The Sensitivity of Counterfactual Fairness to Unmeasured Confounding

Niki Kilbertus
MPI for Intelligent Systems
University of Cambridge

Philip J. Ball
University of Cambridge
Microsoft Research

Matt J. Kusner
University of Oxford
The Alan Turing Institute

Adrian Weller
University of Cambridge
The Alan Turing Institute

Ricardo Silva
University College London
The Alan Turing Institute

Abstract

Causal approaches to fairness have seen substantial recent interest, both from the machine learning community and from wider parties interested in ethical prediction algorithms. In no small part, this has been due to the fact that causal models allow one to simultaneously leverage data and expert knowledge to remove discriminatory effects from predictions. However, one of the primary assumptions in causal modeling is that you know the causal graph. This introduces a new opportunity for bias, caused by misspecifying the causal model. One common way for misspecification to occur is via *unmeasured confounding*: the true causal effect between variables is partially described by unobserved quantities. In this work we design tools to assess the sensitivity of fairness measures to this confounding for the popular class of non-linear additive noise models (ANMs). Specifically, we give a procedure for computing the maximum difference between two counterfactually fair predictors, where one has become biased due to confounding. For the case of bivariate confounding our technique can be swiftly computed via a sequence of closed-form updates. For multivariate confounding we give an algorithm that can be efficiently solved via automatic differentiation. We demonstrate our new sensitivity analysis tools in real-world fairness scenarios to assess the bias arising from confounding.

1 INTRODUCTION

Most work on fairness in machine learning focuses on discrimination against subpopulations in high-stakes decisions such as in criminal justice, lending, and insurance (Kamiran & Calders, 2009; Kamishima et al., 2012; Hardt et al., 2016; Zafar et al., 2017; Berk et al., 2017). These subpopulations are defined by one or more *protected attributes* such as race, gender, age, and sexual orientation. More recently, causal reasoning has been introduced as a valuable tool for the detection and mitigation of harmful biases and disparities in machine learning systems. Notably, it helped refine our understanding of two particular issues.

First, it has been shown by Kleinberg et al. (2016) and Chouldechova (2017) that various subsets of popular observation-based parity notions, which are based only on the joint distribution of all variables involved, can only be satisfied simultaneously in unrealistically trivial situations. This leaves us choosing among criteria when all of them seem desirable. However, since two different data generation mechanisms can give rise to the same observed joint distribution, observation-based notions cannot distinguish scenarios that may have very different fairness interpretations (Hardt et al., 2016).

Second, an earlier approach to individual fairness by Dwork et al. (2012) based on the appealing postulate that "similar individuals should be treated similarly" has proven hard to operationalize in practice. Specifically, it shifts the issue from defining what is fair to defining similarity with respect to the task at hand both between individuals as well as between outcomes.

By (a) explicitly modelling the underlying data generating mechanism with causal models, and (b) using causal primitives such as interventions and counterfactuals to formalize "similar individuals", causality provides valuable insights to resolve these conceptual roadblocks (Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Zhang & Bareinboim, 2018). In this work, we will focus on *counterfactual fairness* as introduced by Kusner et al. (2017), an individual-specific criterion aimed at answering the counterfactual question: "What would have

been my prediction if—all else held causally equal—I was a member of another protected group?". Despite the utility of such causal criteria, they are often contested, because they are based on strong assumptions that are hard to verify in practice. First and foremost, all causal fairness criteria proposed in the literature assume that the causal structure of the problem is known. Typically, one relies on domain experts and methods for causal discovery from data to construct a plausible causal graph. While it is often possible with few variables to get the causal graph approximately right, one often needs untestable assumptions to construct the full graph. The most common untestable assumption is that there is no unmeasured confounding between some variables in the causal graph. Because we cannot measure it, this confounding can introduce bias that is unaccounted for by causal fairness criteria.

In this work we propose a solution. We introduce tools to measure the sensitivity of the popular *counterfactual fairness* criterion to unmeasured confounding. Our tools are designed for the commonly used class of non-linear additive noise models (ANMs, Hoyer et al., 2008). Specifically, they describe how counterfactual fairness changes under a given amount of confounding. The core ideas here described can be adapted for sensitivity analysis of other measures of causal effect, such as the average treatment effect (ATE), itself a topic not commonly approached in the context of graphical causal models. A discussion will be left for a future journal version of this paper. Note that counterfactual fairness poses extra challenges compared to the ATE, as it requires the computation of counterfactuals in the sense of Pearl (2000). Concretely, our contributions are:

- For confounding between two variables, we design a fast procedure for estimating the worst-case change in counterfactual fairness due to confounding. It consists of a series of closed-form updates assuming linear models with non-linear basis functions. This family of models is particularly useful in graphical causal models where any given node has only few parents.
- For more than two variables, we fashion an efficient procedure that leverages automatic differentiation to estimate worst-case counterfactual fairness. In particular, compared to standard sensitivity analysis (typically applied to ATE problems, see e.g. Dorie et al., 2016), we formulate the problem in a multivariate setting as opposed to the typical bivariate case. The presence of other modeling constraints brings new challenges not found in the standard literature.
- We demonstrate that our method allows us to understand how fairness guarantees degrade based on different confounding levels. We also show that even under high levels of confounding, learning counterfactually fair predictors has lower fairness degradation than standard predictors using all features or using all features save for the protected attributes.[1]

2 BACKGROUND

2.1 CAUSALITY AND FAIRNESS

We begin by describing key background in causal inference and reviewing the notion of counterfactual fairness (Kusner et al., 2017).

Causality. We will use the *structural causal model* (SCM) framework described in Pearl (2000), and look at a popular subclass of these models called *additive noise models* (ANMs) (Hoyer et al., 2008). Specifically, an SCM is a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ model, with nodes $\mathcal{V}$ and edges $\mathcal{E}$. Each node $X \in \mathcal{V}$ is a random variable that is a non-linear function f_X of its direct parent nodes $\text{pa}_{\mathcal{G}}(X)$ in $\mathcal{G}$, plus additive error (noise) ϵ as follows: $X = f_X(\text{pa}_{\mathcal{G}}(X)) + \epsilon$. To make model fitting efficient we will consider (a) functions f_X that derive all their non-linearity from an embedding function $\boldsymbol{\phi}$ of their direct parents, and are linear in this embedding; and (b) Gaussian error (noise) ϵ so that:

$$X = \boldsymbol{\phi}(\text{pa}_{\mathcal{G}}(X))^\top \boldsymbol{w}_X + \epsilon, \quad s.t. \quad \epsilon \sim \mathcal{N}(0, \sigma_X),$$

where $\boldsymbol{w}_X$ are weights. Later on, we will consider ANMs over observed variables, where the errors may be correlated. Note that this class of ANMs is not closed under marginalization. For a more detailed analysis of the testable implications of the ANM assumption, see (Peters et al., 2017). Neither of our choices (a) and (b) are a fundamental limitation of our framework: the framework can easily be extended to general non-linear, or even non-parametric functions f_X, as well as non-Gaussian errors. In this work, we make this choice to balance flexibility and computational cost.

