Counterfactual fairness: removing direct effects through regularization

Pietro G. Di Stefano*
pietro.distefano@experian.com
Experian UK&I and EMEA DataLabs
London, UK

James M. Hickey
james.hickey@experian.com
Experian UK&I and EMEA DataLabs
London, UK

Vlasios Vasileiou
vlasios.vasileiou@experian.com
Experian UK&I and EMEA DataLabs
London, UK

ABSTRACT

Building machine learning models that are *fair* with respect to an unprivileged group is a topical problem. Modern fairness-aware algorithms often ignore causal effects and enforce fairness through modifications applicable to only a subset of machine learning models. In this work, we propose a new definition of fairness that incorporates causality through the Controlled Direct Effect (CDE). We develop regularizations to tackle classical fairness measures and present a causal regularization that satisfies our new fairness definition by removing the impact of unprivileged group variables on the model outcomes as measured by the CDE. These regularizations are applicable to any model trained using by iteratively minimizing a loss through differentiation. We demonstrate our approaches using both gradient boosting and logistic regression on: a synthetic dataset, the UCI Adult (Census) Dataset, and a real-world credit-risk dataset. Our results were found to mitigate unfairness from the predictions with small reductions in model performance.

CCS CONCEPTS

• **Computing methodologies** → *Regularization*; *Boosting*; *Modeling methodologies*.

KEYWORDS

Machine Learning, Fairness, Causality

1 INTRODUCTION

One of the most famous concepts in computer science is the one of "garbage in, garbage out". Applied to machine learning algorithms, this phrase captures the concept that the ability to train and the quality of the output from a machine learning model is dependent on the quality of the training data presented to it. Advances in recent decades have resulted in machine learning algorithms that can leverage a larger variety of data types and sources for ever more complex learning tasks. This increased capacity to learn from data also raises risks of incorporating undesired biases from the training data into a machine learning model [36, 40]. Furthermore, even when the data is completely unbiased, machine learning systems can produce biased results for specific groups, as can the case where the sensitive groups form a small minority of the training data. This problem goes beyond "bias in, bias out" becoming in the worst case "fairness in, bias out". It has presented itself in a range of domains from credit-risk assessment to determining an individual's propensity for criminal recidivism [10, 26], and has attracted the attention of both regulators and the media [30].

After the "fairness through unawareness" paradigm was shown to be flawed [11, 14], new approaches to handle discrimination in machine learning have been explored. Recent advances enable practitioners to specify which groups, usually derived from one or more sensitive attributes, they are concerned about possibly treating unfairly and design a system that ameliorates potential bias (unfairness) in the outputs. These adaptations cover the full pipeline of the machine learning training problem, with the main approaches including pre-processing of the data [6, 15, 19, 41], post-processing of outputs [18, 20, 33] and incorporation of fairness constraints directly into the objective function [2, 5, 7, 17, 27, 42]. Unfortunately, all of these methods suffer from drawbacks [12, 23] such as requiring access to sensitive attributes even after model training (an undesirable scenario in many circumstances), ignoring causal structures in the data, and being specific to a particular machine learning algorithm.

These drawbacks pose serious issues for anyone wishing to ensure fairness in their machine learning practices. Firstly, the wide diversity of statistical fairness measures makes it difficult to select an appropriate metric and corresponding fairness-aware algorithm. This is exacerbated by the fact that while many metrics seemingly address the same underlying notion of fairness, those metrics cannot be mathematically optimized simultaneously for a given task [16]. Furthermore, the statistical nature of the metrics make it difficult to discern correlation from causation when examining decisions and addressing fairness bias. The role of causality [9, 24, 32] when reasoning about fairness should not be understated. It is considered of serious importance by social-choice theorists and ethicists and it has been argued that causal frameworks would decisively improve reasoning about fairness [25]. This has resulted in several causal definitions of fairness that take advantage of the concept of *counterfactuals* to address the potential unfair *causal effects* on model outcomes. However, the implementation of these causal worldviews typically focused on generative models and not on how causality may be incorporated into popular discriminative machine learning models. Moreover, there is a general lack of comparison between models constrained by statistical fairness metrics and causal effects.

In this paper, we address each of the presented issues in turn within a *Counterfactual Fairness* framework. In this framework, we first describe a novel definition of fairness that seeks to remove the controlled direct effect (CDE) of the sensitive attribute from the model's predictions. We then propose to enforce such definition through regularization. This is achieved through propensity-score matching [34, 35] and the development of a mean-field theory. Our fairness-via-regularization approach is applicable to any model trained by minimizing a loss function through differentiation. We will focus on the case of binary classification, with a single binary

*To whom correspondence should be addressed

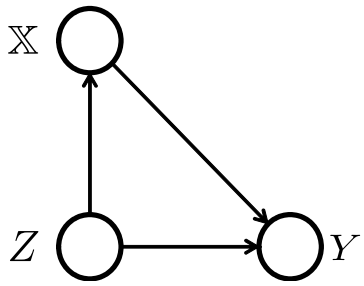

Figure 1: Assumed causal graph for our data generating process, where Z is a parent attribute potentially having a direct effect and an indirect effect through a vector-valued set of features $\mathbb{X}$.

indicator defining un-/privileged groups, and are exemplified using gradient boosted trees and logistic regression.

To further allow evaluation of our Counterfactual Fairness regularization framework, we compare its results against a regularization strategy aimed at satisfying full equality of outcomes between groups. This allows us to highlight how the differences between these worldviews manifest in the optimized model.

The effectiveness and drawbacks of our methods are illustrated using a public benchmark dataset, a private commercial credit-risk dataset, and a synthetic dataset. The inclusion of a private commercial credit-risk dataset provides insight into how these approaches could work in an industrial setting. We consider this form of evaluation to be of critical importance, as the main risks of machine learning fairness are beared by end-consumers and businesses through high-velocity decisioning systems built on such datasets.

