the FairML context, providing a general framework which (1) accommodates effectively any combination of fairness metric and bias that can be posed in the "oblivious setting"; (2) allows researchers to investigate combinations of biases, resulting in non-linear sensitivity; and (3) enables flexible encoding of domain-specific constraints and assumptions. Employing this framework, we analyze the sensitivity of the most common parity metrics under 3 varieties of classifier across 14 canonical fairness datasets. Our analysis reveals the striking fragility of fairness assessments to even minor dataset biases. We show that causal sensitivity analysis provides a powerful and necessary toolkit for gauging the informativeness of parity metric evaluations. Our repository is available here.

5.1 Introduction

Fair machine learning (FairML) is a theoretical approach to studying and remediating disparities in prediction and allocation systems based on machine learning algorithms. A core focus of the field has been to develop, evaluate, and train models to satisfy a number of "fairness metrics". These metrics operationalize the social ideal of fairness as a statistical quantification of some performance measure compared across demographic groups. Such evaluations often play an important role in auditing ML systems [33, 170, 153] to certify whether models satisfy some tolerable level of statistical disparity.

Real-world data, however, is frequently plagued by a variety of measurement biases and other violated assumptions which can undermine the validity of fairness metrics[16, 107]. While such biases come in many forms, in this work we focus on the following: noisy or poorly-defined outcome measures (proxy bias) [81], the observation of samples or outcomes from only a subset of the population (selection bias) [19, 111], or causal impacts on outcomes through background policies within a firm's control [50], which we term *extra-classificatory policies*, or ECPs. We focus on these varieties of bias due to their ubiquity in FairML applications, which we demonstrate through an analysis of their prevalence and magnitude in a range of benchmark datasets (see Table 5.3.1 and App. 5.D).

Motivated by the problems posed by such measurement biases, we offer a framework based on graphical causal inference to operationalize assumptions about data quality issues, alongside methods adapted from causal sensitivity analysis for statistical quantification of their impacts on fairness evaluations. This framework enables both ML practitioners and auditors to empirically gauge the sensitivity of parity metrics to assumption violations for specific combinations of metrics, datasets, and use cases. Causal inference is particularly apt for this problem, as it provides a formal language within which to precisely identify the goals of a particular study: what is the quantity we seek to estimate, and in which population? This accounts for the success of causal inference in the social sciences and makes it similarly well-suited for use in the arsenal of ML auditing tools.

We leverage recent developments in automated discrete causal inference, particularly the autobounds framework of Duarte et al. [62], to provide a unified causal sensitivity analysis framework for the "oblivious" setting, as laid out in Hardt et al. [93]. In this setting, we only have access to protected attributes A, the true target labels Y, and the predicted labels $\hat{Y}$, but not to covariates X. For example, in evaluating racial discrimination in loan granting, one has access to the race attribute A, the true repayment rate Y, and the predictions $\hat{Y}$, but not to input features X nor the form of $\hat{Y}(X)$.

This lends us a straightforward procedure for performing sensitivity analyses for any combination of measurement bias and suitably well-behaved metric that can be posed

obliviously: (i) Express the bias in terms of a causal graph—a directed acyclic graph, henceforth DAG, (ii) choose a sensitivity parameter to control the degree of bias, (iii) provide any additional probabilistic assumptions or relevant structural knowledge. The problem of bounding a statistic under a given degree of bias can then be converted to solving a given constrained optimization problem [62], which is achieved via a branch and bound solver [22], leading to valid bounds even when a global optimum is not reached.

We apply this framework to systematically explore the sensitivity of different metrics to the three biases—proxy label bias, selection bias, and extra-classificatory policy bias—for different datasets and classifiers. Our results reveal that many fairness metrics are in fact *fragile*: realistic violations of core underlying assumptions can imply vacuously wide sensitivity bounds. In other words, features of typical deployment contexts can easily render fairness evaluations useless or uninformative.

The fragility of well-known fairness metrics to pervasive biases represents one key empirical finding. Our second core result demonstrates the existence of tradeoffs between the complexity and fragility of fairness metrics. The robustness of parity notions scales inversely with their dependence on predictive outcomes and the intricacy of this dependence function. We find demographic parity to be most robust to measurement biases, while predictive parity metrics exhibit the most fragility to bias. In light of known incommensurability results [120], we urge the importance of understanding these tradeoffs and their practical implications, for both practitioners and auditors alike.

With these experiments, we aim to demonstrate that the biases we describe are an unavoidable aspect of the FairML problem, not a mere addendum. As such, we have hopes that our unified sensitivity analysis framework can enable both auditors and practitioners to understand how robust their "fairness" evaluations are to various measurement biases by precisely articulating the quantity they wish to evaluate and its divergence from what has been measured. Finally, we hope this work will inspire greater emphasis going forward on the realities of real-world deployment scenarios, such as measurement biases, and their impacts on fairness evaluations.

5.2 Related work

Proxy Label Bias Proxy Label bias is a foundational problem in FairML, endemic within criminal justice and legal applications of ML, which the fairness literature originally arose to address [16, 82]. Fogliato et al. [81] presents one of the earliest considerations of sensitivity analysis within FairML, algebraically deriving sensitivity bounds for proxy label bias. As we demonstrate in 5.6, our approach is capable of re-deriving and extending these results. In a similar vein, Adebayo et al. [2] studied the effects of label noise on fairness metric evaluation, although this work was largely

empirical. Guerdan et al. [90] propose several causal models for reasoning about proxy labels in human—algorithm joint decision making, which can be rendered compatible with our sensitivity analysis framework. Further work has explored alternate aspects of fairness evaluations under proxy label bias [216, 222].

Selection Bias Selection bias was first considered in FairML as "selective labels" by Lakkaraju et al. [129], focused on the scenario in which a policy determines which outcomes are observed. Kallus and Zhou [111] study the effects such a biased policy can have on equalized odds for the unselected population when predictors are trained only on the selected population. Various works now link selection bias in fairness to causal inference [76, 188] with Goel et al. [89] providing a summary of the different types of selection in terms of causal graphs. Coston et al. [51] take an importance-weighting approach to train FairML classifiers with selective labels, under assumptions on the structure of the missingness. Zhang and Long [230] study how to assess the accuracy parity in unselected data, from the selected data, under assumptions on selection structure and the FairML model class.

Extra Classificatory Policy Bias We investigate the impacts of extra classificatory policies, that is, policies under the control of the predicting agent which causally affect the outcome of interest. This work is related to counterfactual risk assessments [50], and subsequent work on counterfactual equalized odds [152]. Sensitivity analysis approaches have further been developed for unmeasured confounding [34, 171]. Our methodology differs from these works in our focus on the oblivious setting. As such, we focus less on identification, but rather on how influential a policy would need to be before it could significantly impact a fairness evaluation.

Additional Related Work Beyond the above work on sensitivity analysis there are other general approaches to understanding the robustness of algorithmic fairness to data bias [136], such as adversarial robustness [38, 126] and distributional robustness [204]. These are very flexible in terms of the types of bias they can represent, however, this renders them less interpretable and less able to incorporate additional assumptions. One measurement bias we did not consider is proxy attribute bias [39], for which there exist quite comprehensive sensitivity analysis results [113]. There is also a selection of work performing sensitivity analysis for unmeasured confounding in causal fairness [118, 187, 223]

5.3 Measurement Biases

This section introduces the measurement biases we consider via two recurring examples, demonstrating how they arise in the wild and emphasizing the role of practitioner choice. We first deliver a conceptual illustration of said biases (3.1-3.3) before presenting an

empirical analysis of their prevalence across a range of FairML benchmark datasets [132].

5.3.1 Proxy Label Bias

We first discuss proxy label bias, introducing it via the following example:

> **An Algorithmic Hiring System:** *Suppose that a company receives thousands of applications for every job they advertise. To handle this, they elect to build an ML-based system to assist in sifting through candidates. They opt to construct a model for predicting employees' performance review scores from their resumés, which is then used to assign scores to applicants based on predicted performance reviews. Finally, to ensure that the model is fair, they check against standard fairness metrics on a held-out subset of the training data.*

The company described has taken steps to ensure that its job candidate filtering model is "fair." However, it has only run parity tests relative to the variable chosen as its target of prediction. The latent variable of interest in this scenario, what the firm ideally strives to predict, is "employee quality." However, "employee quality" is multifaceted and socially constructed; it is not a phenomenon that can be directly and objectively measured.[1]. Instead, engineers leverage a *proxy label*. Unlike the nebulous property of employee quality, employee performance reviews are readily available. It is facially not unreasonable to take performance reviews to stand in for employee quality; after all, the one exists, at least putatively, to track the other. However, there is a problem with this strategy: It has been extensively documented that performance reviews are often discriminatory—in other words, performance reviews are both a worse signal of the true underlying latent for certain demographics, and are skewed negative for those demographics relative to the true latent [61, 176].

If an outcome is biased, then classifiers optimizing predictive accuracy on a proxy can appear to satisfy a fairness metric when, in reality, the metric has inherited the biases of the outcome. Practitioners must understand whether a particular outcome is well-suited to the underlying decision problem, alongside any skew or measurement bias that may be induced by using such a measure. Often the use of a proxy outcome is unavoidable. If so, sensitivity analysis is a key tool for understanding what impact biases in the proxy could have on parity metric evaluations.

5.3.2 Selection Bias:

Introducing our second example to discuss selection bias:

[1] For discussion of how mismatches between unobservable latents and their proxies play in fair ML, see [107].

An Automated Loan-Approval Algorithm: *A large bank has an online lending platform that receives thousands of loan applications per day, the vast majority of them for under USD $1,000. The bank cannot afford to assign employees to assess all the applications, and elects to automate the process. The bank uses its data on loan repayment to fit a model for likelihood of repayment, with loans automatically approved if the estimated probability of repayment sits above some threshold. The model is once again assessed via standard fairness metrics on holdout data.*

As in the previous example, the bank only possesses repayment information for the population that has historically been approved for loans; the subset of the broader population that was not granted a loan is therefore unobserved. Each firm trains a classifier on the subpopulation for which it possesses outcome data. The classifiers' performance on the entire population of job applicants and loan applicants, respectively, are unknown. Further, fairness guarantees on only selected populations can fail to meaningfully extrapolate to deployed models, especially when historical selection procedures encode biases. This phenomenon has been referred to as *selective labels* [129] or *prejudiced data* [111], and now more commonly under the catch-all term of *selection bias* [89].

While selection bias covers the selective labels case, it encompasses a broader class of examples. An important question in the evaluation of fairness metrics is what population the practitioner would ideally want to evaluate the statistic in. We refer to this as the *reference population*. At bare minimum, the reference population should be the population the model is deployed on, not the training population—as the selective labels literature points out. However, there are situations where arguments could be made for broader reference populations. For example, should fairness metrics for an algorithmic hiring system be assessed on the local pool of applicants to that company, or the global pool of applicants to similar roles? Dai et al. [54] demonstrate that, with reputational dynamics in play, applicants may strategically apply to firms based on their chances of success. A firm's hiring practices can therefore satisfy fairness desiderata relative to their applicant pool having radically skewed the demographics of that population via a history of discriminatory hiring policies. Put crudely: the firm that "women candidates know not to apply to" is not non-discriminatory, but an evaluation of hiree demographics relative to the applicant pool might deceptively depict it as such.

There are no universalisable solutions to such issues, as the choice of reference population must depend on precisely what task the model is being trained to perform, and where it will ultimately be deployed. This further points to a need for practitioners to be clear about what population they have chosen, how it may differ from the data they have measured, and why this specific reference population is preferable to others for the task at hand.

5.3.3 Extra-Classificatory Policy Bias

A third issue emerges when considering additional context in the loan-approval case study:

> **Automated Loan-Approval (cont):** *Our bank now has a model for loan approval. However, they must also set interest rates for approved loans. For this, they employ a legacy model, which works off of a number of factors including credit score, loan amount, lendee address, and various macroeconomic variables.*

While the bank has determined that the classifier is fair in the sense that it does not discriminate according to known demographic features and relative to repayment history, these notions of fairness fail to account for the interest rate that the bank determines. As this has a direct and powerful impact on a lendee's likelihood of repayment, if it is set in an (intentionally or unintentionally) discriminatory manner, it can have a large and unaccounted-for effect on the evaluation of fairness metrics.

This is not specific to the loan approval setting; in many scenarios, firms control a number of "levers" that causally impact outcomes for classified populations—we call these *extra-classificatory policies*. Such extra-classificatory policies can drastically affect outcomes, and so shape the data on which models are trained. Through such policies, a firm can create the appearance of demographic base rate discrepancies which then, via predictive models trained on historical data, serve to justify differential lending policies across demographic subpopulations.

The issues pointed towards by extra-classificatory policies are multi-faceted and context-specific. As such, a key challenge to the FairML community is detailing such issues and forming mathematical models of them. We focus on binary policies (e.g. [50]), leaving more general settings to future work.

5.3.4 Cross-Dataset Analysis

To demonstrate the prevalence of these issues, we analyze the presence of each measurement bias for the FairML benchmark datasets given in Le Quy et al. [132]. We remove any datasets not associated with a concrete decision problem, leaving 14 datasets from a variety of different domains, including financial risk, criminal justice, and employment/university admissions. We summarise the results in Table 5.3.1, with the full table of datasets and biases along with rationales available in Appendix 5.D. Our results demonstrate the scope of the problem created by measurement bias, with all the datasets displaying at least one of the problems and 60% displaying all three simultaneously. From this, we see that such biases are themselves a key part of the FairML problem, not optional complications. Moreover, this points to the need for methodological solutions for evaluating and training FairML models in settings where multiple biases are present simultaneously.

Proxy Bias	Selection Bias	ECP Bias	One Bias	Two Biases	Three Biases
69%	85%	85%	100%	92%	61%

Table 5.3.1: We document the proportion of realistic FairML benchmark datasets (from Le Quy et al. [132]) that exhibit each of the three biases we discuss. For more details see App. 5.D.

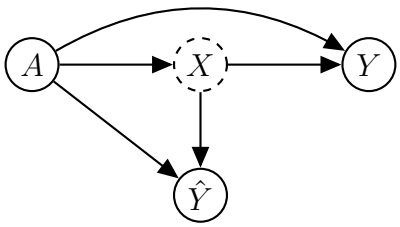
5.4 Graphical Causal Sensitivity Analysis

Here we outline some technical background on graphical causal sensitivity analysis, from graphical sensitivity analysis to the autobounds framework [62] which automates discrete sensitivity analysis. In the following section, we apply this to FairML to construct a sensitivity analysis tool for oblivious settings [93], where we do not observe covariates and all other variables are discrete.

Notation We let Y denote the outcome the practitioner wishes to measure, X the observed covariates, A the protected/sensitive attribute, and $\hat{Y}$ the prediction of Y with domains $\mathcal{Y}, \mathcal{X}, \mathcal{A}, \hat{\mathcal{Y}}$.