Counterfactual fairness. A recent definition of predictive fairness is counterfactual fairness (CF) (Kusner et al., 2017), which to facilitate exposition focuses on total effects. See (Nabi & Shpitser, 2018; Chiappa & Gillam, 2018) for an exploration of path-specific effects. Intuitively, CF states that a predictor $\hat{Y}$ of some target outcome Y gives you a fair prediction if, given that you are a member of group a (i.e., race, gender, sexual orientation), it would have given you the same prediction (in

[1] Code to reproduce the results can be found at github.com/nikikilbertus/cf-fairness-sensitivity.

probability) had you been a member of a different group a'. This is formalized via the causal notion of *counterfactuals* as follows:

$$P(\hat{Y}_{A\leftarrow a'} = y \mid X = \boldsymbol{x}, A = a) = \quad (1)$$
$$P(\hat{Y}_{A\leftarrow a} = y \mid X = \boldsymbol{x}, A = a),$$

where $\hat{Y}_{A\leftarrow a'}$ is the counterfactual prediction, imagining $A = a'$ (note that, because in reality $A = a$, we have that $\hat{Y}_{A\leftarrow a} = \hat{Y}$), and $\boldsymbol{x}$ is a realization of other variables in the causal system. In ANMs $\hat{Y}_{A\leftarrow a'}$ can be computed in four steps: 1. Fit the parameters of the causal model using the observed data: $\mathcal{D} = \{\boldsymbol{x}_i, a_i\}_{i=1}^n$; 2. Using the fitted model and data $\mathcal{D}$, compute all error variables $\boldsymbol{\epsilon}$; 3. Replace A with counterfactual value a' in all causal model equations; 4. Using parameters, error variables, and a', recompute all variables affected (directly or indirectly) by A, and recompute the prediction $\hat{Y}$. To learn a CF predictor satisfying eq. (1) it is sufficient to use any variables that are non-descendants of A, such as the error variables $\boldsymbol{\epsilon}$ (Kusner et al., 2017). Loftus et al. (2018) provide further arguments for the importance of causality in fairness as well as a review of existing methods.

Unmeasured confounding. One key assumption on which CF relies is that there is no *unmeasured confounding* relationship missing in the causal model. In this work, we formalize unmeasured confounding as non-zero correlations between any two error variables in $\boldsymbol{\epsilon}$ which are assumed to follow a multivariate Gaussian distribution. Without accounting for this, the above counterfactual procedure will compute error variables that are not guaranteed to be independent of A. Thus any predictor trained on these exogenous variables is not guaranteed to satisfy counterfactual fairness eq. (1). This setup captures the idea that often we have a decent understanding of the causal structure, but might overlook confounding effects, here in the form of pairwise correlations of noise variables. At the same time, such confounding is often unidentifiable (save for specific parameterizations). Thus assessing confounding is not a model selection problem but a sensitivity analysis problem. To perform such analysis we propose tools to measure the worst-case deviation in CF due to unmeasured confounding. Before describing these tools, we first place them in the context of the long tradition of sensitivity analysis in causal modeling.

2.2 TRADITIONAL SENSITIVITY ANALYSIS

Sensitivity analysis, for quantities such as the average treatment effect, can be traced back at least to the work by Jerome Cornfield on the General Surgeon study concerning the smoking and lung cancer link (Rosenbaum, 2002). Rosenbaum cast the problem in a more explicit statistical framework, addressing the question on how the ATE would vary if some degree of association between a treatment and a outcome was due to unmeasured confounding. The logic of sensitivity analysis can be described in a simplified way as follows: i) choose a level of "strength" for the contribution of a latent variable to the structural equation(s) of the treatment and/or outcome; ii) by fixing this confounder contribution, estimate the corresponding ATE; iii) vary steps i) and ii) through a range of "confounding effects" to report the level of unmeasured confounding required to make the estimate ATE be statistically indistinguishable from zero; iv) consult an expert to declare whether the level of confounding required for that to happen is too strong to be plausible, and if so conclude that the effect is real to the best of one's knowledge. This basic idea has led to a large literature, see (Dorie et al., 2016; Robins et al., 2000) as examples among many of the existing state-of-the-art papers on this topic.

Note the crucial difference between sensitivity analysis and just fitting a latent variable model: we are not learning a latent variable distribution, as the confounding effect for a single cause-effect pair is *unidentifiable*. By holding the contribution of the confounder as constant and known, the remaining parameters become identifiable. We can vary the sensitivity parameter without assuming a probability measure on the confounding effect. The hypothesis test mentioned in the example above can be substituted by other criteria of practical significance.

Much of the work in the statistics literature on sensitivity analysis addresses pairs of cause-effects as opposed to a causal system with intermediate outcomes, and focuses on the binary question on when an effect is non-zero. The *grid search* idea of attempting different levels of the confounding level does not necessarily translate well to a full SCM: grid search grows exponentially with the number of pairs of variables. In our problem formulation described in the sequel, we are interested in bounding the maximum magnitude $p_{\max}$ of the error correlation matrix entries, while maximizing a measure of counterfactual unfairness to understand how it varies by the presence of unmeasured confounding. The solution is not always to set all entries to $p_{\max}$, since among other things we may be interested in keeping a subset of error correlations to be zero. In this case, a sparse correlation matrix with all off-diagonal values set to either 0 or $p_{\max}$ is not necessarily positive-definite. A multidimensional search for the entries of the confounding correlation matrix is then necessary, which we will do in Section 4 by encoding everything as a fully differentiable and unconstrained optimization problem.

3 TOOL #1: GRID-BASED

The notion of sensitivity analysis in a SCM can be complex, particularly when the estimated quantity involves counterfactuals. In this section, we first describe a tool that estimates the effect of confounding on counterfactual fairness, when the confounding is limited to two variables (i.e., *bivariate confounding*). This procedure is computationally efficient for this setting. For the general setting of confounding between any number of variables (*multivariate confounding*) we will introduce a separate tool in Section 4. Below we describe our fast two-variable tool using a real-world example.

3.1 A MOTIVATING EXAMPLE

To motivate our approach, let us revisit the example about law school success analyzed by Kusner et al. (2017). In this task, we want to predict the first year average grade (Y) of incoming law school students from their grade-point average (G) before entering law school and their law school admission test scores (L). In the original work, the goal was to train a predictor $\hat{Y}$ that was counterfactually fair with respect to race.

To evaluate any causal notion of fairness, we need to first specify the causal graph. Here we assume $G \to L$ with errors ϵ_G, ϵ_L, where G and L are both influenced by the sensitive attribute A, see **Model A** in Figure 1. Given this specification, the standard way to train a counterfactually fair classifier is using ϵ_G, ϵ_L—the non-descendants of A. To do so, we first learn them from data as the residuals in predicting G and L from their parents.

The validity of causal estimates rely on the assumption that the constructed causal model and its respective graph (here Model A) captures the true data-generating mechanism. While previous work addressed how to enforce counterfactual fairness across a small enumeration of identifiable competing models (Russell et al., 2017), in this work we consider misspecification in the lack of *unidentifiable* unmeasured confounding. In our example, this means violation of the assumed independence of the error variables ϵ_G and ϵ_L.