The structure of the papers is as follows: in Section 2 we introduce our notation, in Section 3 we provide a background on some key concepts of causality and mediation effects, in Section 4 we provide a discussion on algorithmic fairness, while in Section 5 we present our regularization strategies. We discuss the results of the experiments in Section 6 and relationships with existing literature in Section 7. Finally, we state our conclusions in Section 8.

2 NOTATION AND SETTING

Before beginning our discussions on how fairness is measured and the connection to causal modelling, we introduce our notation. We use Y to refer to the observed label in the data while the sensitive attribute is denoted Z. Throughout this work, our groups of interest will be defined by a single, binary sensitive attribute and we only consider binary classification tasks, i.e. $Y, Z \in \{0, 1\}$. We denote the covariates that do not define the groups of interest (the insensitive covariates) by $\mathbb{X}$. We denote probabilities with $P(\bullet)$ and probability densities with $p(\bullet)$. We assume that our models provide probability estimates $\tilde{Y} = P(Y = 1|\mathbb{X})$, and $\hat{Y} \in \{0, 1\}$ are the binary outcomes obtained by thresholding $\tilde{Y}$.

3 BACKGROUND ON CAUSALITY

The aim of this section is to introduce counterfactual quantities and causal effects, especially mediated ones. We follow the causality literature in denoting counterfactual quantities using subscripts, i.e. we define the counterfactual outcome $Y_{(Z=Z^*,\ \mathbb{X}=\mathbb{X}^*)}$ as the outcome that would have been realized had an individual been assigned the values $Z = Z^*$ and $\mathbb{X} = \mathbb{X}^*$. We assume an unconfounded causal graph of the form of Fig. 1, which embeds our assumption of sequential ignorability:

Assumption 1. *(Sequential ignorability) The following conditional independence relations hold for each realization Z^* and $\mathbb{X}^*$:*

$$\begin{aligned} Y_{Z=Z^*} &\perp\!\!\!\perp Z \\ Y_{\mathbb{X}=\mathbb{X}^*} &\perp\!\!\!\perp \mathbb{X}|Z \end{aligned} \tag{1}$$

The former of the above conditions is equivalent to assuming the absence of confounders that would open a "backdoor" path between Z and Y [32]. We argue that for our purposes the absence of such paths is justifiable in a wide variety of cases, as the sensitive attribute is usually measured at birth, and is therefore naturally a parent variable.

To avoid cases in which the protected attribute is completely specified by the mediators, we also assume the following:

Assumption 2. *(Strong ignorability) There is overlap between the two groups: $0 < P(Z = 1|\mathbb{X}) < 1,\ \forall \mathbb{X}$.*

We start by defining the Average Treatment Effect (ATE):

$$\begin{aligned} \text{ATE} &= \mathbb{E}[Y_{Z=1} - Y_{Z=0}] \\ &= \mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0], \end{aligned} \tag{2}$$

where the second line is a consequence of Sequential ignorability and the law of counterfactuals [32] $P(Y_{Z=Z^*}|Z = Z^*) = P(Y|Z = Z^*)$.

The ATE represents the total causal effect of the protected attribute. In this work, as we shall see in the next section, we are interested in *mediation effects*, i.e. direct and indirect effects, instead of the ATE. We define the point-wise Controlled Direct Effect (CDE) as:

$$\begin{aligned} \text{CDE}(\mathbb{X}) =& \mathbb{E}[Y_{(Z=1,\mathbb{X})} - Y_{(Z=0,\mathbb{X})}] \\ =& \mathbb{E}[Y|Z = 1, \mathbb{X}] - \mathbb{E}[Y|Z = 0, \mathbb{X}], \end{aligned} \tag{3}$$

where, as before, the second line follows from Sequential ignorability. The CDE represents the effect of changing the value of the protected attribute while keeping the value of the covariates fixed. Eq. 3 provides a means for directly estimating the CDE from the data, e.g., via regression techniques [3, 37].

Another important mediation effect is the Natural Direct Effect (NDE), defined as follows. Given a "baseline" value Z^* for the protected attribute, e.g. "female" for a gender discrimination problem, we define the NDE as the difference in Y that would be attained by changing Z, with the value of the mediating variables $\mathbb{X}$ set to what they would have attained had Z been Z^*, i.e. $\mathbb{X}_{Z^*}$. This yields:

$$\begin{aligned} \text{NDE}(Z^*) &= \mathbb{E}[Y_{(Z=1-Z^*,\mathbb{X}=\mathbb{X}_{Z^*})} - Y_{Z=Z^*}] \\ &= (-1)^{(Z^*-1)} \int \mathrm{d}\mathbb{X}\ \text{CDE}(\mathbb{X}) p(\mathbb{X}|Z = Z^*), \end{aligned} \tag{4}$$

where the second line is proved in Ref. [31]. The natural direct effect has a few nice properties. First, it has straightforward implications

in discrimination problems. Second, contrary to the CDE case, it has an indirect counterpart, namely the Natural Indirect Effect (NIE), defined as:

$$\text{NIE}(Z^*) = \mathbb{E}[Y_{Z=Z^*} - Y_{(Z=Z^*, \mathbb{X}=\mathbb{X}_{1-Z^*})}]. \tag{5}$$

The NIE is easily interpreted as the effect on the baseline population's response of changing $\mathbb{X}$ to the value it would have naturally obtained in the non-baseline population while keeping the value of the protected attribute fixed.

Natural direct and indirect effects are, in contrast with the CDE, population-wide quantities and depend on the choice of a baseline value for the protected attribute.