5.4.1 Causal Background

We begin by defining the structural causal model (SCM) approach to causality [162, 175]. Here we model causal relationships via deterministic functions of the observed variables and additional latent variables, with the latter representing the unobserved or random parts of the system. We will always label the observed variables $\mathbf{V}$ and the unobserved variables $\mathbf{U}$. These equations lead to a causal graph that has a node for each variable and a directed edge $V_1 \to V_2$ if V_1 is an argument of the function determining the value of V_2. To illustrate this, the following figure demonstrates the SCM and corresponding graph we assume throughout for the relationships between $A, X, Y,$ and $\hat{Y}$:



(a) The directed acyclic graph representing the relationships between $A, Y, \hat{Y}$ and the unobserved X. The latent variables are implicit in the DAG.

$$
\begin{aligned}
X &= f_X(A, U_1) \\
Y &= f_Y(X, A, U_2) \\
\hat{Y} &= f_{\hat{Y}}(X, A, U_3) \\
\mathcal{C} &= \left(\{f_X, f_Y, f_{\hat{Y}}\}, P(\mathbf{U})\right)
\end{aligned}
$$

(b) The SCM $\mathcal{C}$ over the DAG in (a) defined by a set of functions on $\mathbf{V} = \{A, X, Y, \hat{Y}\}$ and a probability distribution over the implicit latent variables.

Figure 5.4.1: Example of a DAG and the corresponding SCM. Unobserved variables are dashed.

We will use $\mathcal{C}$ to depict an SCM, which consists of a collection of functions and a probability distribution over the noise terms. A complete definition of SCMs can be found in Appendix 5.B.1.

Marginalisation In Causal Models Often—and especially in FairML applications—we do not observe all relevant variables. For example, in this work, we assume the covariates X are unobservable. However, this is less of a problem than it originally appears due to latent projection, as introduced by Verma and Pearl [211]. Latent projection allows us to marginalize out any unobserved variables while preserving the causal structure over observed variables, as demonstrated by Evans [70]. We visualize this process in Fig. 5.5.1, where we marginalize over X leaving just $A, \hat{Y}$, and Y and latent variable, U. We outline this procedure in detail in Appendix 5.B.2. The important point is that we can always preserve the causal structure over our observable variables by using a finite number of latent variables. This is the case regardless of how many variables we marginalize out.

5.4.2 Partial Identification and Sensitivity Analysis

We now show how structural causal models can be used to perform sensitivity analyses for causal (or non-causal) queries, first introducing the important concept of partial identification.

Partial Identification In partial identification, the goal is to understand what values a particular statistic can take relative to our assumptions. We call this a *Query* and write it as a function $\mathcal{Q}(\mathcal{C})$ which takes an SCM and returns a real number. For example, if we wanted to measure counterfactual fairness (for binary A), we could

define the query $\mathcal{Q}_{\text{CF}}$ as:

$$\mathcal{Q}_{\text{CF}}(\mathcal{C}) \coloneqq P_{\mathcal{C}}(\hat{Y}(A=1) \neq \hat{Y}(A=0))$$

Where $\mathcal{Q}_{\text{CF}}(\mathcal{C}) = 0$ exactly when the predictor is counterfactually fair according $\mathcal{C}$ [73].

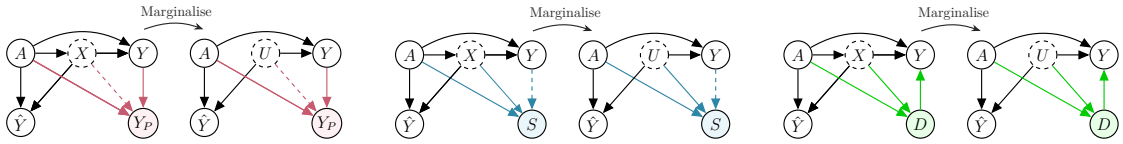
Given a query of interest, $\mathcal{Q}$, the goal of partial identification is to understand what possible values $\mathcal{Q}$ can take given the practitioner's prior knowledge and assumptions. Here we will consider partial identification for a fixed DAG and a set of constraints on the probability distributions and functions defining the SCM. Practitioners can encode these assumptions by defining a set of SCM models, which we will write $\mathcal{M}$. The most natural example here is to let $\mathcal{M}$ contain all possible causal models arising from the graph with the same observational distribution as the measured dataset. Practitioners can restrict this set of SCMs to incorporate more domain-specific information, making the bounds more informative. Using this notation, partial identification is rendered as a pair of optimization problems that lower and upper bound the query of interest over $\mathcal{M}$:

$$\min_{\tilde{\mathcal{C}} \in \mathcal{M}} \mathcal{Q}(\tilde{\mathcal{C}}) \leq \mathcal{Q}(\mathcal{C}) \leq \max_{\tilde{\mathcal{C}} \in \mathcal{M}} \mathcal{Q}(\tilde{\mathcal{C}}) \tag{5.1}$$

Sensitivity Analysis In sensitivity analysis, the goal is to understand how violations of assumptions affect the measure of a statistic. The initial step is to define some sensitivity parameter, which measures the degree to which an assumption is violated. For example, in the proxy attribute literature [39, 113], the goal is to understand how sensitive fairness metrics are to mismeasured protected attributes. For example, as we will discuss later in terms of proxy bias, a natural sensitivity parameter would be $P(Y_P \neq Y)$, where Y_P is the proxy outcome. We may then let $\mathcal{M}_{\text{Prox}}(\delta)$ be the set of causal models that have $P(Y_P \neq Y) \leq \delta$ and comply with the practitioner's assumptions. By repeatedly solving the partial identification problem for different δ to understand how large δ must be, the statistic $\mathcal{Q}$ becomes uninformative.

5.4.3 Discrete Causal Sensitivity Analysis

To perform the partial identification required for causal sensitivity analysis, we have to solve the max/min problem in (5.1). This problem as formulated generically is not tractably solvable, but additional structure on either the query, $\mathcal{Q}$, or set of models, $\mathcal{M}$, can lead to tractable optimization problems and computable bounds. We focus on settings where all variables are discrete, which is particularly helpful for partial identification problems [85, 26, 172] due to the function response framework [15]. This framework takes advantage of the fact that, given a causal graph $\mathcal{G}$, if we fix the latent variables $\mathbf{U}$, the structural equations are deterministic functions of their other inputs in $\mathbf{V}$. If the observed variables are discrete, there are only finitely many such functions. As a result of this, we can represent every single SCM using the one set of



(a) Proxy bias **(b)** Selection bias **(c)** ECP bias

Figure 5.5.1: Causal graphs for each of the biases showing the assumed causal structure over all variables, and the implied structure upon marginalizing out X. Dashed lines denote varying assumptions.

fixed structural equations and a distribution over some discrete latent variables, $\tilde{\mathbf{U}}$. This means that any SCM with discrete observed variables and a fixed graph $\mathcal{G}$ can be represented entirely by the distribution $P(\tilde{\mathbf{U}})$, and so by a point in the probability simplex Δ^k for some k [62, 71].

Duarte et al. [62] showed that this allows partial identification problems (5.1) to be converted into tractable optimization problems, where now the set of causal models $\mathcal{M}$ corresponds to a subset of the probability simplex $\mathcal{M}^\Delta \subset \Delta^k$ and the query $\mathcal{Q}$ becomes a function $f_\mathcal{Q} : \Delta^k \to \mathbb{R}$. Moreover, any statement that can be written as a polynomial in probabilities over factual and counterfactual statements corresponds to a fractional polynomial in $\mathbf{p}_\mathcal{C}$. So, if $f_\mathcal{Q}$ is a polynomial in such probabilities and any prior assumptions can be stated via probabilistic statements and arithmetic operations, the partial identification problem can be converted into a polynomial programming problem. Duarte et al. [62] propose to solve these partial identification problems via branch and bound solvers [22], which ensures that the program produces valid bounds even if convergence is not reached.

5.5 Causal Sensitivity Analysis for FairML

We now apply the methodology outlined in Section 5.4.3 to the FairML setting to create a sensitivity analysis tool for parity metric evaluations. We focus on settings where $(A, Y, \hat{Y})$ are discrete and the auditor does not have access to the covariates X. The following steps lead to a sensitivity analysis tool for any measurement bias that can be stated obliviously and any statistic that can be written as a polynomial in factual and counterfactual probabilities:

1. Determine how the sampled population differs from the target population, expressing the difference in terms of a causal graph over all variables. Marginalize out X, to leave a causal structure over $(A, Y, \hat{Y})$ and any bias-specific variables.
2. Choose a sensitivity parameter to control the degree of measurement bias and provide any additional knowledge relevant to the task at hand.
3. With all this perform the sensitivity analysis by repeatedly solving the optimization

problem in (5.1) for the test statistic using the methodology outlined in 5.4.3.

We apply this procedure to each of the measurement biases given in Section 5.3, using the causal graphs in Fig. 5.5.1 to depict the biases. Causal graphs are context-specific, so we do not expect these to be appropriate in all cases. Instead, we use them as plausible graphs for showcasing our framework.

Proxy Label Bias **(1)** We represent the difference between the measured and target populations using an additional variable Y_P. This denotes the measured proxy of the outcome. We assume that this outcome is a noisy version of the true outcome Y, where the noise depends on A and can optionally depend on X. As we show in Appendix 5.B.3, assuming the proxy depends on additional unobservables leads to the same graph over $(Y, A, \hat{Y})$. **(2)** For the sensitivity parameter, we use the probability that the proxy differs from the outcome the practitioner hopes to measure, so $P(Y_P \neq Y)$.

Selection Bias **(1)** We signal whether or not an individual is selected with a binary variable S, which we assume depends on an individual's protected attribute, covariates, and, in some instances, the outcome. **(2)** For the sensitivity parameter, we choose $P(S = 0)$, the probability of a sample not being selected. This controls the proportion of the population which remains unobserved. However, there are other natural choices for sensitivity parameters, such as statistical measures of sample quality [151]. For selection bias, the practitioner could have significant information about the unselected population which could be used to tighten bounds. For example, selective labels would lead to information on the covariates for the unselected populations, and thus the $(A, \hat{Y})$ proportions.

Extra-Classificatory Policies There are a plurality of plausible ways to mathematically represent the problems arising from ECPs, with different formulations being better suited to different concerns. Here we proceed as follows: **(1)** We use an additional variable, D, to depict the policies in the firm's control, which we take to be binary. We then use counterfactual versions of parity metrics, similar to Coston et al. [50], with formulations given in Appendix 5.A.2. **(2)** We assume the treatment is monotonic, so $Y(D = 1) \geq Y(D = 0)$ and use the average treatment effect, $\mathbb{E}(Y(D = 1) - Y(D = 0))$, as a sensitivity parameter. For ECPs, there are additional constraints that could be added, for example, if a policy is observed we can include that data explicitly.

5.6 Experiments

In this section, we showcase the use of causal sensitivity analysis for fairness applications by performing sensitivity analyses for each of the biases introduced above on a set of benchmark datasets. In doing so, we aim to reveal some of the nuances of performing

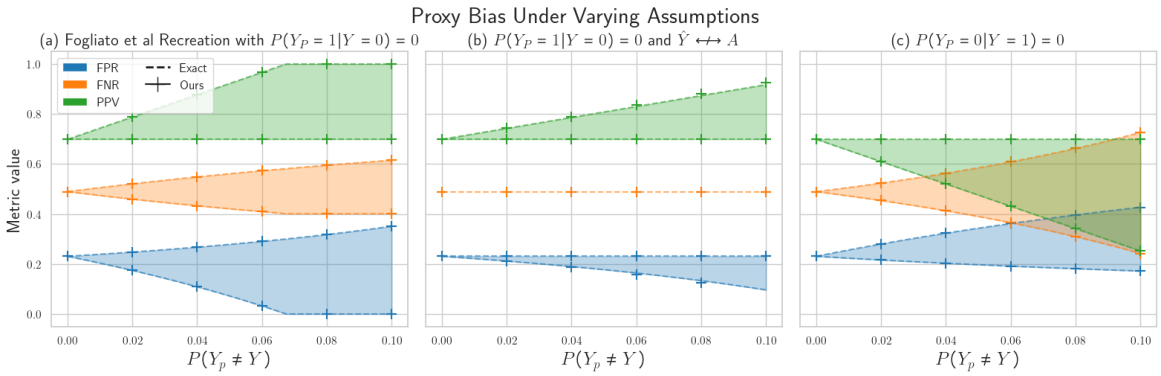


Figure 5.6.1: In this we directly recreate the plots from Fogliato et al. [81] for a predictor trained on the COMPAS dataset, allowing for some probabilistic and causal assumptions to vary. The dashed lines represent exact bounds on each statistic for increasing $P(Y_P \neq Y)$, which follow from Fogliato et al. [81] or our derivations in Appendix 5.C.1.2. (a) represents the original setting, where we have $P(Y_P = 1 \mid Y = 0) = 0$, in (b) we drop the dashed edge between X and Y_P in the causal graph in Fig. 5.5.1a, and finally for (c) we instead take $P(Y_P = 0 \mid Y = 1) = 0$. As we can see, at all points we query, the automatically derived bounds recover the algebraically derived bounds.

sensitivity analyses for various types of bias. These experiments further lend themselves to some general conclusions about the complexity of real-world fairness evaluations. We present additional results in Appendix 5.C, including tests of causal fairness metrics that reveal the effects of measurement bias in a causal setting.

5.6.1 Recreating Fogliato et al. [81] under varying assumptions

Fogliato et al. [81] aims to assess how sensitive the false positive rate (FPR), false negative rate (FNR), and positive predictive value (PPV) are to proxy bias for a predictor trained on the COMPAS dataset [9], under the assumption that $P(Y_P = 1|Y = 0) = 0$. The outcome in the COMPAS dataset is reported re-offense. This outcome a proxy, because we cannot actually observe whether someone has re-offended, we only know if they were convicted of a new offense. The assumption that $P(Y_P = 1|Y = 0) = 0$ implies that whenever someone is reported to have re-offended they actually re-offend; all the measurement error hence comes from people who did re-offend who did not get caught. We start by recreating the results of the original study from just the DAG in Fig. 5.C.1(a). This confirms the correctness of the computational bounds: they always match the true algebraically derived bounds in Fogliato et al. [81].

Next we demonstrate the flexibility of our framework by switching from the assumption that $P(Y_P = 1|Y = 0) = 0$ to $P(Y_P = 0|Y = 1) = 0$. This is probably a less realistic assumption in the COMPAS dataset, implying all measurement error comes from false convictions rather than under-reporting, but in a different context, this might

be a more reasonable assumption. We derive the algebraic bounds to confirm the computational bounds match. The point here is that practitioners can easily encode whatever assumptions make sense in their particular context and quickly get results without needing to algebraically derive bounds, and that these results really do vary under different assumptions.

Finally, in Fig. 5.C.1(c), we drop the dashed edge between X and Y_P in Fig. 5.5.1a. This might seem like an odd experiment: we do not actually observe X, and X already causes Y, so how could dropping the edge from X to Y_P really matter? Without the $X \to Y_P$ edge Y_P is independent of U conditional on Y, which significantly tightens the bounds for FPR and PPV and fully identifies FNR.[2] Once again, we derive the algebraic bounds and find that the computational bounds match. This final experiment simultaneously demonstrates potential drawbacks of causal sensitivity analysis and the enormous benefit of being able to easily run analyses under varying assumptions. Sensitivity analysis can be dangerously misleading if we include unrealistic assumptions, giving us a false sense of security. But the remedy is straightforward: analysts should always run an assumption-lite analysis before incorporating more domain-specific information so that they can understand which assumptions are driving their results and make a decision about whether those assumptions are realistic.