To capture such confounding, we introduce **Model B** in Figure 1. Here the error variables are not independent, they co-vary: $(\epsilon_G, \epsilon_L)^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where,

$$\Sigma = \begin{pmatrix} \sigma_G^2 & p\,\sigma_G\,\sigma_L \\ p\,\sigma_G\,\sigma_L & \sigma_L^2 \end{pmatrix}.$$

Here, $\sigma_\bullet$ is the standard deviation of $\bullet$ and $p \in [-1, 1]$ is the correlation, such that the overall covariance matrix Σ is positive semi-definite. Before going into the detailed procedure of our sensitivity analysis, let us give a general description of what we mean by Model A and Model B throughout this work.

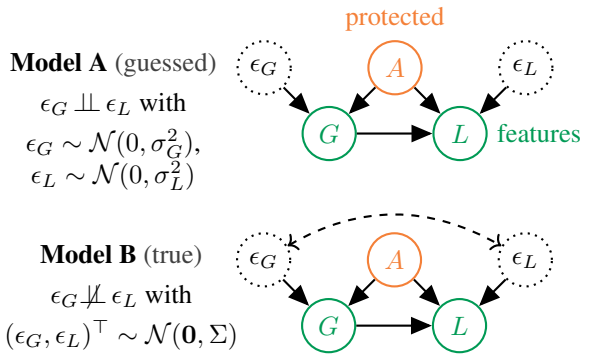

Figure 1: Causal models for the law school example. Model A is the guessed model that has no unobserved confounding. Model B includes confounding via the covariance matrix Σ, which is captured by a bidirected edge using the standard acyclic directed mixed graph notation (ADMG, Richardson, 2003). Our techniques will estimate the worst case difference in the estimation of counterfactual fairness due to such confounding (we will consider a more complicated setup in Section 5).

Model A is the "guessed" causal graph model used to build a counterfactually fair predictor.

Model B is a version of Model A that allows for further unobserved confounding between pairs of error variables not originally featured in A. Model B will play the role of a hypothetical ground truth that simulates "true" counterfactual versions of the predictions based on Model A.

Our tool allows us to answer the following question: how does a predictor that is counterfactually fair under Model A perform in terms of counterfactual unfairness under the confounded Model B? Our goal is to quantify how sensitive counterfactual unfairness is to misspecifications of the causal model, in particular to unobserved confounding. To do so, we will introduce a measure which we will call *counterfactual unfairness* (CFU). Given this, we describe how to compute the worst-case violation of counterfactual fairness within a certain confounding budget, which we characterize by the correlation $-1 \leq p_{\max} \leq 1$ in Model B. By varying the confounding budget, we can assess how robust Model A is to different degrees of model misspecification. Like in classical sensitivity analysis, we can alternatively start from a level of unacceptable CFU, search for the minimum $p_{\max}$ whose worst-case CFU reaches this level, and leave it to domain experts to judge the plausibility of such a degree

of unmeasured confounding $p_{\max}$.

3.2 NOTATION AND PROBLEM SETUP

For both Model A and B the model equations are:

$$G = \boldsymbol{\phi}_G(A)^\top \boldsymbol{w}_G + \epsilon_G, \quad L = \boldsymbol{\phi}_L(A, G)^\top \boldsymbol{w}_L + \epsilon_L, \tag{2}$$

where $\boldsymbol{\phi}_G : \mathcal{A} \to \mathbb{R}^{d_G}$ and $\boldsymbol{\phi}_L : \mathcal{A} \times \mathbb{R} \to \mathbb{R}^{d_L}$ denotes *fixed* embedding functions for A and A, G respectively, $A \in \mathcal{A}$ indicates the membership in a protected group (where $\mathcal{A}$ is the set of possible groups), and $\boldsymbol{w}_G \in \mathbb{R}^{d_G}$, $\boldsymbol{w}_L \in \mathbb{R}^{d_L}$ are the weights of the model.

In order to simplify notation, for observed data $\{(a_i, g_i, l_i)\}_{i=1}^n$, we define

$$\begin{aligned} \boldsymbol{x}_i &= \begin{pmatrix} g_i \\ l_i \end{pmatrix} \in \mathbb{R}^2, \quad \boldsymbol{w} = \begin{pmatrix} \boldsymbol{w}_G \\ \boldsymbol{w}_L \end{pmatrix} \in \mathbb{R}^{d_G + d_L}, \\ \Phi_i &= \begin{pmatrix} \boldsymbol{\phi}_{G_i}^\top & \mathbf{0}^\top \\ \mathbf{0}^\top & \boldsymbol{\phi}_{L_i}^\top \end{pmatrix} \in \mathbb{R}^{2 \times (d_G + d_L)}, \end{aligned} \tag{3}$$

where we write $\boldsymbol{\phi}_{G_i} = \boldsymbol{\phi}_G(a_i)$ and $\boldsymbol{\phi}_{L_i} = \boldsymbol{\phi}_L(a_i, g_i)$ for brevity. In eq. (3) as well as the remainder of this work, equations and assignments with subscripts i on both sides hold for all $i \in \{1, \ldots, n\}$.[2]

3.3 MODEL A: FIT CF PREDICTOR

First, we build a counterfactually fair predictor with our guessed unconfounded Model A via the following steps.

1. Fit Model A via regularized maximum likelihood:

$$\begin{aligned} \min_{\boldsymbol{w}, \sigma_G, \sigma_L} \sum_{i=1}^n (\boldsymbol{x}_i - \Phi_i \boldsymbol{w})^\top \Sigma^{-1} (\boldsymbol{x}_i - \Phi_i \boldsymbol{w}) \\ + \lambda \|\boldsymbol{w}\|_2^2 + n \log\det(\Sigma), \end{aligned} \tag{4}$$

where

$$\Sigma = \begin{pmatrix} \sigma_G^2 & 0 \\ 0 & \sigma_L^2 \end{pmatrix}.$$

Note that we can alternately solve for $\boldsymbol{w}$ and σ_G, σ_L as follows. First fix $\sigma_G = \sigma_L = 1$ and compute

$$\tilde{\boldsymbol{w}}^\dagger = \left(\sum_{i=1}^n \Phi_i^\top \Phi_i + \lambda \boldsymbol{I} \right)^{-1} \left(\sum_{i=1}^n \Phi_i^\top \boldsymbol{x}_i \right).$$

The optimal standard deviations σ_G, σ_L are then simply given by the empirical standard deviations of the residuals under $\tilde{\boldsymbol{w}}^\dagger$. Thus, the optimum of eq. (4) is

$$\boldsymbol{w}^\dagger = \left(\sum_{i=1}^n \Phi_i^\top \Sigma^{-1} \Phi_i + \lambda \boldsymbol{I} \right)^{-1} \left(\sum_{i=1}^n \Phi_i^\top \Sigma^{-1} \boldsymbol{x}_i \right),$$

where $\Sigma = \mathrm{diag}(\sigma_G^2, \sigma_L^2)$.