We conclude this section by citing two equalities derived by Pearl [31], which will be useful later and describe the relation between ATE and Natural mediated effects:

$$\begin{cases} \text{ATE} &= \text{NIE}(1) - \text{NDE}(0) \\ \text{ATE} &= \text{NDE}(1) - \text{NIE}(0) \end{cases} \tag{6}$$

4 ALGORITHMIC FAIRNESS

4.1 Statistical Fairness

The traditional definitions of algorithmic fairness are statistical in nature and reflect underlying worldviews on how the "true" unmeasured target is recorded, its relationship to Z and how it is predicted by a machine learning model. In this framework, statistical measures are derived from one's belief of the latent space structure [16] rather than by specifying causal pathways between the unobserved "true" label and the collected data. Under this construct, two prevailing worldviews of statistical fairness emerge: "We're all equal" and "What you see is what you get". The former focusses on outcomes defined by group membership and seeks to ensure groups are treated equally through balanced outcomes. Contrastingly, the latter favours is a worldview where the data is accurate and so it seeks to offer similar individuals similar outcomes as informed by the data. We focus here on one of the most popular fairness metrics, namely statistical parity difference.

SPD is a group fairness measure on an algorithm's outcome, $\hat{Y}$, and it is 0 (maximally fair) only when $P[\hat{Y} = 1|Z = 1] = P[\hat{Y} = 1|Z = 0]$. It is defined as follows:

$$\text{SPD} = |P[\hat{Y} = 1|Z = 1] - P[\hat{Y} = 1|Z = 0]|. \tag{7}$$

We note that this measure can also be applied to the data by changing $\hat{Y}$ to Y in Equation 7. As there is no link between the algorithm outcome and measurement, this difference can be minimized by changing the outcomes of members of each group independent of all other attributes and so can be viewed as a "lazy penalization".

4.2 Causal Modelling and Fairness

The measures of fairness presented in the Section 4.1 emerge from *a priori* worldviews on how the underlying "true" label is related to Z and the veracity of the recorded data. They do not formally encode the causal relationships between the Z, $\mathbb{X}$ and Y. Consequently, how much causal effect on $\hat{Y}$ can be attributed to Z, either directly or indirectly, is handled implicitly in the worldview employed.

The need to explicitly distinguish between direct and indirect effect is well illustrated by the 1973 University of California, Berkeley gender discrimination scandal [4]. In that case, the data showed significant bias in admissions for male and female applicants. However, after controlling for the department chosen by the applicants, that bias disappeared. It was actually found that female applicants had lower overall admission rates not because they were discriminated against, but simply because they were applying to more competitive departments.

In the Counterfactual Fairness worldview we examine here, we are only concerned with biases that are consequences of *direct effects*. Specifically, we seek to identify, and correct for, how much an outcome for an individual assigned a specific value of Z would change compared to a *counterfactual* world where they had been assigned the alternative value for Z but all other factors had remained the same [1].

This worldview requires in-depth understanding of the data collection and generation process. In particular, it is important to define what information can be recorded in $\mathbb{X}$. We require that variables $\mathbb{X}$ are "fair" in the sense specified by the following requirements:

(1) They were not measured before Z.
(2) They do not, either individually or in combination, directly measure discriminatory attributes.
(3) They are relevant to the problem at hand.

For example, in a financial application where the sensitive attribute is race, "income" would be typically considered fair while "race of the applicant's mother" or "blood pressure" would not be. Taking all of this together, we propose the following definition:

Definition 1. (Counterfactual Fairness) Given a set of fair covariates $\mathbb{X}$, a sensitive attribute Z and a target Y, a fair model is a model that does not learn the controlled direct effect of Z on Y.

We stress that any effect mediated by unobserved variables or variables not included in $\mathbb{X}$ during training will be embedded in the CDE and hence the $\mathbb{X}$ will determine the size of the bias we seek to remove.

We finally note that our definition can be seen as an instance of path-specific counterfactual fairness [9, 29, 39, 43] with the additional requirements on the covariates $\mathbb{X}$ highlighted above.

5 FAIR LOSSES

To improve the fairness aspects of a model's output, in both the Statistical Fairness (Sec. 4.1) and Counterfactual Fairness (Sec. 4.2) worldviews, we modify the loss function and apply regularization. The modifications to the loss function can be applied to all algorithms trained by iteratively minimizing a loss through differentiation, e.g. through gradient descent or gradient boosting.

We combine an original loss function ($\mathcal{L}_o$), which captures our utility objectives, with a fairness penalty ($\mathcal{R}_f$) using a regularization weight (λ). The generic form of fairness-aware loss function ($\mathcal{L}_f$) is as follows:

$$\mathcal{L}_f = (1 - \lambda)\mathcal{L}_o + \lambda\mathcal{R}_f, \tag{8}$$

where the modifications $\mathcal{R}_f$ are differentiable. As we're focussing on binary classification, we take $\mathcal{L}_o$ to be the binary cross-entropy of Y and $\tilde{Y}$.

It is important to note that Z is only required at training time to evaluate $\mathcal{R}_f$ and is not required for prediction. This is a very useful feature for real-world applications, as obtaining Z can be difficult.

5.1 Statistical Fairness Regularization

For illustrative purposes and to compare our causal worldview with the statistical "we are all equal" one, we propose a very simple loss aimed at reducing the SPD of the average scores:

$$\mathcal{R}_f^{\text{SPD}} = \{\mathbb{E}[\tilde{Y}|Z=1] - \mathbb{E}[\tilde{Y}|Z=0]\}^2 \tag{9}$$

5.2 Counterfactual Fairness Regularization

To satisfy Definition 1, the CDE must not be learned by the model. The point-wise CDE is given by Eq. (3), which can be estimated using regression techniques [3, 37]. However, our algorithm, as we shall see, would require us to fit such regressions iteratively, which can be computationally inefficient if we condition on the full set of covariates. To circumvent this possible computational bottleneck, we employ a "balancing score". Balancing scores are functions $b(\mathbb{X})$ of the covariates that make $\mathbb{X}$ and Z conditionally independent, i.e.:

$$\mathbb{X} \perp\!\!\!\perp Z | b(\mathbb{X}). \tag{10}$$

In our work we use one such score, namely the "propensity score" $b(\mathbb{X}) = P(Z=1|\mathbb{X})$. Traditionally this score is utilized in scenarios where $\mathbb{X}$ play the role of confounders rather than mediators. Although the latter case is true in our instance, utilizing the propensity score solves our key computation problem whenever Assumptions 1 and 2 are true.