5.6.2 Intersection of Biases

Motivated by Section 5.3.4, we now consider settings exhibiting multiple simultaneous biases. In Fig. 5.6.2, we provide a sensitivity plot for proxy and selection bias simultaneously for an equalized odds predictor trained on the *Adult* dataset [21]. We plot the upper bound of the equalized odds value as the sensitivity parameters for both biases vary, with a maximum of 5% proxy labels and 5% of the population being unmeasured. This exercise demonstrates that the presence of multiple biases can quickly render parity evaluation meaningless, with the upper bound reaching the maximum possible value of the statistic. Whilst an upper bound inherently represents the worst-case scenario, this experiment showcases that practitioners must provide additional context-specific assumptions to imbue parity evaluations with

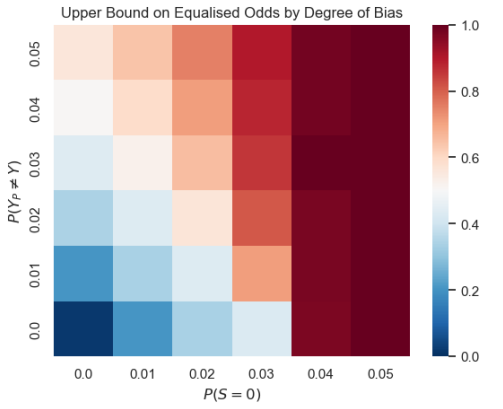


Figure 5.6.2: Combination of Proxy and Selection Bias for an equalized odds predictor on the Adult dataset.

[2] On a technical level, the identification of the FNR points to the utility of sensitivity analysis for discovering new FairML-specific identification theory, an area which has received little attention to date.

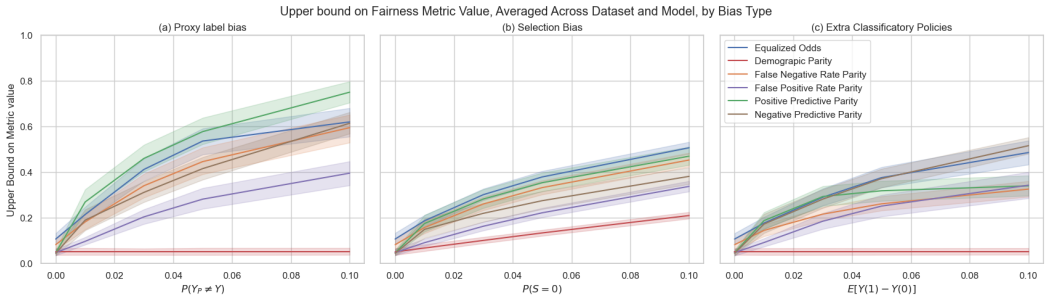


Figure 5.6.3: Results of our cross-dataset study, in which we assess the sensitivity of multiple ML predictors trained to satisfy various parity constraints on the fairness benchmarking datasets listed in Appendix 5.D. We can see different metrics are susceptible to bias in different ways, with notably demographic parity being more robust than more complicated, outcome-dependent metrics.

meaning in the presence of measurement biases, especially when multiple biases are present simultaneously. We include additional results in Appendix 5.C.2.2, which show that biases compound in unpredictable, nonlinear ways.

5.6.3 Cross-Dataset Experiments

Finally, we leverage our framework to systematically explore the sensitivity of the most ubiquitous parity metrics to the measurement biases here surveyed on real benchmark datasets. We run sensitivity analyses on 14 commonly used fairness datasets from [132] training logistic regression, naïve Bayes, and decision trees to satisfy different parity metrics. In Fig. 5.6.3, we provide the results of this experiment, plotting the average upper bound for each metric value at different levels of sensitivity. This further supports our thesis that measurement biases present a severe issue to the informativeness of parity metric evaluation, as we see that even small amounts of bias can render the original parity evaluation meaningless relative to what it seeks to measure. Secondly, we find that this fragility is not uniform across metrics. We can see that demographic parity is, unsurprisingly, more robust than more complicated outcome-dependent metrics. Equalized odds is less robust than FPR/FNR due to the fact that it is measured as the maximum of both. Finally, predictive parity metrics are less robust to biased outcomes than metrics which involve conditioning on the outcome, such as FPR/FNR and equalized odds. We present full experiment details, further analysis, and plots per dataset in Appendix 5.E. We also include a sensitivity for the Folktables dataset [59], which represents one of the most popular datasets for modern FairML.

5.7 Codebase and Web Interface

We have developed both a codebase and a web interface to ensure our framework is as usable as possible. The core of both tools are our bias configs – described in App. 5.F and our codebase documentation—which allow for portable, modular, and reproducible sensitivity analysis. Our codebase is essentially a parser for these configs along with a set of fairness metrics we have implemented. Biases and metrics are designed to allow flexibility to suit particular use cases. The codebase wraps around these biases and fairness metrics and parses them into optimization problems which can be solved to produce bounds, currently using the Autobounds backend. We also provide a website that acts as a user interface for less technical users. The website allows for configs to be loaded or exported, and every element of the config can be edited via the interface. Users can also upload datasets and analyze the sensitivity to their chosen fairness metric/bias combinations.

Discussion & Limitations

In this work, we have described three prevalent measurement biases, argued that they are almost always present in FairML applications, and put forward a toolkit based on newly available methods in causal inference [62] for understanding how these biases impact parity metric evaluation. To apply this methodology, we have focused on the discrete, oblivious setting, however, causal sensitivity analysis is not limited to this domain [224, 40] and therefore its usefulness to FairML should not be either. We hope that the present work will encourage the causal inference community to look to FairML as a key application area for sensitivity analysis. Finally, additional biases that are likely topical to FairML cannot be readily expressed in this framework. For example, the issue of interference, where individuals within a classified population can affect the outcomes of others, would seem to hold in many FairML applications but can be a challenge to express graphically.

Acknowledgments

JF gratefully acknowledges funding from the EPSRC. NF thanks the Rhodes Trust for supporting their studies at Oxford, where they conducted a portion of this research. MA is indebted to the Machine Learning Department at Carnegie Mellon University and the ACMI lab for support of this work.

Appendix

5.A Disparity Metric Definitions

5.A.1 Observational Metrics

False Positive Rate Parity Definition: $\hat{Y} \perp A \mid Y = 0$

Measured as: $P(\hat{Y} = 1 \mid A = 0, Y = 0) - P(\hat{Y} = 1 \mid A = 1, Y = 0)$

False Negative Rate Parity Definition: $\hat{Y} \perp A \mid Y = 1$

Measured as: $P(\hat{Y} = 1 \mid A = 0, Y = 1) - P(\hat{Y} = 1 \mid A = 1, Y = 1)$

Positive Predictive Parity Definition: $Y \perp A \mid \hat{Y} = 1$

Measured as: $P(Y = 1 \mid A = 0, \hat{Y} = 1) - P(Y = 1 \mid A = 1, \hat{Y} = 1)$

Negative Predictive Parity Definition: $Y \perp A \mid \hat{Y} = 0$

Measured as: $P(Y = 1 \mid A = 0, \hat{Y} = 1) - P(Y = 1 \mid A = 1, \hat{Y} = 0)$

Equalized Odds Definition: $Y \perp A \mid \hat{Y}$

Measured as: $\max\left\{\text{FPR}(Y, A, \hat{Y}), \text{FNR}(Y, A, \hat{Y})\right\}$ for false positive rate (FPR) and false negative rate (FNR) given above.

5.A.2 ECP Parity Metric Definitions

Counterfactual False Positive Rate Parity Definition: $\hat{Y} \perp A \mid Y(D = 1) = 0$

Measured as: $P_{\mathcal{C}}(\hat{Y} = 1 \mid A = 0, Y(D = 1) = 0) - P(\hat{Y} = 1 \mid A = 1, Y(D = 1) = 0)$

Counterfactual False Negative Rate Parity Definition: $\hat{Y} \perp A \mid Y = 1$

Measured as: $P(\hat{Y} = 1 \mid A = 0, Y(D = 1) = 1) - P(\hat{Y} = 1 \mid A = 1, Y(D = 1) = 1)$

Counterfactual Positive Predictive Parity Definition: $Y(D=1) \perp A \mid \hat{Y}=1$

Measured as: $P(Y(D=1)=1 \mid A=0, \hat{Y}=1) - P(Y(D=1)=1 \mid A=1, \hat{Y}=1)$

Counterfactual Negative Predictive Parity Definition: $Y(D=1) \perp A \mid \hat{Y}=0$

Measured as: $P(Y(D=1)=1 \mid A=0, \hat{Y}=1) - P(Y(D=1)=1 \mid A=1, \hat{Y}=0)$

Counterfactual Equalised Odds Definition: $Y(D=1) \perp A \mid \hat{Y}$

Measured as: $\max\left\{\mathrm{CF_FPR}(Y, A, \hat{Y}), \mathrm{CF_FNR}(Y, A, \hat{Y})\right\}$ for counterfactual false positive rate ($\mathrm{CF_FPR}$) and counterfactual false negative rate ($\mathrm{CF_FNR}$) given above.

5.B Technical Description

5.B.1 Structural Causal Model definition

Definition 18. *A **structural causal model** (SCM) over the variables $\mathbf{V} = \{V_1, \cdots, V_d\}$ with latent variables $\mathbf{U} = (U_1, \cdots, U_k)$ consists of a set of structural assignments so that every variable $V_i \in \mathbf{V}$ can be written as:*

$$f_{V_i}(\mathrm{Pa}(V_i)), \;\; i = 1, ..., d,$$

*Where $\mathrm{Pa}(V_i) \subset \mathbf{V} \cup \mathbf{U}$ are the **parents** of V_i respectively, f_{V_i} is the **structural equation** for V_i, and we have a distribution $P(\mathbf{U})$ which we assume factorises as $P(\mathbf{U}) = \prod_{i=1}^{k} P(U_i)$. The **causal graph**, $\mathcal{G}$, arising from the structural causal model consists of a vertex for each variable in $\mathbf{V} \cup \mathbf{U}$, and an edge $Z \to V$ if $Z \in \mathrm{Pa}(V)$ for $Z \in \mathbf{V} \cup \mathbf{U}$ and $V \in \mathbf{V}$; we assume throughout that $\mathcal{G}$ is acyclic. Finally, letting $\mathcal{F}$ be the set of structural equations, we can denote a causal model by $\mathcal{C} = (\mathcal{F}, P(\mathbf{U}))$ and the set of all causal models over $\mathbf{V}$ as $\mathbb{M}_{\mathbf{V}}$.*

From the structural causal model we get the *observational distribution*, $P(\mathbf{V})$, by propagating the noise through the structural equations. The *potential outcomes* (counterfactuals) when intervening on a set $\mathbf{A} \subset \mathbf{V}$ are defined via recursive substitution [178], so that when intervening to set $\mathbf{A} = \mathbf{a}$ we take $U_j(\mathbf{a}) = U_j$ and $\mathbf{A}(\mathbf{a}) = \mathbf{a}$ and then defining the general potential outcome as $V_i(\mathbf{a}) \coloneqq f_{V_i}(\{Z(\mathbf{a}) | Z \in \mathrm{Pa}(V_i)\})$. For any event $\star$ involving factual or counterfactual versions of variables in $\mathbf{V}$, we use $P_{\mathcal{C}}(\star)$ to denote the probability of this event under the structural causal model $\mathcal{C}$.

5.B.2 Margnalisation in DAGs

Marginalisation Operation Suppose $\mathbf{V}$ can be split as $\mathbf{V} = \tilde{\mathbf{V}} \cup \tilde{\mathbf{U}}$ where we are interested in the causal structure over $\tilde{\mathbf{V}}$ and do not observe the variables $\tilde{\mathbf{U}}$. We start

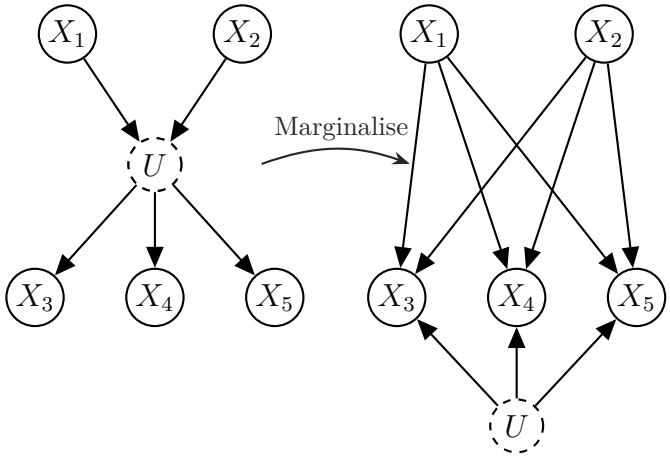


Figure 5.B.1: Example of step one in the marginalization, taken from Evans [70].

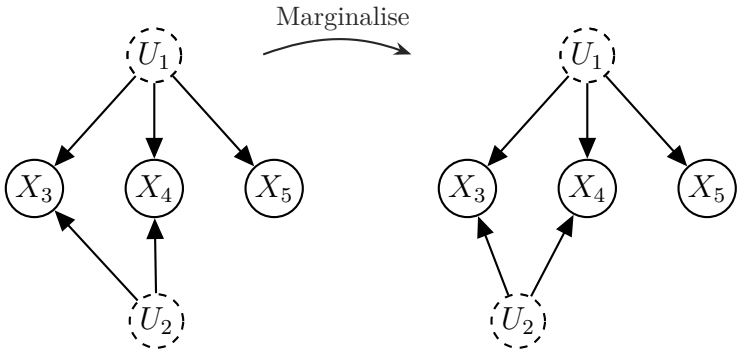


Figure 5.B.2: Example of step two in the marginalization.

from a causal graph $\mathcal{G}$, with unobserved $\tilde{\mathbf{U}}$ we marginalise to get to a graph $\mathcal{G}'$ which is of the form of Definition 18 by doing the following:

1. For all $U \in \tilde{\mathbf{U}}$, add an edge $Z \to \tilde{Z}$ if the current graph contains $Z \to U \to \tilde{Z}$ and then delete any edges $Z \to U$,
2. After completing the first step for all variables in $\tilde{\mathbf{U}}$, delete any U if there exists another $\tilde{U} \in \tilde{\mathbf{U}}$ that influences all of the variables U influences.

Evans [70] showed that there is a structural causal model over the resulting graph which preserves the causal structure over the variables $\tilde{\mathbf{V}}$. Importantly, due to the deletion step, this model has a bounded number of unobserved variables, regardless of how large the set $\tilde{U}$ is.

5.B.3 Alternative Causal Graphs for Proxy Bias

Here we provide the following result demonstrating that a wide variety of graphs can give the same outcome under proxy bias:

Proposition 26. *So long as any additional unobserved variables U' satisfy the following:*

1. *U' does not cause A.*
2. *There is no direct arrow from U' to $\hat{Y}$.*

Then marginalizing over U' will lead to the same graph as Fig. 5.5.1a.