2. Given fitted weights $\boldsymbol{w}^\dagger$, estimate the errors ϵ_G, ϵ_L,

$$\hat{\boldsymbol{\epsilon}}_i \equiv (\hat{\epsilon}_{g_i}, \hat{\epsilon}_{l_i})^\top \equiv \boldsymbol{x}_i - \Phi_i \boldsymbol{w}^\dagger.$$

3. Fit a counterfactually fair predictor $\hat{y}_i \equiv f_{\boldsymbol{\theta}}(\hat{\boldsymbol{\epsilon}}_i)$ with parameters $\boldsymbol{\theta}$ to predict outcomes y_i via

$$\boldsymbol{\theta}^\dagger = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n \mathcal{L}(f_{\boldsymbol{\theta}}(\hat{\boldsymbol{\epsilon}}_i), y_i),$$

for some loss function $\mathcal{L}$. While virtually any predictive model can be used in the two-variable case, in the general case we require the counterfactually fair predictor to be differentiable, such that it is amenable to gradient-based optimization. The definition of counterfactual fairness constrains the optimization for any loss function. Here, we use the sufficient condition for counterfactual fairness that the predictor $\hat{y}$ depends only on the error terms, which are non-descendants of A (Kusner et al., 2017).

3.4 MODEL B: EVALUATE CFU

Next, we evaluate how the predictor $f_{\boldsymbol{\theta}^\dagger}$ obtained in the previous section breaks down in the presence of unobserved confounding, i.e., in Model B. To do so, we fit Model B and generate "true" counterfactuals $\boldsymbol{x}'$. If we were handed these counterfactuals and we wanted to make predictions using $f_{\boldsymbol{\theta}^\dagger}$ we would compute their error terms $\hat{\boldsymbol{\epsilon}}'$ using Step 2 above. If Model A was in fact the model that generated the counterfactuals $\boldsymbol{x}'$ then the predictions on the error terms for the real data and the counterfactuals would be *identical*: $f_{\boldsymbol{\theta}^\dagger}(\hat{\boldsymbol{\epsilon}}) = f_{\boldsymbol{\theta}^\dagger}(\hat{\boldsymbol{\epsilon}}')$.

However, because the counterfactuals were generated by the true weights $\boldsymbol{w}^*$ of Model B, not the weights $\boldsymbol{w}^\dagger$ of Model A, there will be a difference between the real data and counterfactual predictions $f_{\boldsymbol{\theta}^\dagger}(\hat{\boldsymbol{\epsilon}}) \neq f_{\boldsymbol{\theta}^\dagger}(\hat{\boldsymbol{\epsilon}}')$. It is this discrepancy we will quantify with our measure of counterfactual unfairness (CFU). Here is how we compute it for a given confounding budget $p_{\max}$.

1. Fit model B via regularized maximum likelihood:

$$\begin{aligned} \min_{\boldsymbol{w}, \sigma_G, \sigma_L} \sum_{i=1}^n (\boldsymbol{x}_i - \Phi_i \boldsymbol{w})^\top \Sigma^{-1} (\boldsymbol{x}_i - \Phi_i \boldsymbol{w}) \\ + \lambda^\dagger \|\boldsymbol{w}\|_2^2 + n \log\det(\Sigma), \end{aligned} \tag{5}$$

where

$$\Sigma \equiv \begin{pmatrix} \sigma_G & 0 \\ 0 & \sigma_L \end{pmatrix} \underbrace{\begin{pmatrix} 1 & p_{\max} \\ p_{\max} & 1 \end{pmatrix}}_{P} \begin{pmatrix} \sigma_G & 0 \\ 0 & \sigma_L \end{pmatrix}.$$

[2] Note that A need not be exogenous. Since we would need to include additional—standard but occluding—steps in the algorithm to handle discrete variables, this assumption is solely to simplify the presentation.

As before we can alternately solve for $\boldsymbol{w}$ (closed-form) and σ_G, σ_L (via coordinate descent).[3] Let $\boldsymbol{w}^*$ be the final weights after optimization.

2. Given weights $\boldsymbol{w}^*$, estimate the errors of Model B,

$$\hat{\boldsymbol{\delta}}_i = (\hat{\delta}_{g_i}, \hat{\delta}_{l_i})^\top = \boldsymbol{x}_i - \Phi_i \boldsymbol{w}^*.$$

3. For a fixed counterfactual value $a' \in \mathcal{A}$, compute the Model B counterfactuals of G and L for all i,

$$g'_i = \boldsymbol{\phi}_G(a'_i)^\top \boldsymbol{w}^*_G + \hat{\delta}_{g_i}\ ,$$
$$l'_i = \boldsymbol{\phi}_L(a'_i, g'_i)^\top \boldsymbol{w}^*_L + \hat{\delta}_{l_i},$$

where $\boldsymbol{w}^* = (\boldsymbol{w}^*_G, \boldsymbol{w}^*_L)^\top$. If $\boldsymbol{x}'_i \equiv (g'_i, l'_i)^\top$, we can write the above equation as

$$\boldsymbol{x}'_i = \Phi'_i \boldsymbol{w}^* + \hat{\boldsymbol{\delta}}_i,$$

and $\Phi'_i \equiv \text{diag}(\boldsymbol{\phi}_G(a'_i), \boldsymbol{\phi}_L(a'_i, g'_i))$ is defined in general by sequential propagation of counterfactual values according to the ancestral ordering of the SCM.

4. Compute the (incorrect) error terms of the counterfactuals using the same procedure as in step 2 of Section 3.3, using weights $\boldsymbol{w}^\dagger$ of Model A:

$$\hat{\boldsymbol{\epsilon}}'_i = (\hat{\epsilon}_{g'_i}, \hat{\epsilon}_{l'_i})^\top = \boldsymbol{x}'_i - \Phi'_i \boldsymbol{w}^\dagger.$$

Again, the predictions on the above quantity $f_{\theta^\dagger}(\hat{\boldsymbol{\epsilon}}'_i)$ will differ from those made on the real-data error terms $f_{\theta^\dagger}(\hat{\boldsymbol{\epsilon}}_i)$ (unless the counterfactuals were also generated according to model A).

5. To measure the discrepancy, we propose to quantify counterfactual unfairness as the squared difference between the above two quantities:

$$\text{CFU}_i = (f_{\boldsymbol{\theta}^\dagger}(\hat{\boldsymbol{\epsilon}}_i) - f_{\boldsymbol{\theta}^\dagger}(\hat{\boldsymbol{\epsilon}}'_i))^2.$$

Ultimately, to summarize the aggregate unfairness, we will compute the average counterfactual unfairness:

$$\text{CFU} = \frac{1}{n} \sum_{i=1}^{n} \text{CFU}_i\,. \tag{6}$$

A quick note: in the two-variable setting, given a confounding budget $p_{\max}$, the worst-case CFU occurs precisely at $p_{\max}$ (which need not be the case for multivariate confounding as we show in Appendix A). Thus, the above procedure computes the maximum CFU with bivariate confounding budget equal to $p_{\max}$. CFU measures how the counterfactual responses $\hat{Y}(a)$ and $\hat{Y}(a')$, defined using model A, differ "in reality", i.e., if model B is "true". What qualifies as bad CFU is problem dependent and requires interaction with domain experts, who can make judgment calls about the plausibility of the misspecification $p_{\max}$ that is required to reach a breaking point. Here, a breaking point could be the CFU of a predictor that completely ignores the causal graph.