Given assumptions 1 and 2, we can define a "mean field" CDE given by:

$$\begin{aligned}\text{MFCDE}(b) &= \mathbb{E}[Y|Z=1,b] - \mathbb{E}[Y|Z=0,b] \\ &= \frac{\int_{\mathcal{X}_b} \mathrm{d}\mathbb{X}\, p(\mathbb{X})\text{CDE}(\mathbb{X})}{\int_{\mathcal{X}_b} \mathrm{d}\mathbb{X}\, p(\mathbb{X})},\end{aligned} \tag{11}$$

where $\mathcal{X}_{b^*} := \{\mathbb{X} \text{ s.t. } b(\mathbb{X}) = b^*\}$. The second line of Eq. (11) is proven in Appendix A. The MFCDE can thus be interpreted as an average of the CDE across a volume with constant $b = b^*$. We also note that, as in a classic result by Rosenbaum and Rubin [34] which we extend to the CDE in Appendix A, the population-wide averages of MFCDE and CDE are the same, i.e.

$$\int \mathrm{d}b\; p(b)\,\text{MFCDE}(b) = \int \mathrm{d}\mathbb{X}\; p(\mathbb{X}).\text{CDE}(\mathbb{X}) \tag{12}$$

A stronger statement can be made about the MFCDE if we assume the following:

Assumption 3. *(Mean field approximation) Wherever* $p(\mathbb{X}) > 0$, $\text{CDE}(\mathbb{X})$ *is approximately constant in* $\mathcal{X}_{b(\mathbb{X})}$.

This allows us to approximate the MFCDE with the point-wise CDE, i.e. $\text{CDE}(\mathbb{X}) \simeq \text{MFCDE}(b(\mathbb{X}))$. A key result is then that minimization of the CDE (per Counterfactual Fairness) is approximately equivalent to minimizing the MFCDE. Thus, our goal is now to iteratively train a model conditioned on the fair covariates $\mathbb{X}$ only and which does not learn the MFCDE.

As a first step, we use our training data to estimate $\mathbb{E}[Y|Z,b]$ with a regression model such as:

$$\mathbb{E}[Y|Z,b] = \sum_{k=0}^{N_1} \alpha_k b^k + Z \sum_{k=0}^{N_2} \beta_k b^k, \tag{13}$$

for arbitrary N_1 and N_2. We then use this to define a "fair" target Y_f such that:

$$\begin{aligned}\mathbb{E}[Y_f|Z,b] &= \mathbb{E}[Y|Z,b] + \frac{(-1)^Z}{2}\text{MFCDE}(b) \\ &= \sum_{k=0}^{\max(N_1,N_2)} \gamma_k b^k,\end{aligned} \tag{14}$$

where we used the fact that $\text{MFCDE}(b) = \sum_{k=0}^{N_2} \beta_k b^k$ and, employing the indicator function $I(\bullet)$, we defined

$$\gamma_k = \alpha_k\, I(k \le N_1) + \frac{1}{2}\beta_k\, I(k \le N_2). \tag{15}$$

Y_f is easily interpreted as a version of a target corrected by a symmetrized version of the CDE, where privileged and unprivileged groups receive anti-symmetrical corrections.

We can also show (see Appendix A) that:

$$\mathbb{E}[Y_f|Z=1] - \mathbb{E}[Y_f|Z=0] = \frac{\text{NIE}(1) - \text{NIE}(0)}{2}, \tag{16}$$

meaning that the ATE of Z on Y_f is equal to symmetric version of the indirect effect on Y, thereby justifying the definition of fair target.

Having defined our "fair" target, and how it relates to the MFCDE (hence, to the CDE), we now describe the procedure to remove the CDE during model training. Firstly, at each iteration of the optimization algorithm, we extract the probability scores estimated by our model at that iteration, the balancing scores and the protected attributes for each training example to define a surrogate training set $\{(\tilde{Y}_i, b_i, Z_i)\}$. Then, we use this latter to estimate the coefficients of surrogate model:

$$\mathbb{E}[\tilde{Y}|Z,b] = \sum_{k=0}^{\max(N_1,N_2)} \tilde{\alpha}_k b^k + Z \sum_{k=0}^{N_2} \tilde{\beta}_k b^k. \tag{17}$$

To remove the direct effect, we want to limit the coefficients of this surrogate model to those of the fair target Y_f. Our aim will then be to achieve $\tilde{\beta}_k \to 0$ and, for $k > 0$, $|\tilde{\alpha}_k| < |\gamma_k|$, which leads us to the following loss:

$$\mathcal{R}_f^{\text{CDE}} = \sum_{k=1}^{\max(N_1,N_2)} I(|\tilde{\alpha}_k| > |\gamma_k|)(\tilde{\alpha}_k - \gamma_k)^2 + \sum_{k=0}^{N_2} \tilde{\beta}_k^2 \tag{18}$$

In order for our loss to be differentiable, we require that the coefficients $\tilde{\alpha}_k$ and $\tilde{\beta}_k$ are estimated as functions which can be differentiated w.r.t. all the $\tilde{Y}_i$. In our experiments, we used Ordinary Least Square (OLS) regression. Although a wide array of models can be used to define $\mathcal{R}_f^{\text{CDE}}$, we recommend that the selected model is collapsible [28, 38] in order to give the correct causal interpretation to the coefficient.