Proof. To show this we need to demonstrate that once we have performed the marginalization operations, no additional edges or nodes will be added to the graph. We do this step by step:

1. This step will add edges if we have two vertices V, V' such that $V \to U' \to V'$. However, if neither of V, V' are $\hat{Y}$ then these vertices will already be adjacent in the graph. As the graph is acyclic that means we cannot add any edges.
2. After removing all edges in step one, we will be left so that U' has no parents and affects a subset of vertices in the graph. However, as U' does not cause A, this must be a subset of $\{\hat{Y}, Y_P, Y\}$. As these are the vertices caused by U this will lead to the deletion of U'.

□

5.C Additional Results

5.C.1 Proxy Label Results

5.C.1.1 Plots from Fogliato et al. [81] under varying assumptions

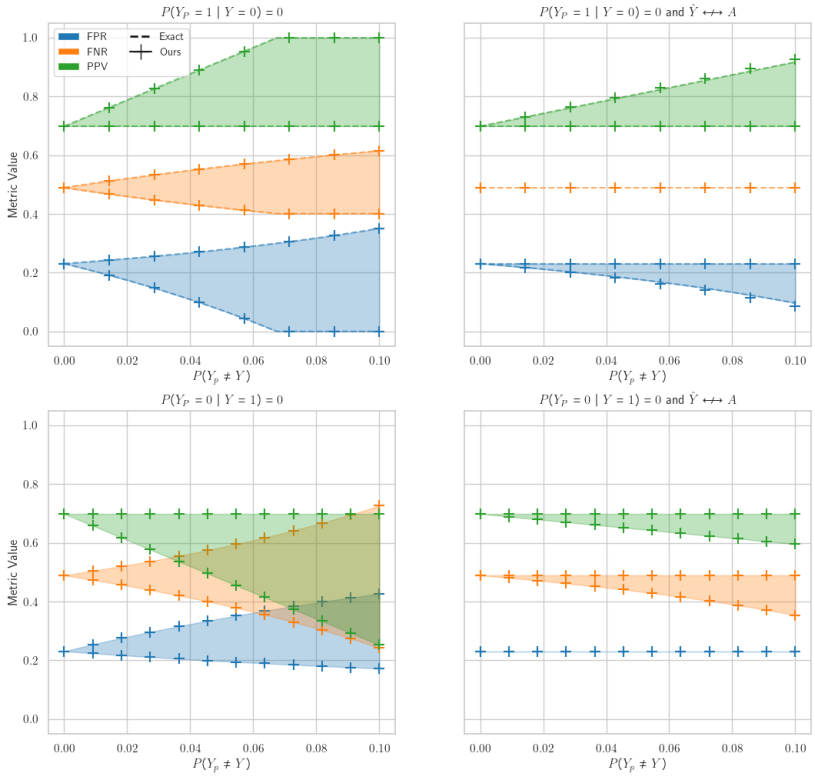


Figure 5.C.1: In this plot, we recreate the results from Fogliato et al. [81], where we are interested in the false positive rate (FPR), false negative rate (FNR), and positive predictive value (PPV) for a classifier trained on the COMPAS dataset. In this plot, we consider varying for which j we have $P(Y_P = 1 - j \mid Y = j)$, and we can see that doing so greatly changes the shape of the sensitivity set. Moreover, when we pair these assumptions by dropping of the red dashed edge in Fig. 5.5.1a we see we can identify some of the metrics of interest under any degree of bias. For $j = 1$ we identify the FNR and for $j = 0$ we identify the FPR. We prove these identification results in Appendix 5.C.1.2

.

5.C.1.2 Proxy Identification Results

In the setup of Fogliato et al. [81], the objective is to obtain the false positive/negative rate in a group $A = a$, where it is assumed that $P(Y = 1, Y_P = 0) = 0$. Now declaring

the following parameters:

$$\begin{aligned} p_{ij} &= P(Y_P = i, \hat{Y} = j \mid A = a) \\ \alpha_j &= P(Y = 1, Y_P = 0, \hat{Y} = j \mid A = a) \\ \alpha &= \alpha_0 + \alpha_1 \end{aligned}$$

Under these assumptions α_0, α_1 are sufficient to parameterise the distribution, $P(Y, Y_P, \hat{Y} \mid A = a)$. Now, following [81] we have that:

$$\begin{aligned} \text{FPR}_Y &= \frac{p_{01} - \alpha_1}{p_{00} + p_{01} - \alpha} \\ \text{FNR}_Y &= \frac{p_{10} + \alpha_0}{p_{10} + p_{11} + \alpha} \\ \text{PPV}_Y &= \frac{p_{11} + \alpha_1}{p_{01} + p_{11}} \end{aligned}$$

Now, with the absence of the dashed edge, the DAG in Fig. 5.5.1a implies the independence $\hat{Y} \perp Y_P \mid Y, A$. Therefore we get the following:

$$\begin{aligned} \alpha_j &= P(Y = 1, Y_P = 0, \hat{Y} = j \mid A = a) \\ &= \frac{P(Y = 1, Y_P = 0 \mid A = a) P(Y = 1, \hat{Y} = j \mid A = a)}{P(Y = 1 \mid A = a)} \\ &= \frac{\alpha(p_{1j} + \alpha_j)}{p_{10} + p_{11} + \alpha} \end{aligned}$$

Solving for α_j, we get $\alpha_j = \alpha \left(\frac{p_{1j}}{p_{10} + p_{11}} \right)$. Now, inputting this for α_0 in the expression for FNR_Y we get:

$$\begin{aligned} \text{FNR}_Y &= p_{10} \left(\frac{1 + \frac{\alpha}{p_{10} + p_{11}}}{p_{10} + p_{11} + \alpha} \right) \\ &= \frac{p_{10}}{p_{10} + p_{11}} \\ &= \text{FNR}_{Y_P} \end{aligned}$$

Therefore, under the assumptions given, the true false negative rate is identified and equal to the observed false negative rate on the proxy labels. Inputting the value for α_1 into FPR_Y we instead get:

$$\text{FPR}_Y = \frac{p_{01} - \alpha \left(\frac{p_{10}}{p_{10} + p_{11}} \right)}{(p_{00} + p_{01} - \alpha)}$$

As this is a decreasing function of α we can see that for $\alpha \leq \alpha_0$, FPR_Y is bounded as:

$$\frac{p_{01}}{(p_{00} + p_{01})} \leq \text{FPR}_Y \leq \frac{p_{01} - \alpha \left(\frac{p_{10}}{p_{10} + p_{11}} \right)}{(p_{00} + p_{01} - \alpha)}$$

For PPV, we again input α_1 to give:

$$\text{PPV}_Y = \frac{p_{11} + \alpha\left(\frac{p_{11}}{p_{10}+p_{11}}\right)}{p_{01} + p_{11}}$$

Leading to the bounds:

$$\text{PPV}_{Y_P} \leq \text{PPV}_Y \leq \frac{p_{11} + \alpha\left(\frac{p_{11}}{p_{10}+p_{11}}\right)}{p_{01} + p_{11}}$$

The statements for the identification of the false positive rate and false negative rate are as follows:

Proposition 27. *Suppose we have $P(Y_P = 1 \mid Y = 0) = 0$. Then under the conditional independence statement $\hat{Y} \perp Y_P \mid Y, A$, for all level of proxy bias $P(Y_P \neq Y)$:*

$$\text{FNR}_{Y|A=a} = \text{FNR}_{Y_P|A=a}$$

Where $\text{FNR}_{Y|A=a}$ is the true false negative rate for the group $A = a$ and $\text{FNR}_{Y_P|A=a}$ is the proxied false negative rate.

Proof. Follows from the above derivations. □

Now the equivalent statement for the false positive ratio:

Proposition 28. *Suppose we have $P(Y_P = 0 \mid Y = 1) = 0$. Then under the conditional independence statement $\hat{Y} \perp Y_P \mid Y, A$, for all level of proxy bias $P(Y_P \neq Y)$:*

$$\text{FPR}_{Y|A=a} = \text{FPR}_{Y_P|A=a}$$

Where $\text{FPR}_{Y|A=a}$ is the true false negative rate for the group $A = a$ and $\text{FPR}_{Y_P|A=a}$ is the proxied false negative rate.

Proof. This follows from considering the distribution where Y, Y_P and $\hat{Y}$ are all flipped as any statement about the false positive rate in the original distribution translates to a statement about the false negative rate in the flipped distribution. The assumption $P(Y_P = 0 \mid Y = 1)$ in the original distribution translates to $P(Y_P = 1 \mid Y = 0)$ in the flipped distribution, whereas all other assumptions are symmetric to the flipping operation. Therefore we can apply proposition 27 to see that the flipped FNR is constant under any degree of proxy noise. This leads us to conclude that under these assumptions the FPR in the original distribution must also be constant under any degree of proxy noise. □

5.C.2 Selection Results

5.C.2.1 Selective labels under MNAR

Here we include an experiment applying the framework to selective labels under the missing not at random assumption (MNAR)[183]. This supposes that we only see the outcome on a subset of the full dataset, with the outcome on the rest of the dataset free to vary arbitrarily. We work with the Dutch census dataset Van der Laan [206], first fitting an unconstrained logistic regression, then forming the selected population as those who have a predicted probability higher than 0.3.

Once we have formed the selected subset, we then train four classifiers, each to satisfy a different parity metric. We train to false negative rate parity, false positive rate parity, positive predictive parity, and negative predictive parity. False negative/positive rate parity are trained using the reductions approach [3], whereas for positive predictive parity and negative predictive parity, we train 100 predictors, each weighting different parts of the distribution, taking the one with the lowest parity score above a given accuracy threshold. The plots are shown in Fig. 5.C.2.

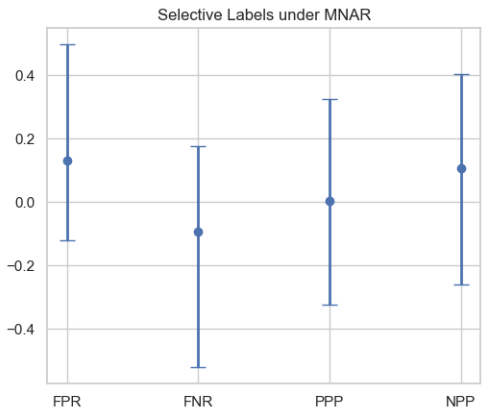


Figure 5.C.2: This plot demonstrates a sensitivity analysis for selective labels on the Dutch dataset under the missing, not at random assumption.

5.C.2.2 Selection and Proxy Plots

Here we demonstrate the effect of selection and proxy bias jointly on the Adult dataset. We include the results in Fig. 5.C.3, which show that the occurrence of multiple biases acts differently for different parity metrics.

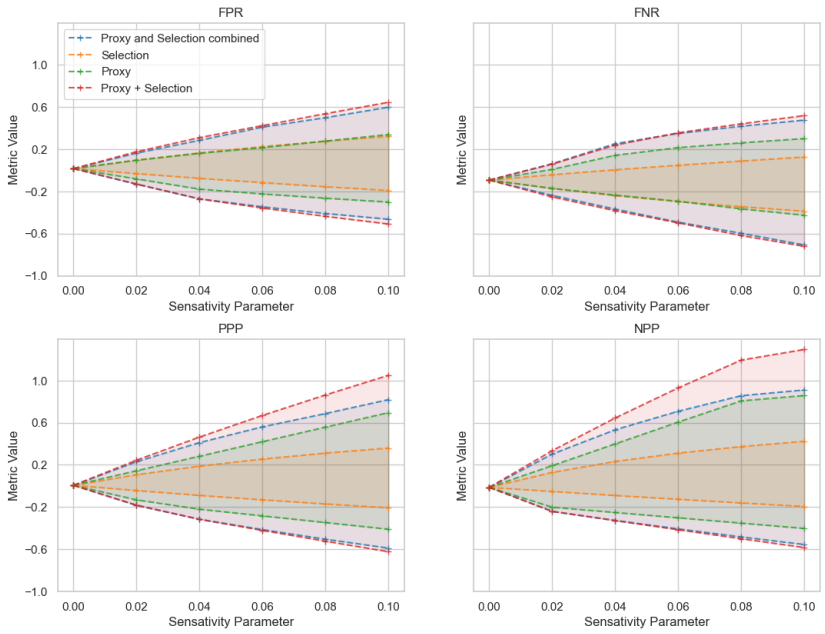


Figure 5.C.3: In these plots, we can see the effect of doing a sensitivity analysis jointly for selection and proxy bias. We can see that for the false positive rate parity (FPR) and false negative rate parity (FNR) the combined biases behave roughly as the sum of both biases, however, for positive predictive parity (PPP) and negative predictive parity (NPP) the combination behaves differently with the combined bias amounting to a smaller possible range for the metrics than the sum of the range of both biases individually.

.

5.C.3 ECP bias results

5.C.3.1 ECP experimental set up

For this experiment, we focus on finding the possible ranges for the counterfactual parity metrics from the given observational statistics. We use the sensitivity parameter of $P(Y(1) \neq Y(0))$, adding additional causal assumptions such as monotonicity ($Y(1) \geq Y(0)$) and if the policy is observed or not. When simulating the policy, we draw ECP $\sim$ Ber$(\frac{1}{2} + c * A)$ for $c = 0.2$ to skew the policy in one direction. Results show in Fig. 5.C.4

5.C.4 Causal Fairness Experiments

In this section, we deliver some results on applying our sensitivity analysis framework to causal fairness metrics of the variety detailed in [166]. Before doing so, we add some technical comments on these types of interventions in FairML and some nuances of measurement bias in the context of causal inference.

Firstly, we would like to note that our framework can still be applied to perform

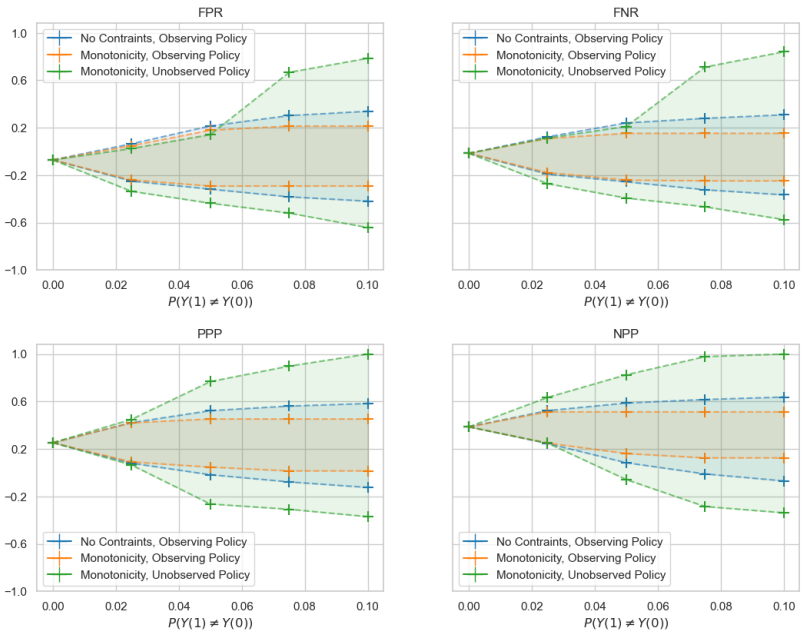


Figure 5.C.4: In these plots, we perform a sensitivity analysis for the value of the counterfactual parity metrics given in Appendix 5.A.2. In each case, we work under 3 differing levels of assumption and information

sensitivity analysis for measurement bias for other fairness metrics *without* having to consider counterfactuals relative to protected characteristics such as race or gender, thereby avoiding difficulties with intervention on such traits [125, 104, 115]. In this case, A could be seen as denoting membership to a group and indexing different graphs for each group as in Bright et al. [30]. In this case, the arrows leading from A would only express conditional independence relationships as opposed to causal ones. Notably, in the graphs, we suggest they are unconstrained.