To summarize: we learn $\hat{Y} \equiv f_{\boldsymbol{\theta}^\dagger}$ as function of X and A, where X and A are implicit in the expression of the (estimated) error terms $\boldsymbol{\epsilon}$ that are computed using the assumptions of the working Model A. We assess how "unfair" $\hat{Y}$ is by comparing for each data point the two counterfactual values $\hat{Y}(a) \equiv f_{\boldsymbol{\theta}^\dagger}(\hat{\boldsymbol{\epsilon}}_i)$ and $\hat{Y}(a') \equiv f_{\boldsymbol{\theta}^\dagger}(\hat{\boldsymbol{\epsilon}}'_i)$ where the "true" counterfactual is generated according to the world assumed by Model B. The space of models to which Model B belongs is a continuum indexed by $p_{\max}$, which will allow us to visualize the sensitivity of Model A by a one-dimensional curve. We will do this by finding the best fitting model (in terms of structural equation coefficients and error variances) at different values of $p_{\max}$, so that the corresponding CFU measure is determined by $p_{\max}$ only (results on the above law school model are shown in Section 5). We assume that the free confounding parameter is not identifiable from data (as it would be the case if the model was linear and the edge $A \to L$ was missing, the standard instrumental variable scenario).

4 TOOL #2: OPTIMIZATION-BASED

In this section, we generalize the procedure outlined for the two-variable case in Section 3 to the general case.

4.1 NOTATION AND PROBLEM SETUP

Besides the protected attribute A and the target variable Y, let there be m additional observed feature variables X_j in the causal graph $\mathcal{G}$ each of which comes with an unobserved error term variable ϵ_j.

As before, we express the assignment of the structural equations for a specific realization of observed features $\boldsymbol{x} = (x_1, \ldots, x_m)^\top$ and error terms $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)^\top$ as the following operation, i.e., $\boldsymbol{x} = \Phi \boldsymbol{w} + \boldsymbol{\epsilon}$. Here Φ has m rows and $d = \sum_{V \in \text{has-parents}(\mathcal{G})} d_V$ columns, where d_V is the dimensionality of embedding $\phi_V : \mathbb{R}^{|\text{pa}_\mathcal{G}(V)|} \to \mathbb{R}^{d_V}$ for each node $V \in \mathcal{G}$ that has parent nodes. Without loss of generality, we assume the nodes $\{A\} \cup \{X_j\}_{j=1}^m$ to be topologically sorted with respect to $\mathcal{G}$ with A always being first. We combine the individual weights as, $\boldsymbol{w} = (\boldsymbol{w}_{X_1}, \boldsymbol{w}_{X_2}, \ldots, \boldsymbol{w}_{X_m}) \in \mathbb{R}^d$, and represent Φ once evaluated on a specific sample $(a, \boldsymbol{x})$ of the vari-

[3] In fact we optimize $\log(\sigma_G), \log(\sigma_L)$ to ensure the standard deviations are positive.

ables $A, X_1, \ldots, X_m$ as,

$$\Phi = \begin{pmatrix} \boldsymbol{\phi}_{X_1}^\top & & \mathbf{0}^\top \\ & \ddots & \\ \mathbf{0}^\top & & \boldsymbol{\phi}_{X_m}^\top \end{pmatrix},$$

where $\boldsymbol{\phi}_{X_j}$ is based on the parents of X_j, $(a, \boldsymbol{x}_{\mathrm{pa}_G(X_j)})$. The covariance matrix of the error terms is given by

$$\Sigma \equiv \mathrm{diag}(\sigma_1, \ldots, \sigma_m) P \,\mathrm{diag}(\sigma_1, \ldots, \sigma_m),$$

where $\sigma_1, \ldots, \sigma_m$ are the standard deviations of each variable and P is a correlation matrix.

4.2 THE OPTIMIZATION PROBLEM

In the general case our goal is to find a correlation matrix P that satisfies a "confounding budget" $p_{\max}$. In particular we would like to constrain the correlation P_{jk} between any two different variables X_j and X_k for $j \neq k$, while allowing $P_{jj} = 1$ for all j. Additionally, we want to take into account any prior knowledge that certain variable pairs should have no correlation, if available. The most intuitive way to budget the amount of confounding is to limit the absolute size of any correlation by $p_{\max}$ as: $|P_{jk}| \leq p_{\max}$ for all $j \neq k$. This captures a notion of "restricted unobserved confounding" and leads to the following optimization problem

$$\begin{aligned} \max_P \quad & \sum_{i=1}^{n} \mathrm{CFU}_i \qquad (7) \\ \text{subject to} \quad & P_{jj} = 1 \quad \text{for } j \in \{1, \ldots, m\}, \\ & |P_{jk}| \leq p_{\max} < 1 \quad \text{for all } (j,k) \in \mathcal{C}, \\ & P_{jk} = 0 \quad \text{for all } (j,k). \notin \mathcal{C}, j \neq k. \end{aligned}$$

$\mathcal{C}$ is the set of correlations that should be non-zero. A quick aside: the setting where there are zero correlations can be captured using the standard acyclic directed mixed graph notation (ADMG, Richardson, 2003). Specifically, this can be represented by ADMGs by removing bidirected edges between any two error terms whose correlation is fixed to zero.

As in the bivariate case, CFU is a direct function of P only: all other parameters will be determined given the choice of correlation matrix by maximizing likelihood. Note that eq. (7) contains multiple nested optimization problems (for the counterfactually fair model weights $\boldsymbol{\theta}^\dagger$, and the weights and standard deviations of Models A and B). To solve it efficiently, we will parameterize P in a way that facilitates optimization via off-the-shelf, unconstrained, automatic differentiation tools. We provide more details about computational bottlenecks in Appendix B.

4.3 ALGORITHM

We use the following approach to accommodate the constraints in eq. (7) in a way such that our algorithm does not require a constrained optimization subroutine. Assume first that P has no correlations that should be zero. We compute $LL^\top$ for a matrix $L \in \mathbb{R}^{m \times m}$, whose entries are the parameters we eventually optimize. To constrain the off-diagonals to a given range and ensure that P has 1s on the diagonal, we define P as,

$$P := \tanh_{p_{\max}}(LL^\top) := \boldsymbol{I} + p_{\max}\,(\boldsymbol{J} - \boldsymbol{I}) \odot \tanh(LL^\top),$$

where $\odot$ denotes element-wise multiplication of matrices and $\boldsymbol{J}$ is a matrix of all ones. This way P is symmetric, differentiable w.r.t. the entries of L, has 1s on the diagonal, and its off-diagonal values are squashed to lie within $(-p_{\max}, p_{\max})$. While it is natural to directly mask and clamp the diagonal, there are various ways to squash the off-diagonals to a fixed range in a smooth way, which bears close resemblance to barrier methods in optimization. We choose $\tanh()$ because of its abundance in ML literature, but other forms of P may work better for specific applications. Note that this formulation does not guarantee P to be positive-semidefinite.

In Algorithm 1, we describe our procedure to maximize counterfactual unfairness given a confounding budget $p_{\max}$ and observational data $\{\boldsymbol{x}_i, y_i, a_i, \}_{i=1}^n \subset \mathbb{R}^m \times \{0,1\}^2$. The algorithm closely follows the procedure described in Section 3 for the bivariate case. Since we use automatic differentiation provided by PyTorch (Paszke et al., 2017) to obtain gradients, we only show the forward pass in Algorithm 1. For the initialization INITIALIZEPARAMETERS(), we simply populate L as a lower triangular matrix with small random values for the off-diagonals and 1s on the diagonal.