We emphasize how computing the OLS regression coefficients has a complexity that scales quadratically with the number of covariates. This means that, especially in cases where the number of covariates $\mathbb{X}$ is large, as is often the case in today's applications, our mean field solution allows for a dramatic speed-up.

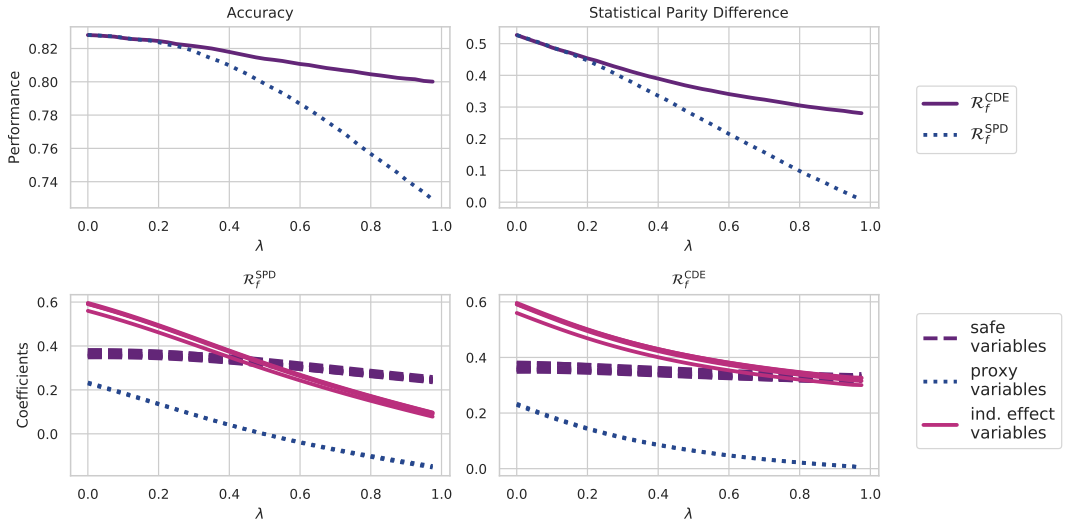

Figure 2: Results for the synthetic dataset and logistic regression models using the SPD Loss of Eq. (9) and the CDE Loss of Eq. (18). For the CDE loss, we used $N_1 = 1, N_2 = 0$. All the results are plotted against the regularization strength λ. Top panel: Accuracy (left) and SPD (right) using a threshold equal to 0.5. Bottom panel: Coefficients of the logistic regression models for the safe, indirect effect and proxy variables (see Section 6.1.1)

6 EXPERIMENTS

We evaluated our algorithms on three binary classification tasks using a synthetic dataset, which we included for illustrative purposes, the UCI adult dataset [13], and a commercial credit-risk dataset . We briefly describe these datasets in Section 6.1 and provide a summary in Table 1.

For each dataset, algorithm and loss, we computed results by sweeping the regularization parameter λ from 0 to 0.975 using a step size of 0.025. We evaluated every model trained by its fairness, as measured by SPD, and accuracy (or precision, for the credit-risk models). For the synthetic dataset, we evaluated our results using only logistic regression, while on the other two datasets we used both logistic regression and XGBoost [8].

For the synthetic dataset, we show results for both the SPD loss [Eq. (9)], and CDE loss [Eq. (18)], comparing how the results highlight the differences in worldviews that these losses entail for a problem for which we know the causal story. Contrastingly, for the adult and credit-risk we only display results for the CDE loss.

We employed logistic regression with L_1 loss to estimate the propensity scores b. We also mention that, for higher values of λ, we found it beneficial to first pre-train our models using small values of λ until convergence, and only after that slowly increase λ up to the desired value.

We note that our losses are twice differentiable and we supply the diagonal of the Hessian during training to the XGBoost algorithm in every case.

6.1 Datasets

6.1.1 Synthetic Data. We generated a synthetic dataset mimicking the graph of Fig. (1). We sampled the protected attribute out of a Bernoulli distribution and three sets of covariates: "safe" covariates $\mathbb{X}_s \sim \mathcal{N}(0, 1)$, "proxy" covariates and "indirect effect" covariates sampled from $\mathbb{X}_p, \mathbb{X}_i \sim \mathcal{N}(Z, 1)$. We then define the log-odds of the binary target as $S_Y = 0.25\mathbf{w} \cdot (\mathbb{X}_i + \mathbb{X}_s) + 1.25Z$. Here, $\mathbf{w}$ is a vector of ones. We finally sample $Y \sim \text{Bern}\left(\frac{1}{1+e^{-S_Y}}\right)$. For our experiments, we used 10 safe, 4 indirect effect and 2 proxy variables.

Dataset	n. rows	n. features	SPD
Synthetic	100,000	16	0.54
UCI Adult	48,842	9	0.20
Credit-risk	71,809	58	0.15

Table 1: Summary of datasets included

6.1.2 UCI Adult Dataset. For this dataset, the goal is to predict whether a person will have an income below or above 50K USD. In this dataset, we are interested in removing gender bias and so we consider gender as our protected attribute. Furthermore, we excluded the following additional sensitive attributes: race, marital status, native country and relationship from our models. The other covariates primarily relate to financial information, occupation and education.

6.1.3 Private Credit-Risk Dataset. In this dataset, we are trying to infer the probability that a customer will not default on their credit given curated information on their current account transactions. Here, we're interested in removing bias related to age. We binarized the age variable dividing our examples in two groups, an "older" group of people over 50 and an "younger" group of people under 50.