Secondly, measurement biases and, specifically, selection bias in causal fairness entail additional complications. This is because almost always, membership in such a dataset is causally downstream of the protected attribute, meaning that when conditioning on individual presence in a dataset, we are introducing selection bias in some form. As Fawkes et al. [76] argue, this means that DAG models will be unable to correctly capture the causal structure in most datasets we come across in FairML. Failing to account for such effects can lead to erroneous causal conclusions.

Having said this, we will proceed with applying the causal graphs in Fig. 5.5.1 to do causal fairness analysis for the following metrics:

Counterfactual Fairness (CF) [128] We measure this as $P_{\mathcal{C}}(\hat{Y}(A=1) \neq \hat{Y}(A=0))$ which is equal to 0 exactly when $\hat{Y}$ is counterfactually fair [73].

Total Effect (TE) [166] Measured as $P_{\mathcal{C}}(\hat{Y}(A=1)) - P_{\mathcal{C}}(\hat{Y}(A=0))$.

Spurious Effect (SE) [166] Measured as $P_{\mathcal{C}}(\hat{Y}(A=a)) - P_{\mathcal{C}}(\hat{Y} \mid A=a)$.

Results are shown in Fig. 5.C.5, where we have assumed that counterfactual fairness is identified at a particular value. We can see that all causal fairness metrics recover a linear relationship under selection in this context.

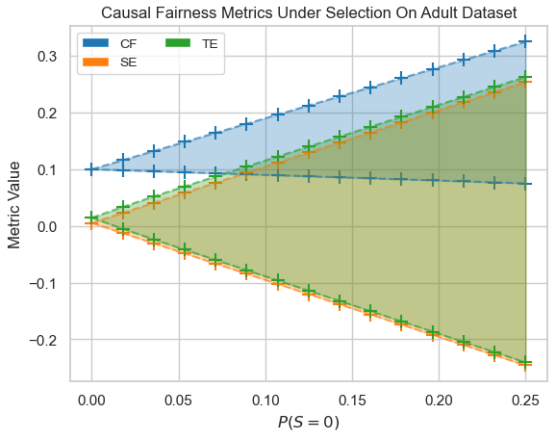


Figure 5.C.5: Causal fairness metrics under selection. We show plots for Counterfactual Fairness (CF), Total Effect (TE), and Spurious Effect (SE) using graph 5.5.1b where we have additionally assumed that counterfactual fairness is the point identified.

5.D Details of cross dataset bias analysis

In this section, we analyze the datasets presented in Le Quy et al. [132] for the three biases we present in Section 5.3. We describe each dataset, describe the task that most closely relates to the use of this dataset, and, relative to this task, we describe the measurement biases present. For each bias, we justify our decision.

Synthetic tasks Synthetic tasks are difficult to discuss, since biases are contextual and these tasks are purely theoretical. Given a downstream task, they may or might not exhibit the biases we discuss. We therefore drop them from the analysis.

Bank Marketing Dataset The goal here is to target current clients for the bank to open more accounts. Since the outcome, in this case, is exactly what the bank seeks to maximize, this dataset does not exhibit proxy or ECP bias. However, contacts were made via phone, so there is selection bias in whether customers answered the phone.

Dataset	Task	Proxy Bias	Selection Bias	ECP Bias
Adult	Synthetic			
KDD Census-Income	Synthetic			
German credit	Credit risk		✓	✓
Dutch census	Synthetic			
Bank marketing	Client		✓	
Credit card clients	Default Risk		✓	✓
COMPAS recid.	Risk prediction	✓	✓	✓
COMPAS viol. recid.	Risk prediction	✓	✓	✓
Communities&Crime	Neighborhood risk	✓	✓	✓
Diabetes	Re-admission risk	✓		✓
Ricci	Promotion Prediction	✓	✓	✓
Student-Mathematics	Admissions	✓	✓	✓
Student-Portuguese	Admissions	✓	✓	✓
OULAD	Admissions	✓	✓	✓
Law School	Admissions	✓	✓	✓

Table 5.D.1: Analysis of the datasets from Le Quy et al. [132], split by task. The explanation for the biases are given in Appendix 5.D.

German credit and Credit card clients For both of these datasets, the goal is to predict whether customers face default risk. The aim is to use this to decide if applying customers presents a risk to the bank or not. As a result, there will be selection bias, since defaults are only observed for a firms' prior customers. Finally, as with the example in the main text, this exhibits ECP bias, since the firm sets the credit limit, which impacts likelihood of default.

COMPAS recid. and COMPAS viol. recid. and Communities and Crime Datasets built off COMPAS have been well documented to exhibit all these biases and more [16]. Such issues are not unique to COMPAS, and are exhibited in all recidivism and crime prediction datasets. We see similar issues in the Communities and Crime dataset, where the aim is to predict the number of historical crimes per hundred thousand population for a number of cities. Both under-reporting of crimes and over-criminalization render the per-capita crime estimates proxy variables. Due to controversy over the reporting of rape statistics many midwestern communities were excluded, leading to sampling bias. Finally, police practices affect the likelihood of a crime being reported, and police often act in discriminatory ways, so ECP bias seems likely to be a problem as well.

Diabetes For this dataset, the goal is to predict if a patient will be readmitted in the next 30 days. The aim is to use this to assess patient health risk upon leaving the hospital, to determine if they should be discharged. The population is a sample of the

patient pool, and so there should not be selection bias. Readmissions differ from the underlying recurring illness, so this does represent a proxy, albeit a fairly reasonable one. In this case, ECP bias is a cause for concern due to the differences in quality of care by demographic group [67].

Ricci The Ricci dataset is an employment dataset, where the goal is to predict the likelihood of a promotion based on a selection of available covariates. A model trained on this data would then be used to predict the potential of applying candidates in order to decide if they are invited to interview or do additional tests. This application would fall prey to all biases we have presented and, as such, strong justification would be required as to the usefulness of the model. Going through biases one by one, proxy bias is exhibited in a similar way to the example presented in the main text, selection bias is present as the model is evaluated on a different population to the one it is trained on and, finally, the firm's policies will have an impact on who succeeds and is promoted at the company.

Admissions datasets The final datasets can all be grouped under admissions to academic institutions. Similarly to the employment example, these are prone to exhibit all the biases we have outlined. This is due to the challenges of having a perfectly objective measure of performance, deployment of models on applying populations but fit on accepted populations, and universities' policies affecting the success of students. Therefore, when using these predictors, arguments should be made as to why using such a measure would not induce demographic skew.

5.E Details of cross-dataset experiment

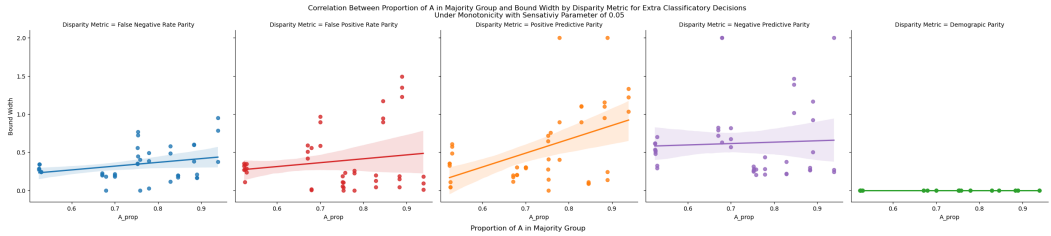
For this experiment, we train numerous predictors across a variety of common fairness benchmark datasets [132] to satisfy parity constraints. For each dataset we train 18 classifiers total, where the ML model is one of logistic regression, naïve Bayes, and a decision tree and the parity constraint is false negative rate parity, false positive rate parity, positive predictive parity and negative predictive parity, demographic parity, and equalized odds. With the exception of positive/negative predictive parity, we train all classifiers to satisfy these constraints using the reductions approach [3]. For positive/negative predictive parity, we train 100 predictors, each weighting different parts of the distribution, taking the one with the lowest parity score above a given accuracy threshold. We vary the sensitivity parameter over a range of realistic values for many real-world settings, computing the sensitivity bounds for each level of the parameter. We find that, except for demographic parity, all parity measures we evaluate exhibit significant sensitivity over these parameter ranges. This makes it hard to understand what satisfying, e.g., equalised odds, means on a given dataset. The caveat is that equalised odds is only satisfied as long as there are no significant

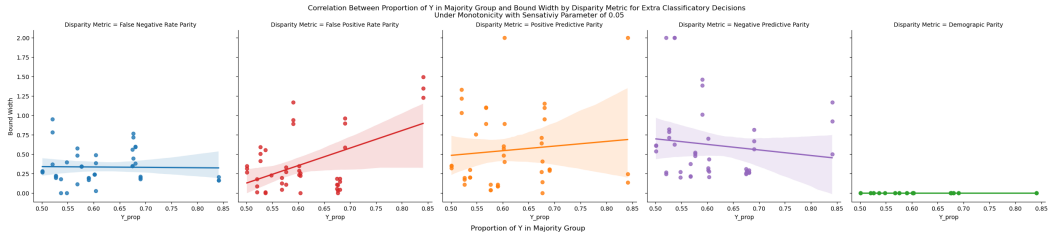
measurement biases in the underlying data, which is almost never the case in FairML audits.

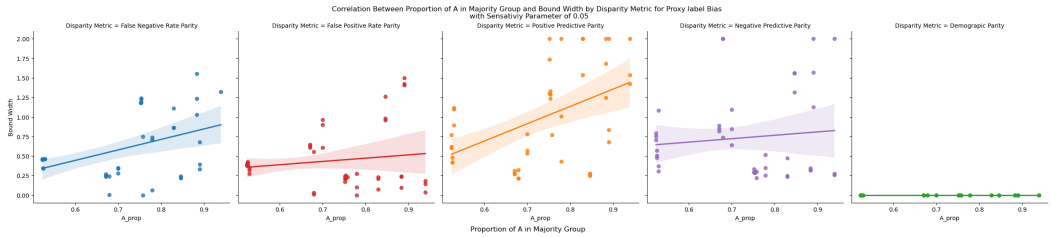
5.E.1 Analysis of Results

5.E.1.1 Correlational Plots

Here we explore how sensitivity varies according to class imbalance in either A or Y. We find that for some metrics (Negative Predictive Parity) class imbalance seems to make little difference to the sensitivity of metrics, with next to no correlation observed between imbalance and sensitivity. This sits in contrast to other metrics (Positive Predictive Parity) where we can see a clear, positive, correlation between class imbalance and sensitivity.


Correlation Between Proportion of A in Majority Group and Bound Width by Disparity Metric for Extra Classificatory Decisions
Under Monotonicity with Sensativiy Parameter of 0.05
Disparity Metric = False Negative Rate Parity
Disparity Metric = False Positive Rate Parity
Disparity Metric = Positive Predictive Parity
Disparity Metric = Negative Predictive Parity
Disparity Metric = Demograpic Parity
Bound Width
A_prop
Proportion of A in Majority Group


Correlation Between Proportion of Y in Majority Group and Bound Width by Disparity Metric for Extra Classificatory Decisions
Under Monotonicity with Sensativiy Parameter of 0.05
Disparity Metric = False Negative Rate Parity
Disparity Metric = False Positive Rate Parity
Disparity Metric = Positive Predictive Parity
Disparity Metric = Negative Predictive Parity
Disparity Metric = Demograpic Parity
Bound Width
Y_prop
Proportion of Y in Majority Group


Correlation Between Proportion of A in Majority Group and Bound Width by Disparity Metric for Proxy label Bias
with Sensativiy Parameter of 0.05
Disparity Metric = False Negative Rate Parity
Disparity Metric = False Positive Rate Parity
Disparity Metric = Positive Predictive Parity
Disparity Metric = Negative Predictive Parity
Disparity Metric = Demograpic Parity
Bound Width
A_prop
Proportion of A in Majority Group

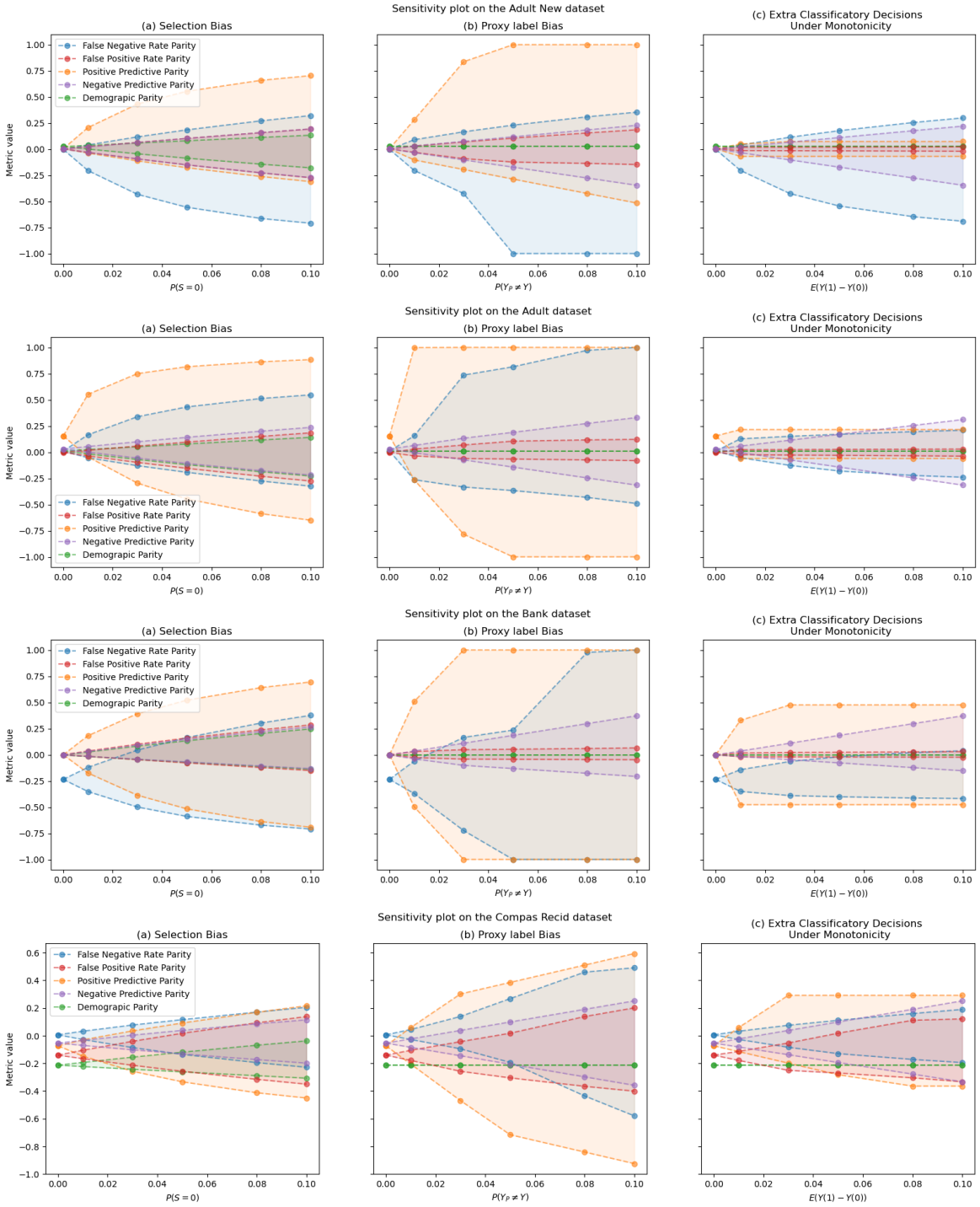

Correlation Between Proportion of Y in Majority Group and Bound Width by Disparity Metric for Proxy label Bias
with Sensativiy Parameter of 0.05
Disparity Metric = False Negative Rate Parity
Disparity Metric = False Positive Rate Parity
Disparity Metric = Positive Predictive Parity
Disparity Metric = Negative Predictive Parity
Disparity Metric = Demograpic Parity
Bound Width
Y_prop
Proportion of Y in Majority Group