If $\mathcal{E}$ indicates some correlations should be zero, we suggest the following standard "clique parameterization": L is a $m \times c$ matrix where c is the number of cliques in $\mathcal{E}$, with L_{ik} being a non-zero parameter if and only if vertex i is in clique k. $L_{ik} \equiv 0$ otherwise. It follows that such a matrix will have zeros at precisely the locations not in $\mathcal{E}$.[4] See Silva et al. (2007) and Barber (2009) for examples of applications of this idea. For large cliques, further refinements are necessary to avoid unnecessary constraints, such as creating more than one row per clique of size four or larger. In the interest of space, details are left for an expanded version of this paper and our experiments will not make use of sparse P (note that this parameterization also assumes that the number of cliques is tractable). Note that individual parameters L_{ik} may not be identifiable, but identifiability is not necessary here, all we care

[4] Barring unstable parameter cancellations that have measure zero under continuous measures on $\{L_{ik}\}$.

about is the objective function: CFU. As a matter of fact, multiple globally optimal solutions are to be expected even in the space of P transformations. A more direct parameterization of sparse P, with exactly one parameter per non-zero entry of the upper covariance matrix, is discussed by Drton & Richardson (2004). Computationally, this minimal parameterization does not easily lead to unconstrained gradient-based algorithms for optimizing sparse correlation matrices with bounded entries. We suggest the clique parameterization as a pragmatic alternative. Special cases may be treated with more efficient specialized approaches. See (Cinelli et al., 2019) for a thorough discussion of fully linear models.

In Section 5 we will demonstrate this approach on a 3-variable-confounding scenario to showcase our approach. As this paper is aimed at describing the methodology, we will leave more complex confounding scenarios to an extended version of this work and only briefly describe an extension to path-specific effects in Appendix C.

5 EXPERIMENTS

We compare the grid-based and the optimization-based tools introduced in Sections 3 and 4 on two real datasets.

In all experiments our embedding ϕ is a polynomial basis up to a fixed degree. The degree is determined via cross validation (5-fold) jointly with the regularization parameter $\lambda^\dagger$. Our counterfactually fair predictor is regularized linear regression on the noise terms $\boldsymbol{\epsilon}$:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \boldsymbol{\phi}(\hat{\boldsymbol{\epsilon}}_i)^\top \boldsymbol{\theta})^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \,.$$

For this model, counterfactual unfairness is:

$$\mathrm{CFU}_i = \left(\left(\boldsymbol{\phi}(\hat{\boldsymbol{\epsilon}}_i) - \boldsymbol{\phi}(\hat{\boldsymbol{\epsilon}}'_i) \right)^\top \boldsymbol{\theta}^\dagger \right)^2 .$$

For comparison, we also train two baselines that also use regularized ridge regression (degree and regularization are again selected by 5-fold cross-validation):

unconstrained: an unconstrained predictor using all observed variables as input $f_{\mathrm{uc}} : (A, X_1, \ldots, X_m) \mapsto Y$.

blind unconstrained: an unconstrained predictor using all features, but not the protected attribute, as input $f_{\mathrm{buc}} : (X_1, \ldots, X_m) \mapsto Y$.

Analogous to our definition of CFU in eq. (6), we compute the unfairness of these baselines as the mean squared difference between their predictions on the observed data and the predictions of the counterfactually fair predictor on the observed data: $\frac{1}{n} \sum_{i=1}^{n} (f_{\boldsymbol{\theta}^\dagger}(\hat{\boldsymbol{\epsilon}}_i) - \hat{y}_i^{(\mathrm{b})\mathrm{uc}})^2$, where $\hat{y}_i^{\mathrm{uc}} = f_{\mathrm{uc}}(a_i, \boldsymbol{x}_i)$ and $\hat{y}_i^{\mathrm{buc}} = f_{\mathrm{buc}}(\boldsymbol{x}_i)$. This choice is motivated by the fact that in practice we care about how much potential predictions deviate from predictions satisfying a fairness measure. For our grid-based approach we repeatedly fix $p_{\max} \in [0, 1)$ to a particular value and then use the procedure in Section 3 to compute CFU. For the optimization approach we similarly fix $p_{\max} \in [0, 1)$ in the constraint of eq. (7). For efficiency we use the previously found correlation matrix P as initialization for the next setting of $p_{\max}$.

Algorithm 1 MAXCFU: Maximize counterfactual unfairness under a certain confounding budget constraint.

Require: data $\{\boldsymbol{x}_i, y_i, a_i, \}_{i=1}^n$, confounding budget $p_{\max}$, learning rate α, minibatch size B

$\{\hat{\boldsymbol{\epsilon}}_i\}_{i=1}^n, \boldsymbol{w}^\dagger, \boldsymbol{\theta}^\dagger \leftarrow$ FITMODELA$(\{\boldsymbol{x}_i, y_i, a_i, \}_{i=1}^n)$
$\mathcal{D} \leftarrow \{\boldsymbol{x}_i, y_i, a_i, \Phi_i, \hat{\boldsymbol{\epsilon}}_i\}_{i=1}^n$ ▷ full dataset
$L \leftarrow$ INITIALIZEPARAMETERS()
for $t = 1 \ldots T$ **do** ▷ iterations
  $\mathcal{D}^{(t)} \leftarrow$ SAMPLEMINIBATCH$(\mathcal{D}, B)$
  $\Delta \leftarrow \nabla_L \mathrm{CFU}(\mathcal{D}^{(t)}, \boldsymbol{w}^\dagger, \frac{B}{n}\lambda^\dagger, \boldsymbol{\theta}^\dagger, L)$ ▷ autodiff
  $L \leftarrow L + \alpha\Delta$ ▷ gradient ascent step
return $\mathrm{CFU}(\mathcal{D}, \boldsymbol{w}^\dagger, \lambda^\dagger, \boldsymbol{\theta}^\dagger, L)$

function FITMODELA$(\{\boldsymbol{x}_i, y_i, a_i, \}_{i=1}^n)$
  $\boldsymbol{w}^+ \leftarrow \left(\sum_{i=1}^n \Phi_i^\top \Phi_i + \lambda^\dagger \boldsymbol{I} \right)^{-1} \left(\sum_{i=1}^n \Phi_i^\top \boldsymbol{x}_i \right)$
  $\Sigma \leftarrow \mathrm{diag}(\mathrm{var}(\{\boldsymbol{x}_i - \Phi_i \boldsymbol{w}^+\}_{i=1}^n))$
  $\boldsymbol{w}^\dagger \leftarrow \left(\sum_{i=1}^n \Phi_i^\top \Sigma^{-1} \Phi_i + \lambda^\dagger \boldsymbol{I} \right)^{-1} \times \left(\sum_{i=1}^n \Phi_i^\top \Sigma^{-1} \boldsymbol{x}_i \right)$
  $\hat{\boldsymbol{\epsilon}}_i \leftarrow \boldsymbol{x}_i - \Phi_i \boldsymbol{w}^\dagger$
  $\boldsymbol{\theta}^\dagger \leftarrow \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n \mathcal{L}(f_{\boldsymbol{\theta}}(\hat{\boldsymbol{\epsilon}}_i), y_i)$
  return $\{\hat{\boldsymbol{\epsilon}}_i\}_{i=1}^n, \boldsymbol{w}^\dagger, \boldsymbol{\theta}^\dagger$