Private to Experian

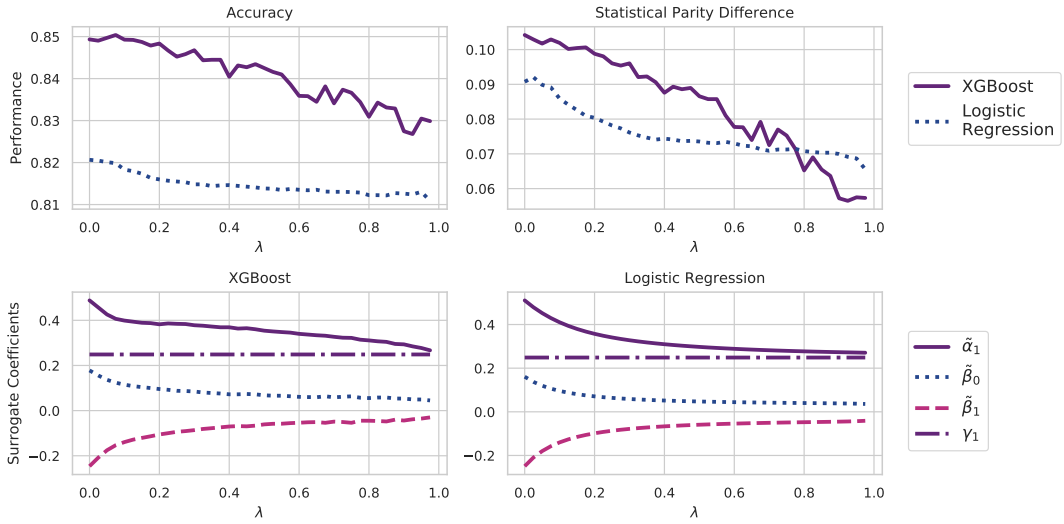

Figure 3: Results for the UCI Adult dataset employing XGBoost and Logistic regression models modified through the loss of Eq. (18). Results are plotted against the regularization strength λ. For these experiments, we employed $N_1 = N_2 = 1$. Top panel: Accuracy (left) and SPD (right) using a threshold equal to 0.5. Bottom panel: surrogate coefficients [see Eq. (17)] estimated on the test set for the XGBoost (left) and Logistic Regression (right) models.

6.2 Results

Results for the synthetic dataset and logistic regression models are shown in Fig. 2. We observe that, as we sweep $\lambda \to 1$, the accuracy of the CDE loss is generally higher than that of the SPD loss, while the SPD is higher. This can be explained by the fact that SPD loss enforces a far more stringent view on fairness than the CDE loss, as it removes both average direct and indirect effects. The SPD loss converges to an almost ideal performance in its target fairness metric. In the bottom panel, which is our main result for this dataset, we show coefficients of the logistic regression model plotted against the regularization strength λ. We find that the coefficients of the variables that were constructed to be correlated both with the target Y and the protected attribute Z (indirect effect variables) and of those correlated with Z alone (proxy variables) become smaller as the fairness regularization increases. Furthermore, we observe that the coefficient changes with λ reflect the different worldviews described in Section 4. The CDE loss is very faithful to the original causal story (see Section 6.1.1), where the causal sampling coefficients of the fair and indirect effect variables are the same and the coefficients of the proxy variables are zero. Contrastingly, the SPD loss cause the coefficients to converge to values very different from the ones of the data generating distribution, which in this worldview is in itself deemed "unfair".

Results for the UCI Adult dataset are shown in Fig. 3. For this dataset, referring to Eq. (18), we used $N_1 = N_2 = 1$. For the XGBoost models, we observe that, as λ is increased, accuracy drops by a tolerable amount, i.e. from roughly 0.85 at $\lambda = 0$ to roughly 0.83 at $\lambda = 0.975$. At the same time, SPD is reduced from 0.10 to 0.06. Results for logistic regression are more nuanced: accuracy goes from 0.82 to 0.81, while SPD decreases from 0.09 to 0.07. In the bottom panel we show how the coefficients of the surrogate model [see Eq. (17)] change as λ increases. We notice how $\tilde{\alpha}_1$ converges to its target γ_1 while the $\tilde{\beta}_i$ coefficients approach zero.

Finally, in Fig. 4 we show results for the credit-risk dataset. Here we used $N_1 = 1, N_2 = 0$. Also, we tailored our results to the specificity of the credit-risk problem, where usually people are approved when their default probability is very low. We therefore used a threshold of 0.85 and, since we are more interested in the cost of false positives than the one of false negatives, we evaluate the performances of the model using precision. Results for XGBoost show that precision went from 0.96 to 0.89, while SPD went from 0.28 to 0.09. Logistic regression did not seem particularly affected by the regularization. In this case, precision went from 0.86 to 0.85, while SPD dropped from 0.05 to 0.04. To explain this observation, we conjecture that logistic regression does not have enough statistical capacity to absorb the direct effect when it's not directly exposed to the protected attribute Z. Results for the surrogate coefficients (bottom panel) show how the XGBoost models display generally good convergence.

7 RELATED WORK

There has been significant advancement in the areas of incorporating fairness into machine learning algorithms and the role of causality in fairness.

Training fair models: In Ref. [42], a fair neural network was built through the use of an adversarial model that tries to predict the sensitive attribute from the model outputs. Similarly, Ref. [27] developed statistical fairness regularizations to debias neural networks. The form of these regularizations restricted them to neural networks. Ref. [17] incorporate convex regularizations directly into training a fair logistic regression, but this approach relies on empirical weights to represent historical bias and is directly related to