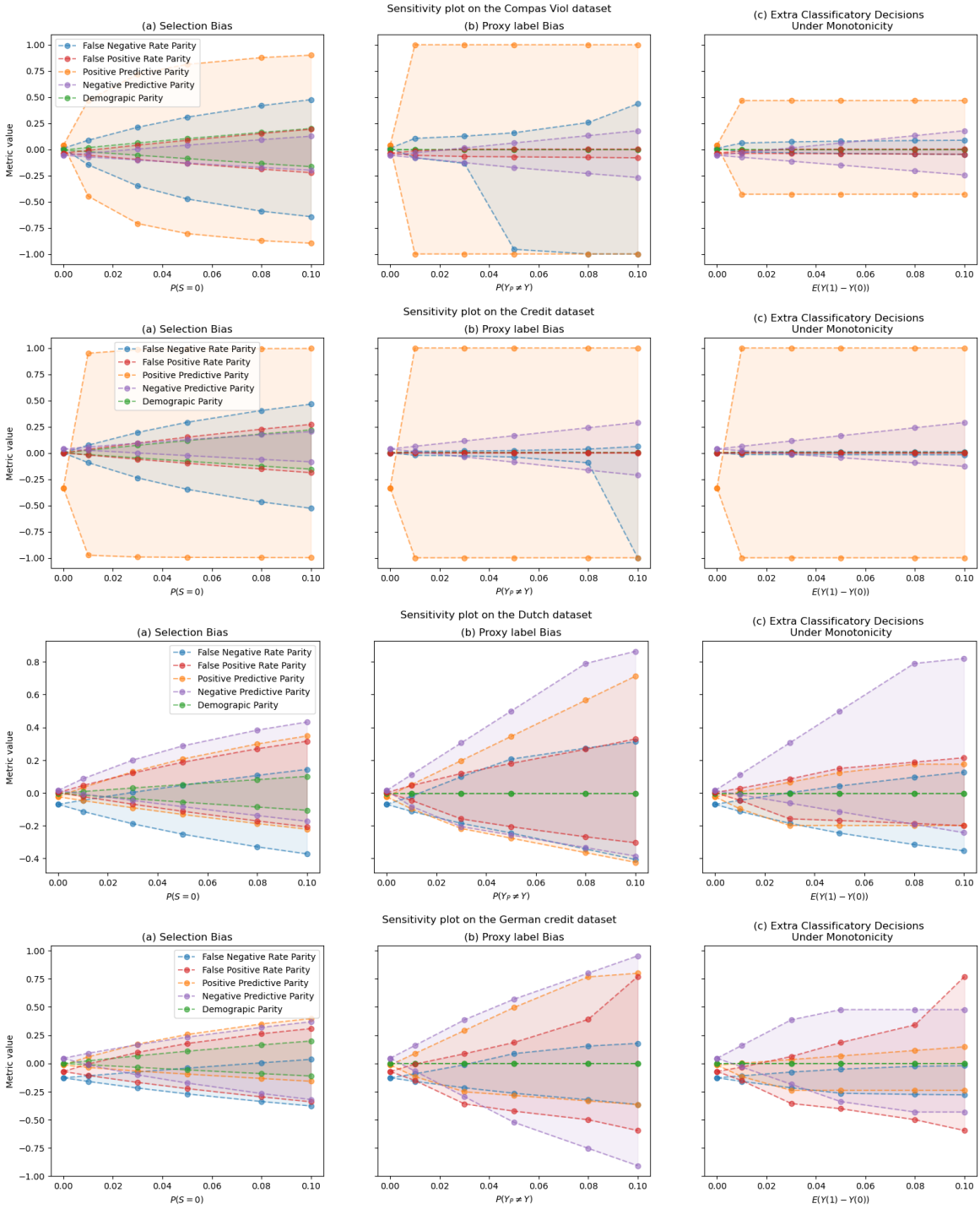

Correlation Between Proportion of A in Majority Group and Bound Width by Disparity Metric for Selection Bias
with Sensativiy Parameter of 0.05
Disparity Metric = False Negative Rate Parity
Disparity Metric = False Positive Rate Parity
Disparity Metric = Positive Predictive Parity
Disparity Metric = Negative Predictive Parity
Disparity Metric = Demograpic Parity
Bound Width
A_prop
Proportion of A in Majority Group

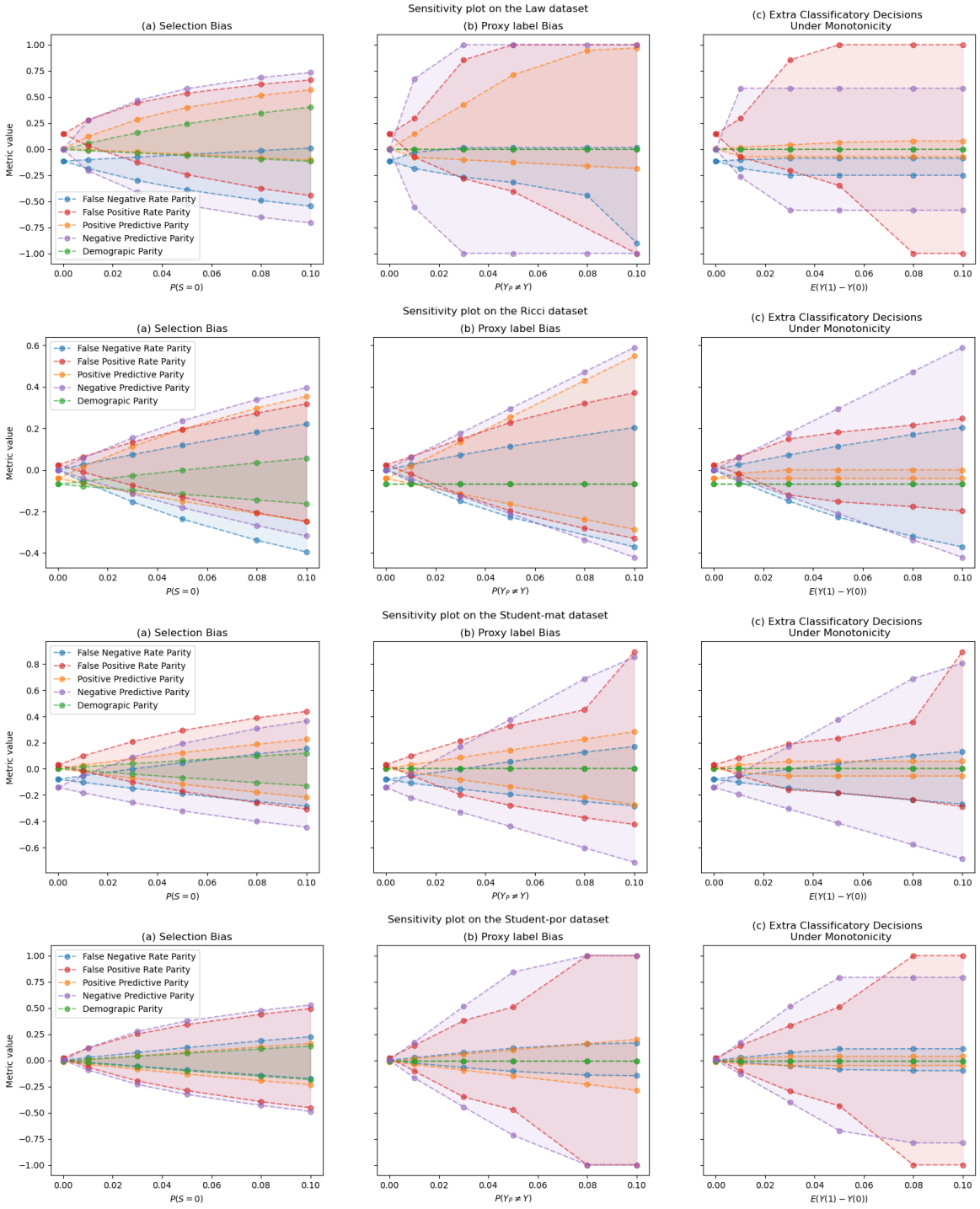

Correlation Between Proportion of Y in Majority Group and Bound Width by Disparity Metric for Selection Bias
with Sensativiy Parameter of 0.05
Disparity Metric = False Negative Rate Parity
Disparity Metric = False Positive Rate Parity
Disparity Metric = Positive Predictive Parity
Disparity Metric = Negative Predictive Parity
Disparity Metric = Demograpic Parity
Bound Width
Y_prop
Proportion of Y in Majority Group

5.E.1.2 Cross-Dataset Analysis

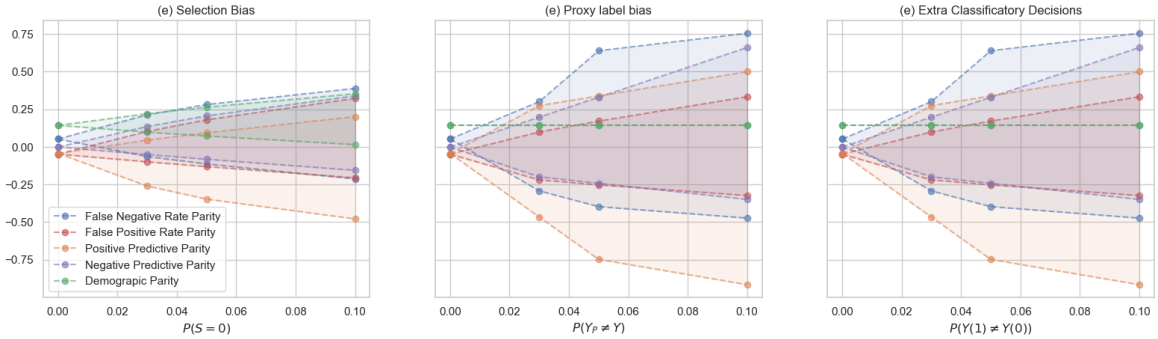
These results reveal some heterogeneity across datasets. While we do not find that parity metrics adhere to a universal ordering in terms of robustness against measurement bias, we find that these metrics can be generally grouped together with respect to their order of complexity, and that fragility to bias scales with this complexity. In particular, on the Adult New, Adult, Bank, COMPAS Recid/Viol, Credit, and Dutch datasets, we observe that positive predictive parity is a standout fragile method across biases, followed by false negative rate parity and negative predictive parity. On the German Credit, Law, and Ricci, Student mat/por we see NPP and FNRP tend to be the worst, followed by FPRP and PPP.


Sensitivity plot on the Adult New dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions Under Monotonicity
False Negative Rate Parity
False Positive Rate Parity
Positive Predictive Parity
Negative Predictive Parity
Demograpic Parity
Metric value
$P(S=0)$
$P(Y_p \neq Y)$
$E(Y(1)-Y(0))$
Sensitivity plot on the Adult dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions Under Monotonicity
Sensitivity plot on the Bank dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions Under Monotonicity
Sensitivity plot on the Compas Recid dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions Under Monotonicity


Sensitivity plot on the Compas Viol dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions Under Monotonicity
False Negative Rate Parity
False Positive Rate Parity
Positive Predictive Parity
Negative Predictive Parity
Demograpic Parity
Metric value
P(S = 0)
P(Yp ≠ Y)
E(Y(1) − Y(0))
Sensitivity plot on the Credit dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions Under Monotonicity
Sensitivity plot on the Dutch dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions Under Monotonicity
Sensitivity plot on the German credit dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions Under Monotonicity



Finally, we also perform a sensitivity analysis for a predictor trained on the folktables [59] dataset:


(e) Selection Bias
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
False Negative Rate Parity
False Positive Rate Parity
Positive Predictive Parity
Negative Predictive Parity
Demographic Parity
0.00
0.02
0.04
0.06
0.08
0.10
$P(S = 0)$
(e) Proxy label bias
0.00
0.02
0.04
0.06
0.08
0.10
$P(Y_P \neq Y)$
(e) Extra Classificatory Decisions
0.00
0.02
0.04
0.06
0.08
0.10
$P(Y(1) \neq Y(0))$

5.F Codebase and web interface

We have developed both a codebase and a web interface to ensure our framework is as usable as possible. The core of both tools are our bias configs, which allow for portability, modularity, and reproducibility.

The bias config file is a JSON file that specifies the DAG, constraints, and other parameters for the analysis. For example, the selection bias config specifies the following:

```
{
    "dag_str": "A->Y, A->P, A->S, U->P, U->Y, U->S, Y->S",
    "unob": ["U"],
    "cond_nodes": ["S"],
    "attribute_node": "A",
    "outcome_node": "Y",
    "prediction_node": "P",
    "constraints": ["P(S = 1) >= 1 - D"]
}
```

The `dag_str` field is an edgelist for the DAG. The `unob` field specifies the unobserved variables in the DAG, which are used to compute the conditional independencies between the observable variables. The `cond_nodes` field specifies the variables that are observed conditional on some value, e.g., in the selection case we only observe individuals conditional on $S = 1$.

The `attribute_node`, `outcome_node`, and `prediction_node` fields specify the nodes for the observed attribute, outcome, and prediction, which will be used in the parity metric. Usually, these will be set to "A", "Y", and "P" respectively, but some biases may require different nodes. For example, in *Proxy Y Bias*, the observed attribute Y is not the true outcome but a proxy for the true outcome, so the `outcome_node` field would be set to the true outcome node. In the *Proxy Y* config, we call this true outcome Z.

The `constraints` field specifies a set of probabilistic constraints on the variables in the DAG. The variable D is a protected variable name that is used to specify the sensitivity parameter, which is a bias-specific level of measurement bias. In the selection bias, we have $P(S = 1) \geq 1 - D$, so D is the probability of observing an individual. In Proxy Y Bias, we have $P(Z = 0 \,\&\, Y = 0) + P(Z = 1 \,\&\, Y = 1) \geq 1 - D$, so D lower bounds the probability that the observed (proxy) outcome equals the true outcome.