function CFU$(\mathcal{D}, \boldsymbol{w}^\dagger, \lambda^\dagger, \boldsymbol{\theta}^\dagger, L)$
  $\boldsymbol{w}^*, \boldsymbol{\sigma}^* \leftarrow \min_{\boldsymbol{w}, \boldsymbol{\sigma}} \sum_{i=1}^n (\boldsymbol{x}_i - \Phi_i \boldsymbol{w})^\top \Sigma^{-1} \times (\boldsymbol{x}_i - \Phi_i \boldsymbol{w}) + \lambda^\dagger \|\boldsymbol{w}\|_2^2 + n \log\det(\Sigma)$
  where $\Sigma = \mathrm{diag}(\boldsymbol{\sigma}) \tanh_{p_{\max}}(LL^\top) \mathrm{diag}(\boldsymbol{\sigma})$
  $\hat{\boldsymbol{\delta}}_i \leftarrow \boldsymbol{x}_i - \Phi_i \boldsymbol{w}^*$
  $a'_i \leftarrow 1 - a_i$ and $\boldsymbol{x}'_i \leftarrow \Phi'_i \boldsymbol{w}^* + \hat{\boldsymbol{\delta}}_i$
  where Φ'_i is computed via iterative assignment
  $\hat{\boldsymbol{\epsilon}}'_i \leftarrow \boldsymbol{x}'_i - \Phi'_i \boldsymbol{w}^\dagger$
  $\mathrm{CFU} \leftarrow \frac{1}{n} \sum_{i=1}^n (f_{\boldsymbol{\theta}^\dagger}(\hat{\boldsymbol{\epsilon}}_i) - f_{\boldsymbol{\theta}^\dagger}(\hat{\boldsymbol{\epsilon}}'_i))^2$
  return CFU

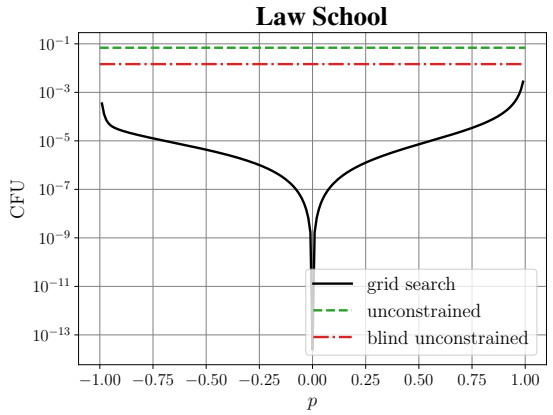

Figure 2: Counterfactual unfairness for the law school dataset. See text for details.

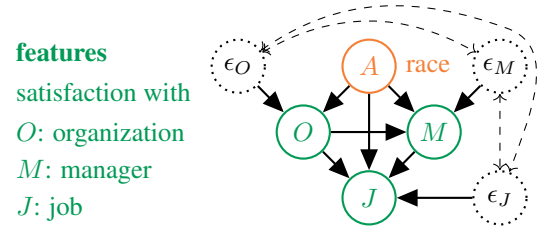

Figure 3: The true causal graph (Model B) for the NHS Staff Survey dataset.

Law School data. Our first experiment is on our motivating example introduced in Section 3 on law school success (recall eq. (2) and Figure 1 for details on the causal models). Our data comes from law school students in the US (Wightman, 1998). As our causal model investigates confounding between two variables we will use the grid-based approach introduced in Section 3 to calculate the maximum CFU. Recall that for the bivariate approach we fix a confounding level $p = p_{\max}$ and then compare predictions between real data based on Model A versus counterfactuals generated from Model B. Figure 2 shows the CFU for the grid-based approach (black), alongside the baselines (green/red), as the correlation p varies. We first note that the confounding is not symmetric around $p = 0$. For the law school data, negative correlations have smaller CFU. In general, this is a data-specific property.

Additionally, we notice that as $p_{\max}$ moves away from 0 it increases noticeably, then plateaus in roughly $[0.1, 0.9]$ and finally increases again. Our suspicion is that the initial jump may be due to a small model misspecification. Specifically, a small change in $p_{\max}$ may cause the generated counterfactuals to have additional error which may dominate for such small $p_{\max} \leq 0.1$ and then becomes insignificant for larger $p_{\max}$. For large $p_{\max} \geq 0.9$ we

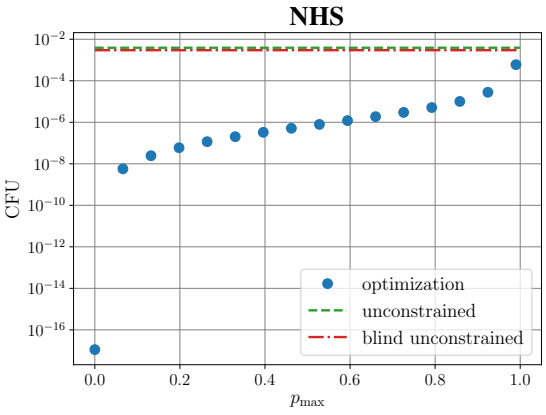

Figure 4: Counterfactual unfairness as a function of $p_{\max}$ for the multivariate NHS dataset.

believe the increase may be due to numeric instability as the covariance matrix becomes nearly negative definite. This could cause the weights of the model to rapidly grow or shrink. Between the small and large regimes, we see the CFU gradually increase as $p_{\max}$ is increased. Finally, we note that both baseline approaches have higher CFU than found with any grid-based setting.

NHS Staff Survey. Our second experiment is based on the 2014 UK National Health Service (NHS) Survey Picker Institute Europe (2015). The goal of the survey was to "gather information that will help to improve the working lives of staff in the NHS". Answers to the survey questions ranged from 'strongly disagree' (1) to 'strongly agree' (5). We averaged survey answers for related questions to create a dataset of continuous indices for: *job satisfaction* (J), *manager satisfaction* (M), *organization satisfaction* (O), and *overall health* (Y). The goal is to predict health Y based on the remaining information. Additionally, we collected the race (A) of the survey respondents. Using this data, we formulate a ground-truth causal graph shown in Figure 3 (equivalent to Model B in Figure 1). This causal graph includes correlations between all error terms $\epsilon_J, \epsilon_M, \epsilon_O$. This model has the following structural equations

$$\begin{aligned} O &= \boldsymbol{\phi}_O(A)^\top \boldsymbol{w}_O + \epsilon_O \\ M &= \boldsymbol{\phi}_M(A, O)^\top \boldsymbol{w}_M + \epsilon_M \\ J &= \boldsymbol{\phi}_J(A, O, M)^\top \boldsymbol{w}_J + \epsilon_J. \end{aligned} \tag{8}$$

Just as in the law school example, we measure the impact of this confounding by comparing this model to the unconfounded model (i.e., all error terms are jointly independent). As there is no general efficient way to grid-search for positive definite matrices that maximize CFU for a given $p_{\max}$, we make use of our optimization-based

procedure for calculating maximum CFU, as described in Algorithm 1. Figure 4 shows the results of our method on the NHS dataset. Note that we only show positive $p_{\max}$ because our optimization problem eq. (7) only constrains the absolute value of the off-diagonal correlations. This allows the procedure to learn whether positive or negative correlations result in greater CFU. As in the law school dataset we see an initial increase in CFU for small $p_{\max}$, followed by a plateau, ending with another small increase. As before, all values have lower values than the two baseline techniques.