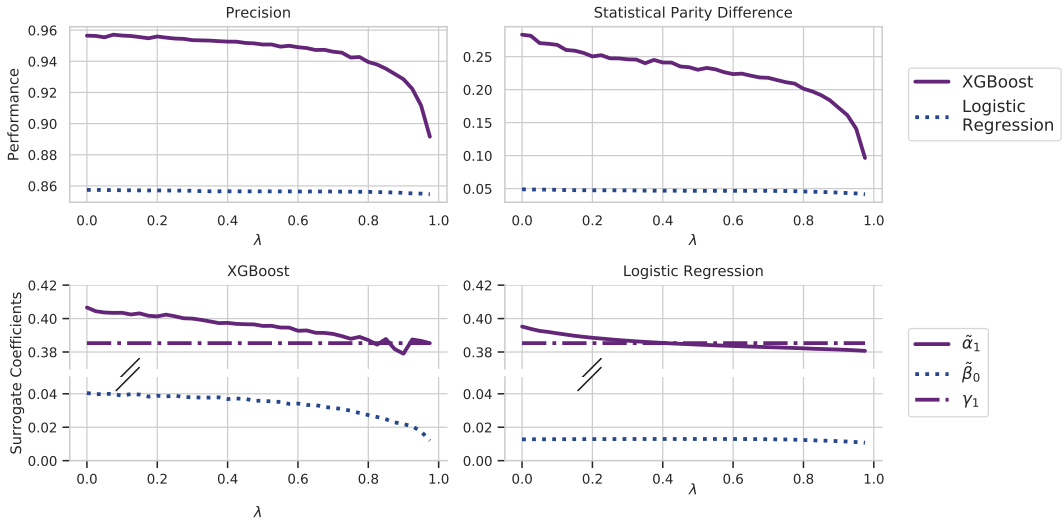

Figure 4: Results for the credit-risk dataset employing XGBoost and Logistic regression models modified through the loss of Eq. (18). Results are plotted against the regularization strength λ. For these experiments, we employed $N_1 = 1, N_2 = 0$. Top panel: Accuracy (left) and SPD (right) using a threshold equal to 0.85. Bottom panel: surrogate coefficients [see Eq. (17)] estimated on the test set for the XGBoost (left) and Logistic Regression (right) models.

proportionally fair classification rather than a traditional fairness measure. Other approaches have instead posed the problem as one of constrained optimization while others still have used multiple models to remove bias [5, 21]. Here, we propose novel regularizations that are applicable to any model trained using gradient-based optimizers and are designed to target both the classic and causal metrics directly in the scores.

Causal Fairness: Several works have recently addressed the problem of tackling some definition of counterfactual fairness [22, 24]. In Refs. [9, 29, 39, 43] the problem of identifying and removing path specific effects is studied. Those papers consider generative (or partly generative [29]) models. Since direct effects are particular cases of path specific effects, the scope of those works is somewhat bigger than ours but, crucially, they do not provide a generic method for incorporation of such effects into standard discriminative machine learning models. Our proposal also benefits from being model-agnostic. Furthermore, to compare our worldview with these path-specific works, we borrow an example from [29]. In this example, the influence of gender on hiring outcomes for a white collar job might be considered fair through a variable such as education, and unfair through a variable such as physical strength. Here, we argue that for most practical purposes our definition of fair covariates $\mathbb{X}$ (see Section 4.2) should suggest that physical strength should not be selected among the $\mathbb{X}$ as it easily arguable that it itself is discriminatory attribute and is irrelevant to the problem: one might not want to discriminate between stronger and weaker people for this particular application, regardless on the effect that gender has on it. This combination of what variables are deemed fair individually irrespective of the pathway in conjunction with the CDE makes our worldview particularly novel and applicable in many industrial settings.

8 CONCLUSIONS

In this work, we extended the literature by proposing a new definition of fairness that focuses on the removal of the controlled direct effect and is causal in nature. Incorporating causal effects into notions of fairness is crucial [4], and we argue that our definition is intuitive and general. We demonstrated how to enforce it through the use of a regularization term. Our solution is particularly appealing with respect to existing ones because it is applicable to any model that uses gradient-based optimization, including popular discriminative models. We exemplified our approaches on three datasets using XGBoost and logistic regression. In all cases, our framework allowed for a realistic trade-off between fairness and predictive performance.

9 ACKNOWLEDGEMENTS

We are indebted to C. Dhanjal, F. Bellosi, G. Jones and L. Stoddart for fruitful discussions. We also wish to acknowledge Experian Ltd and J. Campos Zabala for supporting this work.

A BALANCING SCORES AND MEDIATION EFFECTS

In this Appendix, we wish to justify Eq. (11), (12) and (16). Eq. (11) is proven below.

Theorem 1. *If Assumptions 1 and 2 hold, then, if $b(\mathbb{X})$ is a balancing score, the second line of Eq. (11) holds.*

Proof. In order to prove our hypothesis, it is sufficient to show that

$$\mathbb{E}[Y|b, Z] = \frac{\int_{\mathcal{X}_b} \mathrm{d}\mathbb{X}\, p(\mathbb{X})\mathbb{E}[Y|\mathbb{X}, Z]}{\int_{\mathcal{X}_b} \mathrm{d}\mathbb{X}\, p(\mathbb{X})}$$

We have:

$$
\begin{aligned}
&\mathbb{E}[Y|b,Z] \qquad (19)\\
&= \int \mathrm{d}\mathbb{X}\ \mathbb{E}[Y|b,Z,\mathbb{X}]p(\mathbb{X}|b,Z)\\
&= \int \mathrm{d}\mathbb{X}\ \mathbb{E}[Y|b,Z,\mathbb{X}]\frac{P(Z|\mathbb{X},b)p(\mathbb{X}|b)}{P(Z|b)}\\
&= \frac{\int_{\mathcal{X}_b} \mathrm{d}\mathbb{X}\ \mathbb{E}[Y|b,Z,\mathbb{X}]\frac{P(Z|\mathbb{X},b)p(\mathbb{X})}{P(Z|b)}}{\int_{\mathcal{X}_b}\mathrm{d}\mathbb{X}\ p(\mathbb{X})}\\
&= \frac{\int_{\mathcal{X}_b} \mathrm{d}\mathbb{X}\ p(\mathbb{X})\mathbb{E}[Y|\mathbb{X},Z]}{\int_{\mathcal{X}_b}\mathrm{d}\mathbb{X}\ p(\mathbb{X})}
\end{aligned}
$$