Our codebase is essentially a parser for these configs along with a set of fairness metrics we have implemented. Biases and metrics are designed to allow combinations to suit any particular use case. The codebase wraps around these biases and fairness metrics and parses them into optimization problems which can be solved to produce bounds, currently using the autobounds backend.

The website offers a user interface for constructing and editing bias configs. Configs can be loaded or exported, and every element of the config can be edited via the interface. Users can upload their own datasets and analyze the sensitivity of their dataset to their chosen fairness metric/bias combinations.

5.G Impact Statement

This work aims to broaden the discussion of measurement biases in FairML and provide practical tools for practitioners in the area to use. Our hope is that any potential societal consequences of the work will be positive, corresponding to more equitable algorithmic decision making.

6

Conclusion, Limitations and Future Outlook

6.1 Conclusion

To conclude, we first discuss the limitations of each of the works presented, and potential resolutions that could make up future work. We then finish by presenting a few broad conclusions to be taken from the work in this thesis.

6.1.1 The Hardness of Validating Observational Studies with Experimental Data

In terms of this paper, we identify two main limitations. Firstly, we focused on the problem of creating bands that almost surely contain the function, referred to as confidence bands in Wasserman [217]. Whilst doing this allowed us to show the impossibility of upper bounding a sensitivity analysis parameter, there are limitations to taking this view. Most notably, the problem we point to is not unique to CATE functions, as any outcome function suffers from similar impossibility results. This is unsurprising because if we could construct confidence bands for normal regression tasks, we could simply construct a band for each treatment arm and then add each of the bands' width together to construct a band for the difference between the arms. As a consequence of this, much work in nonparametric statistics considers sets which satisfy some marginal guarantee instead. By this, we mean constructing intervals such that for X sampled from a distribution $P(X)$, the bands contain the true function 95% of the time. Future work should focus on the most efficient way to construct these types of intervals for CATE, possibly drawing on the literature on conformal prediction for

CATE [5].

Secondly, whilst we isolated that smoothness is the assumption required to produce valid bands, we did not resolve this by proposing a method which bakes in exact smoothness constraints. Instead, we used a Gaussian process based approach, which imposes smoothness assumptions implicitly through the choice of kernel. A natural extension would therefore be to propose a method which would allow you to specify a smoothness parameter and directly produce the appropriate bounds, which decay to 0 as the sample size tends to infinity.

6.1.2 Is merging worth it? Securely Evaluating the Information Gain for Causal Dataset Acquisition.

For this paper, we again isolate two main limitations. Firstly, whilst this paper focuses on the problem of data merging - a task commonly done with incredibly large datasets - the methods it builds upon are more commonly employed in small sample settings, with issues scaling GPs and Bayesian causal forests. This also holds for our use of multiparty computation, as when the dataset size increases the function that is to be evaluated becomes more complicated, which creates problems in terms of computational cost and accuracy. Therefore, an important area of future exploration should focus on understanding how such methods can scale up to larger datasets.

Secondly, in order to evaluate the information gain on the causal predictions only, we required that the parameters could be split into causal and non-causal portions. This places fundamental restrictions on what models can be used within our framework. Whilst there are alternative approaches to gauging information gain in CATE predictions [108], these apply to individual data points and cannot easily be extended to datasets. Therefore, an important area of inquiry following this work would be to provide theoretically justified measures of information gain which can be applied at the dataset level when the model does not factorise.

6.1.3 Selection, Ignorability, and Challenges with Causal Fairness

For this work, we identify the main limitation as not precisely identifying the target of the intervention when discussing causal effects of protected attributes. This is a problem for the field of causal fairness as a whole, with Kasirzadeh and Smart [115] and Hu and Kohler-Hausmann [104] arguing the counterfactual contrast is not clear in these contexts. However, there have been works in the social sciences and economics [123, 122, 190] which resolve this by taking significant care to specify the causal contrast. This could be through making the distinction between race and perception of race, where we could imagine in a causal intervention on the latter in, for example, criminal justice. This could also include specifying which point in time we are imagining an

intervention, as discussed in Loftus [141]. As an example of this, in a hiring context there may be a significant difference between correcting for the causal effect of gender from birth, to correcting for the causal effect of gender in the specific hiring process. Incorporating this into the chapter, we could imagine that under different targets of intervention the assumption of an ancestrally closed protected attribute set of the claim that almost all traits are causally downstream of the protected attributes could come under question. To resolve this, the work could be extended to consider longitudinal causal effects or more complicated selection structures, including selection on unobservables.

6.1.4 The Fragility of Fairness: Causal Sensitivity Analysis for Fair Machine Learning

In this chapter, the major limitation to directly applying this methodology comes from the sensitivity analysis we make use of. Firstly, autobounds [62] relies upon solving a non-linear optimisation problem where the number of parameters grows exponentially in the number of variables. This means, it is constrained to settings with small graphs, which in turn constrains how many biases we can perform sensitivity analysis for simultaneously. To resolve this would require constraining the set of causal models in order to make the optimisation problem easier, or relaxing the optimisation problem [197]. Secondly, as autobounds is constrained to working with discrete variables, we cannot easily incorporate biases which depend on continuous values or information from continuous covariates. Again, to incorporate this would require using alternative causal sensitivity analysis tools, such as Padh et al. [160]. Having said this, the aim of this paper was to demonstrate that causal graphs can be a useful tool for reasoning about multiple sources of measurement bias simultaneously in algorithmic fairness.

6.1.5 Concluding Remarks

In this thesis, we have presented numerous works at the intersection of causal inference, machine learning, and data quality. We have demonstrated the importance of thinking about data quality when trying to estimate causal effects using machine learning, and the usefulness of causality for reasoning about measurement biases in fair ML. Through this, we hope we have demonstrated the utility of causality as a tool for reasoning about data in machine learning more generally.

Bibliography

[1] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi. Big healthcare data: preserving security and privacy. *Journal of big data*, 5(1):1–18, 2018.

[2] J. Adebayo, M. Hall, B. Yu, and B. Chern. Quantifying and mitigating the impact of label errors on model disparity metrics. *arXiv preprint arXiv:2310.02533*, 2023.

[3] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[4] A. M. Alaa and M. Van Der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

[5] A. M. Alaa, Z. Ahmad, and M. van der Laan. Conformal meta-learners for predictive inference of individual treatment effects. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] A. Almodóvar, J. Parras, and S. Zazo. Federated learning for causal inference using deep generative disentangled models. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.

[7] M. A. Alvarez, L. Rosasco, N. D. Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3): 195–266, 2012.

[8] H. Andersen. When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, 80(5):672–683, 2013.

[9] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

[10] Y. Annadani, P. Tigas, D. R. Ivanova, A. Jesson, Y. Gal, A. Foster, and S. Bauer. Differentiable multi-target causal bayesian experimental design. *arXiv preprint arXiv:2302.10607*, 2023.

[11] P. Aronow, J. M. Robins, T. Saarinen, F. Sävje, and J. Sekhon. Nonparametric identification is not enough, but randomized controlled trials are. *arXiv preprint arXiv:2108.11342*, 2021.

[12] J. Awan and A. Slavković. Structure and sensitivity in differential privacy: Comparing k-norm mechanisms. *Journal of the American Statistical Association*, 116(534):935–954, 2021.

[13] R. R. Bahadur and L. J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122, 1956.

[14] S. Balakrishnan and L. Wasserman. Hypothesis testing for densities and high-dimensional multinomials. *The Annals of Statistics*, 47(4):1893–1927, 2019.

[15] A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

[16] M. Bao, A. Zhou, S. Zottola, B. Brubach, S. Desmarais, A. Horowitz, K. Lum, and S. Venkatasubramanian. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*, 2021.

[17] E. Bareinboim and J. Tian. Recovering causal effects from selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[18] E. Bareinboim, J. Tian, and J. Pearl. Recovering from selection bias in causal and statistical inference. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[19] E. Bareinboim, J. Tian, and J. Pearl. Recovering from selection bias in causal and statistical inference. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 433–450. 2022.

[20] D. Beaver. Efficient multiparty protocols using circuit randomization. In *Advances in Cryptology—CRYPTO'91: Proceedings 11*, pages 420–432. Springer, 1992.

[21] B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[22] P. Belotti, J. Lee, L. Liberti, F. Margot, and A. Wächter. Branching and bounds tighteningtechniques for non-convex minlp. *Optimization Methods & Software*, 24(4-5):597–634, 2009.

[23] G. Bernstein and D. R. Sheldon. Differentially private bayesian linear regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[24] M. Bertanha and M. J. Moreira. Impossible inference in econometrics: Theory and applications. *Journal of Econometrics*, 218(2):247–270, 2020.

[25] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley. *Statistics and public policy*, pages 113–130, 1977.

[26] B. Bonet. Instrumentality tests revisited. *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence,*, page 226–235, 2001.

[27] S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5): 2885–2915, 2021.

[28] E. V. Bonilla, K. Chai, and C. Williams. Multi-task gaussian process prediction. *Advances in neural information processing systems*, 20, 2007.

[29] S. Bouabid, J. Fawkes, and D. Sejdinovic. Returning the favour: when regression benefits from probabilistic causal knowledge. In *International Conference on Machine Learning*, pages 2885–2913. PMLR, 2023.

[30] L. K. Bright, D. Malinsky, and M. Thompson. Causally interpreting intersectionality theory. *Philosophy of Science*, 83(1):60–81, 2016.

[31] M. L. Brodie. Data integration at scale: From relational data integration to information ecosystems. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications*, pages 2–3. IEEE, 2010.

[32] P. Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3): 404–426, 2020.

[33] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[34] Y. Byun, D. Sam, M. Oberst, Z. Lipton, and B. Wilder. Auditing fairness under unobserved confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 4339–4347. PMLR, 2024.

[35] I. A. Canay, A. Santos, and A. M. Shaikh. On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, 81(6):2535–2559, 2013.

[36] F. Castanedo et al. A review of data fusion techniques. *The scientific world journal*, 2013, 2013.

[37] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical science*, pages 273–304, 1995.

[38] H. Chang, T. D. Nguyen, S. K. Murakonda, E. Kazemi, and R. Shokri. On adversarial bias and the robustness of fair machine learning. *arXiv preprint arXiv:2006.08669*, 2020.

[39] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348, 2019.

[40] V. Chernozhukov, C. Cinelli, W. Newey, A. Sharma, and V. Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, 2022.

[41] S. Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

[42] S. Chiappa and W. S. Isaac. A causal bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*, pages 3–20. Springer, 2018.

[43] S. Chiappa, R. Jiang, T. Stepleton, A. Pacchiano, H. Jiang, and J. Aslanides. A general approach to fairness with optimal transport. In *AAAI*, pages 3633–3640, 2020.

[44] H. A. Chipman, E. I. George, and R. E. McCulloch. Bart: Bayesian additive regression trees. *Annals of Applied Statistics*, 6(1):266–298, 2012.

[45] S.-C. Chow and J.-p. Liu. Design and analysis of bioavailability and bioequivalence studies. 2008.

[46] B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39(1):165–191, 2024.

[47] J. Correa and E. Bareinboim. Causal effect identification by adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[48] J. Correa, J. Tian, and E. Bareinboim. Generalized adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[49] J. D. Correa, J. Tian, and E. Bareinboim. Identification of causal effects in the presence of selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2744–2751, 2019.

[50] A. Coston, A. Mishler, E. H. Kennedy, and A. Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 582–593, 2020.

[51] A. Coston, A. Rambachan, and A. Chouldechova. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, pages 2144–2155. PMLR, 2021.

[52] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

[53] A. Curth and M. Van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.

[54] J. Dai, S. Fazelpour, and Z. Lipton. Fair machine learning under partial compliance. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 55–65, 2021.

[55] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15, 1979.

[56] P. De Bartolomeis, J. Abad, K. Donhauser, and F. Yang. Detecting critical treatment effect bias in small subgroups. *arXiv preprint arXiv:2404.18905*, 2024.

[57] P. De Bartolomeis, J. A. Martinez, K. Donhauser, and F. Yang. Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. In *International Conference on Artificial Intelligence and Statistics*, pages 1045–1053. PMLR, 2024.

[58] R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

[59] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34: 6478–6490, 2021.

[60] A. Doan, A. Halevy, and Z. Ives. *Principles of data integration*. Elsevier, 2012.

[61] F. Dobbin, D. Schrage, and A. Kalev. Rage against the iron cage: The varied effects of bureaucratic personnel reforms on diversity. *American Sociological Review*, 80(5):1014–1044, 2015.

[62] G. Duarte, N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, (just-accepted):1–25, 2023.

[63] C. Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[64] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[65] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[66] A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

[67] L. E. Egede. Race, ethnicity, culture, and disparities in health care. *Journal of general internal medicine*, 21(6):667, 2006.

[68] European Commission. Data protection rules for the public sector. Retrieved from https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-public-sector_en, 2018.

[69] D. Evans, V. Kolesnikov, M. Rosulek, et al. A pragmatic introduction to secure multi-party computation. *Foundations and Trends® in Privacy and Security*, 2 (2-3):70–246, 2018.

[70] R. J. Evans. Graphs for margins of bayesian networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

[71] R. J. Evans. Margins of discrete bayesian networks. 2018.

[72] R. J. Evans and V. Didelez. Recovering from selection bias using marginal structure in discrete models. In *ACI@ UAI*, pages 46–55, 2015.

[73] J. Fawkes and R. J. Evans. Results on counterfactual invariance. *arXiv preprint arXiv:2307.08519*, 2023.

[74] J. Fawkes, R. Hu, R. J. Evans, and D. Sejdinovic. Doubly robust kernel statistics for testing distributional treatment effects. *Transactions on Machine Learning Research*, .

[75] J. Fawkes, L. Ter-Minassian, D. R. Ivanova, U. Shalit, and C. C. Holmes. Is merging worth it? securely evaluating the information gain for causal dataset acquisition. In *The 28th International Conference on Artificial Intelligence and Statistics*, .

[76] J. Fawkes, R. Evans, and D. Sejdinovic. Selection, ignorability and challenges with causal fairness. In *Conference on Causal Learning and Reasoning*, pages 275–289. PMLR, 2022.

[77] J. Fawkes, N. Fishman, M. Andrews, and Z. Lipton. The fragility of fairness: Causal sensitivity analysis for fair machine learning. *Advances in Neural Information Processing Systems*, 37:137105–137134, 2024.

[78] J. Fawkes, M. O'Riordan, A. Vlontzos, O. Corcoll, and C. M. Gilligan-Lee. The hardness of validating observational studies with experimental data. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=LwmhRJEt2r.

[79] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[80] C. Fiedler, C. W. Scherer, and S. Trimpe. Practical and rigorous uncertainty bounds for gaussian process regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7439–7447, 2021.

[81] R. Fogliato, A. Chouldechova, and M. G'Sell. Fairness evaluation in presence of biased noisy labels. In *International conference on artificial intelligence and statistics*, pages 2325–2336. PMLR, 2020.

[82] R. Fogliato, A. K. Kuchibhotla, Z. Lipton, D. Nagin, A. Xiang, and A. Chouldechova. Estimating the likelihood of arrest from police records in presence of unreported crimes. *The Annals of Applied Statistics*, 18(2):1253–1274, 2024.

[83] A. E. Foster. *Variational, Monte Carlo and policy-based approaches to Bayesian experimental design*. PhD thesis, University of Oxford, 2021.