6 CONCLUSION

In this work we presented two techniques to assess the impact of unmeasured confounding in causal additive noise models. We formulated unmeasured confounding as covariance between error terms. We then introduced a grid-based approach for confounding between two terms, and an optimization-based approach for confounding in the general case. We demonstrated our approach on two real-world fairness datasets. As a next step, we plan to write an extended version of this work with experiments on larger graphs with known zero correlations. We would also like to extend these approaches to handle the sensitivity of other quantities such as the structural equations. Overall, we believe the tools in this work are an important step towards making causal models suitable to address discrimination in real-world prediction problems.

Acknowledgments

We thank Chris Russell for useful discussions during the initial phase of this project. AW, MK, RS acknowledge support from the David MacKay Newton research fellowship at Darwin College, The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074, and the Leverhulme Trust via the CFI.

References

Barber, D. Identifying graph clusters using variational inference and link to covariance parametrization. *Philosophical Transactions of the Royal Society A*, 367:4407–4426, 2009.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint:1703.09207*, 2017.

Chiappa, S. and Gillam, T. Path-specific counterfactual fairness. *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Cinelli, C., Kumor, D., Chen, B., Pearl, J., and Bareinboim, E. Sensitivity analysis of linear structural causal models. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Dorie, V., Harada, M., Carnegie, N. B., and Hill, J. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine*, 35(20):3453–3470, 2016.

Drton, M. and Richardson, T. Iterative conditional fitting for Gaussian ancestral graph models. *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, pp. 130–137, 2004.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, pp. 214–226. ACM, 2012.

Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323, 2016.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21:689–696, 2008.

Kamiran, F. and Calders, T. Classifying without discriminating. In *International Conference on Computer, Control and Communication*, pp. 1–6. IEEE, 2009.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, 30:656–666, 2017.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint:1609.05807*, 2016.

Kusner, M., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30:4066–4076, 2017.

Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

Nabi, R. and Shpitser, I. Fair inference on outcomes. *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

Picker Institute Europe. National Health Service national staff survey, 2014. 2015. URL `http://doi.org/10.5255/UKDA-SN-7776-1`.

Richardson, T. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30: 145–157, 2003.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, 2000.

Rosenbaum, P. *Observational Studies*. Springer, 2002.

Russell, C., Kusner, M., Loftus, J., and Silva, R. When worlds collide: integrating different counterfactual assumptons in fairness. *Advances in Neural Information Processing Systems*, 30:6417–6426, 2017.

Shpitser, I. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, 37:10111035, 2013.

Silva, R., Chu, W., and Ghahramani, Z. Hidden common cause relations in relational learning. *Advances in Neural Information Processing Systems*, 20:1345–1352, 2007.

Wightman, L. F. LSAC national longitudinal bar passage study. *LSAC Research Report Series.*, 1998.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *World Wide Web Conference*, 2017.

Zhang, J. and Bareinboim, E. Fairness in decision-making: The causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

A COUNTEREXAMPLE

In this section we show that in the multivariate setting, the worst-case counterfactual unfairness with a confounding budget of $p_{\max}$ is not necessarily obtained when all non-zero entries of the correlation matrix are set to $p_{\max}$. To this end, it suffices to find a symmetric matrix A with 1s on the diagonal that is not positive-semidefinite when all its non-zero off-diagonal entries are set to the same value, which we define to be the considered confounding budget $p_{\max}$. Since each valid correlation matrix must be positive-semidefinite, the correlation matrix for the worst-case counterfactual unfairness must be different from A (while maintaining the zero entries). Because all off-diagonal entries are upper bounded by $p_{\max}$, at least one of them must be smaller than the corresponding value in A.

For example, consider

$$A = \begin{pmatrix} 1 & p_{\max} & p_{\max} \\ p_{\max} & 1 & 0 \\ p_{\max} & 0 & 1 \end{pmatrix}.$$

Since the eigenvalues of A are 1, $1 - \sqrt{2}p_{\max}$, and $1 + \sqrt{2}p_{\max}$, we see that A is not positive-semidefinite for $p_{\max} > 1/\sqrt{2}$.

In general, the matrix $A \in \mathbb{R}^{n \times n}$ with $A_{ii} = 1$ for $i \in \{1, \ldots, n\}$, $A_{1i} = A_{i1} = p_{\max}$ for $i \in \{2, \ldots, n\}$ and $A_{ij} = 0$ for all remaining entries, has the eigenvalues (without multiplicity) 1, $1 - \sqrt{n-1}p$, and $1 + \sqrt{n-1}p$. Therefore, A is not positive-semidefinite for $p_{\max} > 1/\sqrt{n-1}$. We conclude that as the dimensionality of the problem increases, we may encounter such situations for ever smaller confounding budget.

B COMPUTATIONAL CONSIDERATIONS

Step 17 of Algorithm 1 is the main place where code optimization can take place, and alternatives to the (local) penalized maximum likelihood taking place there could be suggested (perhaps using spectral methods). It is hard though to say much in general about Step 20, as counterfactual fairness allows for a large variety of loss functions usable in supervised learning. In the case of linear predictors, it is still a non-convex problem due to the complex structure of the correlation matrix, and for now we leave as an open problem whether non-gradient based optimization may find better local minima.

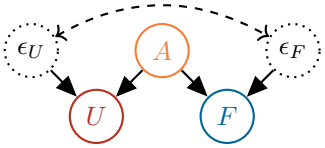

Figure 5: A path-specific model where the path from protected attribute A to feature U is unfair and the path from A to feature F is fair.

C PATH-SPECIFIC SENSITIVITY

Path-specific effects were not originally described by Kusner et al. (2017) as the goal there was to introduce the core idea of counterfactual fairness in a way as accessible as possible (some discussion is provided in the supplementary material of that paper). See Chiappa & Gillam (2018) for one take on the problem, and Loftus et al. (2018) for another take to be fully developed in a future paper. Here we consider an example that illustrates how notions of path-specific effects (Shpitser, 2013) can be easily pipelined with our sensitivity analysis framework.

Consider Figure 5, where the path from $A \to U$ is considered unfair and $A \to F$ is considered fair, in the sense that we do not want a non-zero path-specific effect of A on $\hat{Y}$ that is comprised by a possible path $A \to U \to \hat{Y}$ in the causal graph implied by the chosen construction of $\hat{Y}$. Then a path-specific counterfactually fair predictor is one that uses $\{\epsilon_U, F\}$ as input. Note that the only difference this makes in our grid-based tool is that we only estimate the error ϵ_U for the unfair path in Model A (step 2, Section 3.3) and fit a predictor on $\{\epsilon_U, F\}$ (step 3, Section 3.3). Additionally, we only compute the incorrect error terms of the counterfactuals in Model B, using the weights of Model A (step 4, Section 3.4). For the optimization-based tool we would change lines 13, 14, and 20 in the same way.