[84] T. Fritz. Beyond bell's theorem: correlation scenarios. *New Journal of Physics*, 14(10):103001, 2012.

[85] D. Geiger and C. Meek. Quantifier elimination for statistical problems. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, page 226–235, 1999.

[86] A. Gelman, J. Hill, and A. Vehtari. *Regression and other stories*. Cambridge University Press, 2021.

[87] C. Gilligan-Lee. Causing trouble. *New Scientist*, 246(3279):32–35, 2020.

[88] C. M. Gilligan-Lee, C. Hart, J. Richens, and S. Johri. Leveraging directed causal discovery to detect latent common causes in cause-effect pairs. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):4938–4947, 2022.

[89] N. Goel, A. Amayuelas, A. Deshpande, and A. Sharma. The importance of modeling data missingness in algorithmic fairness: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7564–7573, 2021.

[90] L. Guerdan, A. Coston, Z. S. Wu, and K. Holstein. Ground (less) truth: A causal framework for proxy labels in human-algorithm decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 688–704, 2023.

[91] P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[92] J. Y. Halpern. *Actual causality*. MiT Press, 2016.

[93] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[94] E. Hariton and J. J. Locascio. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716, 2018.

[95] A. Hauser and P. Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

[96] J. He, S. Yalov, and P. R. Hahn. Xbart: Accelerated bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1130–1138. PMLR, 2019.

[97] C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

[98] J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[99] P. D. H. Hofmann. UCI machine learning repository, 1994. URL https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

[100] P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[101] K. D. Hoover et al. *Causality in macroeconomics*. Cambridge University Press, 2001.

[102] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[103] A. S. Householder. The numerical treatment of a single nonlinear equation. *(No Title)*, 1970.

[104] L. Hu and I. Kohler-Hausmann. What's sex got to do with fair machine learning? *arXiv preprint arXiv:2006.01770*, 2020.

[105] Z. Hussain, M.-C. Shih, M. Oberst, I. Demirel, and D. Sontag. Falsification of internal and external validity in observational studies via conditional moment restrictions. In *International Conference on Artificial Intelligence and Statistics*, pages 5869–5898. PMLR, 2023.

[106] Z. M. Hussain, M. Oberst, M.-C. Shih, and D. Sontag. Falsification before extrapolation in causal effect estimation. *Advances in Neural Information Processing Systems*, 35:6161–6174, 2022.

[107] A. Z. Jacobs and H. Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.

[108] A. Jesson, P. Tigas, J. van Amersfoort, A. Kirsch, U. Shalit, and Y. Gal. Causalbald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. *Advances in Neural Information Processing Systems*, 34: 30465–30478, 2021.

[109] O. Jeunen, C. Gilligan-Lee, R. Mehrotra, and M. Lalmas. Disentangling causal effects from sets of interventions in the presence of unobserved confounders. *Advances in Neural Information Processing Systems*, 35:27850–27861, 2022.

[110] A. Kadadi, R. Agrawal, C. Nyamful, and R. Atiq. Challenges of data integration and interoperability in big data. In *2014 IEEE international conference on big data (big data)*, pages 38–40. IEEE, 2014.

[111] N. Kallus and A. Zhou. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, pages 2439–2448. PMLR, 2018.

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

[113] N. Kallus, X. Mao, and A. Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

[114] R. Kanagavelu, Z. Li, J. Samsudin, Y. Yang, F. Yang, R. S. M. Goh, M. Cheah, P. Wiwatphonthana, K. Akkarajitsakul, and S. Wang. Two-phase multi-party computation enabled privacy-preserving federated learning. In *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pages 410–419. IEEE, 2020.

[115] A. Kasirzadeh and A. Smart. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–236, 2021.

[116] E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

[117] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017.

[118] N. Kilbertus, P. J. Ball, M. J. Kusner, A. Weller, and R. Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in Artificial Intelligence*, pages 616–626. PMLR, 2020.

[119] A. Kirsch, J. Van Amersfoort, and Y. Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

[120] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[121] B. Knott, S. Venkataraman, A. Hannun, S. Sengupta, M. Ibrahim, and L. van der Maaten. Crypten: Secure multi-party computation meets machine learning. *Advances in Neural Information Processing Systems*, 34:4961–4973, 2021.

[122] D. Knox and J. Mummolo. Making inferences about racial disparities in police violence. *Proceedings of the National Academy of Sciences*, 117(3):1261–1262, 2020.

[123] D. Knox, W. Lowe, and J. Mummolo. Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637, 2020.

[124] R. Kohavi and B. Becker. UCI machine learning repository, 1994. URL https://archive.ics.uci.edu/ml/datasets/adult.

[125] I. Kohler-Hausmann. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113:1163, 2018.

[126] N. Konstantinov and C. H. Lampert. Fairness-aware pac learning from corrupted data. *The Journal of Machine Learning Research*, 23(1):7173–7232, 2022.

[127] N. Krantsevich, J. He, and P. R. Hahn. Stochastic tree ensembles for estimating heterogeneous effects. In *International Conference on Artificial Intelligence and Statistics*, pages 6120–6131. PMLR, 2023.

[128] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30, 2017.

[129] H. Lakkaraju, J. Kleinberg, J. Leskovec, J. Ludwig, and S. Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.

[130] R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

[131] S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[132] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1452, 2022.

[133] A. Lederer, J. Umlauft, and S. Hirche. Uniform error bounds for gaussian process regression with application to safe control. *Advances in Neural Information Processing Systems*, 32, 2019.

[134] M. Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246, 2002.

[135] L. Li, Y. Fan, M. Tse, and K.-Y. Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020.

[136] N. Li, N. Goel, and E. Ash. Data-centric factors in algorithmic fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410, 2022.

[137] Y. Li, Y. Zhou, A. Jolfaei, D. Yu, G. Xu, and X. Zheng. Privacy-preserving federated learning framework based on chained secure multiparty computing. *IEEE Internet of Things Journal*, 8(8):6178–6186, 2020.

[138] X. Lin and R. J. Evans. Many data: Combine experimental and observational data through a power likelihood. *arXiv preprint arXiv:2304.02339*, 2023.

[139] D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[140] Y. Liu, H. Wang, S. Wang, Z. He, W. Xu, J. Zhu, and F. Yang. Disentangle estimation of causal effects from cross-silo data. *arXiv preprint arXiv:2401.02154*, 2024.

[141] J. R. Loftus. It's about time: counterfactual fairness and temporal depth. In *CEUR Workshop Proceedings*, volume 3442, 2023.

[142] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

[143] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[144] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

[145] D. J. MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[146] O. J. Maclaren and R. Nicholson. What can be estimated? identifiability, estimability, causal inference and ill-posed inverse problems. *arXiv preprint arXiv:1904.02826*, 2019.

[147] D. Malinsky, I. Shpitser, and T. Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3080–3088. PMLR, 2019.

[148] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[149] C. Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 411–418, 1995.

[150] M. M. Mello, J. K. Francer, M. Wilenzick, P. Teden, B. E. Bierer, and M. Barnes. Preparing for responsible sharing of clinical trial data, 2013.

[151] X.-L. Meng. Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, 12(2):685–726, 2018.

[152] A. Mishler, E. H. Kennedy, and A. Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400, 2021.

[153] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[154] J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of machine learning research*, 21(99):1–108, 2020.

[155] K. Muandet, W. Jitkrittum, and J. Kübler. Kernel conditional moment test via maximum moment restriction. In *Conference on Uncertainty in Artificial Intelligence*, pages 41–50. PMLR, 2020.

[156] T. Muazu, Y. Mao, A. U. Muhammad, M. Ibrahim, U. M. M. Kumshe, and O. Samuel. A federated learning system with data fusion for healthcare using multi-party computation and additive secret sharing. *Computer Communications*, 216:168–182, 2024.

[157] V. Mugunthan, A. Polychroniadou, D. Byrd, and T. H. Balch. Smpai: Secure multi-party computation for federated learning. In *Proceedings of the NeurIPS 2019 Workshop on Robust AI in Financial Services*, volume 21. MIT Press Cambridge, MA, USA, 2019.

[158] R. Nabi and I. Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[159] F. Niu, H. Nori, B. Quistorff, R. Caruana, D. Ngwe, and A. Kannan. Differentially private estimation of heterogeneous causal effects. In *Conference on Causal Learning and Reasoning*, pages 618–633. PMLR, 2022.

[160] K. Padh, J. Zeitler, D. Watson, M. Kusner, R. Silva, and N. Kilbertus. Stochastic causal programming for bounding treatment effects. In *Conference on Causal Learning and Reasoning*, pages 142–176. PMLR, 2023.

[161] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[162] J. Pearl. Causal inference in statistics: An overview. 2009.

[163] J. Peters, P. Bühlmann, and N. 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, 2016.

[164] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[165] J. Platt and S. Kardia. Public trust in health information sharing: implications for biobanking and electronic health record systems. *Journal of personalized medicine*, 5(1):3–21, 2015.

[166] D. Plecko and E. Bareinboim. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022.

[167] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

[168] T. Rainforth, R. Cornish, H. Yang, A. Warrington, and F. Wood. On nesting monte carlo estimators. In *International Conference on Machine Learning*, pages 4267–4276. PMLR, 2018.

[169] T. Rainforth, A. Foster, D. R. Ivanova, and F. Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

[170] I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

[171] A. Rambachan, A. Coston, and E. Kennedy. Counterfactual risk assessments under unmeasured confounding. *arXiv preprint arXiv:2212.09844*, 2022.

[172] R. R. Ramsahai and P. Spirtes. Causal bounds and observable constraints for non-deterministic models. *Journal of Machine Learning Research*, 13(3), 2012.

[173] T. Räz. Group fairness: Independence revisited. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 129–137, 2021.

[174] T. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[175] T. S. Richardson and J. 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, 2013.

[176] L. A. Rivera and A. Tilcsik. Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review*, 84(2):248–274, 2019.

[177] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

[178] J. M. Robins and T. S. Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, pages 103–158, 2010.

[179] J. M. Robins, M. A. Hernán, and U. SiEBERT. Effects of multiple interventions. *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors*, 1:2191–2230, 2004.

[180] J. P. Romano. On non-parametric testing, the uniform behaviour of the t-test, and related problems. *Scandinavian Journal of Statistics*, 31(4):567–584, 2004.

[181] P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

[182] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[183] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[184] D. B. Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2):757–763, 1997.

[185] C. Russell, M. Kusner, C. Loftus, and R. Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in neural information processing systems*, volume 30. NIPS Proceedings, 2017.

[186] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019.

[187] M. Schröder, D. Frauen, and S. Feuerriegel. Causal fairness under unobserved confounding: A neural sensitivity framework. In *The Twelfth International Conference on Learning Representations*, 2023.

[188] P. Schulam and S. Saria. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30, 2017.

[189] P. Sebastiani and H. P. Wynn. Maximum entropy sampling and optimal bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000.

[190] M. Sen and O. Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19: 499–522, 2016.

[191] R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. 2020.

[192] A. Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

[193] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

[194] I. Shpitser. *Complete identification methods for causal inference*. PhD thesis, UCLA, 2008.

[195] I. Shpitser and E. Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of Statistics*, 44(6):2433, 2016.

[196] I. Shpitser, T. S. Richardson, and J. M. Robins. Multivariate counterfactual systems and causal graphical models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 813–852. 2022.

[197] M. Shridharan and G. Iyengar. Scalable computation of causal bounds. *Journal of Machine Learning Research*, 24(237):1–35, 2023.

[198] C. Spearman. The proof and measurement of association between two things. 1961.

[199] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2239–2248, 2018.

[200] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[201] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

[202] M. J. Stensrud, J. G. Young, V. Didelez, J. M. Robins, and M. A. Hernán. Separable effects for causal inference in the presence of competing events. *Journal of the American Statistical Association*, pages 1–9, 2020.

[203] S. Tarumi, M. Suzuki, H. Yoshida, S. Miyauchi, and R. Kurazume. Personalized federated learning for institutional prediction model using electronic health records: A covariate adjustment approach. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023.

[204] B. Taskesen, V. A. Nguyen, D. Kuhn, and J. Blanchet. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*, 2020.

[205] C. Toth, L. Lorch, C. Knoll, A. Krause, F. Pernkopf, R. Peharz, and J. Von Kügelgen. Active bayesian causal inference. *Advances in Neural Information Processing Systems*, 35:16261–16275, 2022.

[206] P. Van der Laan. The 2001 census in the netherlands: Integration of registers and surveys. In *CONFERENCE AT THE CATHIE MARSH CENTRE.*, pages 1–24, 2001.

[207] G. Van Goffrier, L. Maystre, and C. M. Gilligan-Lee. Estimating long-term causal effects from short-term experiments and long-term observational data with unobserved confounding. In *Conference on Causal Learning and Reasoning*, pages 791–813. PMLR, 2023.

[208] S. Verma and J. Rubin. Fairness definitions explained. In *2018 ieee/acm International Workshop on Software Fairness (fairware)*, pages 1–7. IEEE, 2018.

[209] T. Verma. Invariant properties of causal models, 1991.

[210] T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier, 1990.

[211] T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 221–236. 2022.

[212] T. V. Vo, A. Bhattacharyya, Y. Lee, and T.-Y. Leong. An adaptive kernel approach to federated learning of heterogeneous causal effects. *Advances in Neural Information Processing Systems*, 35:24459–24473, 2022.

[213] T. V. Vo, T.-Y. Leong, et al. Federated learning of causal effects from incomplete observational data. *arXiv preprint arXiv:2308.13047*, 2023.

[214] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

[215] H. Wang, Z. Kaplan, D. Niu, and B. Li. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE conference on computer communications*, pages 1698–1707. IEEE, 2020.

[216] J. Wang, Y. Liu, and C. Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 526–536, 2021.

[217] L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[218] S. Wellek. *Testing statistical hypotheses of equivalence*. Chapman and Hall/CRC, 2002.

[219] L. F. Wightman. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. 1998.

[220] World Bank. World bank indicators. URL https://data.worldbank.org/indicator.

[221] L. Wu and S. Yang. Integrative r-learner of heterogeneous treatment effects combining experimental and observational studies. In *Conference on Causal Learning and Reasoning*, pages 904–926. PMLR, 2022.

[222] S. Wu, M. Gong, B. Han, Y. Liu, and T. Liu. Fair classification with instance-dependent label noise. In *Conference on Causal Learning and Reasoning*, pages 927–943. PMLR, 2022.

[223] Y. Wu, L. Zhang, and X. Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the twenty-eighth international joint conference on Artificial Intelligence*, 2019.

[224] K. M. Xia, Y. Pan, and E. Bareinboim. Neural causal models for counterfactual identification and estimation. In *The Eleventh International Conference on Learning Representations*, 2022.

[225] S. Yang, D. Zeng, and X. Wang. Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. *arXiv preprint arXiv:2007.12922*, 2020.

[226] A. C. Yao. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*, pages 160–164. IEEE, 1982.

[227] E. N. Zalta, U. Nodelman, C. Allen, and J. Perry. *Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Center for the Study of Language and Information ..., 1995.

[228] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[229] J. Zhang and E. Bareinboim. Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[230] Y. Zhang and Q. Long. Assessing fairness in the presence of missing data. *Advances in neural information processing systems*, 34:16007–16019, 2021.