Where line 4 follows from $p(\mathbb{X}|b) = I(\mathbb{X}\in\mathcal{X}_b)\,p(\mathbb{X})/\int_{\mathcal{X}_b}\mathrm{d}\mathbb{X}\,p(\mathbb{X})$. The final line follows from the definition of the balancing score. □

In order to justify Eqs. (12) and (16) we also need to prove the following:

Theorem 2. *Under the assumptions of Theorem 1, the following equations hold:*

$$\mathbb{E}_{\mathbb{X}}[\mathrm{CDE}(\mathbb{X})] = \int \mathrm{d}b\ \mathrm{MFCDE}(b)p(b) \qquad (20)$$

$$\mathrm{NDE}(Z^*) = \int \mathrm{d}b\ \mathrm{MFCDE}(b)p(b|Z=Z^*), \qquad (21)$$
$$\forall Z^*\in\{0,1\}$$

Proof. The proof of Eqs. (20) and (21) are very similar. We'll prove Eq. (21) and leave the proof of Eq. (20) to the reader. It is sufficient to show that, for each value of $Z', Z^*\in\{0,1\}$ we have

$$\int \mathrm{d}\mathbb{X}\ p(\mathbb{X})\mathbb{E}[Y|\mathbb{X},Z=Z'] =$$
$$\int \mathrm{d}b\ p(b|Z=Z^*)\mathbb{E}[Y|b,Z=Z'].$$

The rest of the thesis follows straightforwardly from the definitions of CDE [Eq. (3)], MFCDE [Eq. (11)] and NDE [Eq. (4)]. We have:

$$
\begin{aligned}
&\int \mathrm{d}b\ \mathbb{E}[Y|b,Z=Z']p(b|Z=Z^*)\\
&= \int \mathrm{d}b\ \mathrm{d}\mathbb{X}\ \mathbb{E}[Y|b,Z=Z',\mathbb{X}]p(\mathbb{X}|b,Z=Z')p(b|Z=Z^*)\\
&= \int \mathrm{d}b\ \mathrm{d}\mathbb{X}\ \Big\{\mathbb{E}[Y|b,Z=Z',\mathbb{X}]\\
&\qquad\qquad \frac{P(Z=Z'|\mathbb{X},b)p(b|\mathbb{X})p(\mathbb{X})}{P(Z=Z'|b)p(b)}p(b|Z=Z^*)\Big\}\\
&= \int \mathrm{d}b\ \mathrm{d}\mathbb{X}\ \Big\{\mathbb{E}[Y|b,Z=Z',\mathbb{X}]\\
&\qquad\qquad \frac{P(Z=Z'|\mathbb{X},b)\delta(b-b(\mathbb{X}))p(\mathbb{X})}{P(Z=Z'|b)p(b)}p(b|Z=Z^*)\Big\}\\
&= \int \mathrm{d}\mathbb{X}\ \Big\{\mathbb{E}[Y|Z=Z',\mathbb{X}]\\
&\qquad\qquad \frac{P(Z=Z'|\mathbb{X},b(\mathbb{X}))p(\mathbb{X})}{P(Z=Z'|b(\mathbb{X}))p(b(\mathbb{X}))}p(b(\mathbb{X})|Z=Z^*))\Big\}\\
&= \int \mathrm{d}\mathbb{X}\ \mathbb{E}[Y|Z=Z',\mathbb{X}]p(\mathbb{X})\frac{P(Z=Z^*|\mathbb{X})}{P(Z=Z^*)}\\
&= \int \mathrm{d}\mathbb{X}\ \mathbb{E}[Y|Z=Z',\mathbb{X}]p(\mathbb{X}|Z=Z^*)
\end{aligned}
$$

where line 3 follows from two applications of Bayes' rule, $\delta(\bullet)$ in line 4 is Dirac's delta distribution, line 5 follows from integrating b out, line 6 again from Bayes' rule and the definition of a balancing score, which entails $P(Z|b(\mathbb{X})) = P(Z|(\mathbb{X})$; the final line follows straightforwardly, from a reverse application of Bayes' rule. □

Eq. (16) is then derived from Eqs. (21) and (6) with minimal algebra.

REFERENCES

[1] 1996. *Carson v. Bethlehem Steel Corporation.* United States Court of Appeals,Seventh Circuit.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, 60–69. http://proceedings.mlr.press/v80/agarwal18a.html

[3] Reuben M. Baron and David A. Kenny. 1986. The moderatorâEUR"mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51, 6 (1986), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

[4] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. 1975. Sex Bias in Graduate Admissions: Data from Berkeley. *Science* 187, 4175 (1975), 398–404. https://doi.org/10.1126/science.187.4175.398 arXiv:https://science.sciencemag.org/content/187/4175/398.full.pdf

[5] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21 (2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3992–4001. http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

[7] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA). Association for Computing Machinery, New York, NY, USA, 319–328. https://doi.org/10.1145/3287560.3287586

[8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD 2016)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[9] Silvia Chiappa and Thomas Gillam. 2018. Path-Specific Counterfactual Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (02 2018). https://doi.org/10.1609/aaai.v33i01.33017801

[10] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047 arXiv:https://doi.org/10.1089/big.2016.0047 PMID: 28632438.

[11] Kevin A. Clarke. 2005. The Phantom Menace: Omitted Variable Bias in Econometric Research. *Conflict Management and Peace Science* 22, 4 (2005), 341–352. https://doi.org/10.1080/07388940500339183 arXiv:https://doi.org/10.1080/07388940500339183

[12] Sam Corbett-Davies, Sharad Goel, Jamie Morgenstern, and Rachel Cummings. 2018. Defining and Designing Fair Algorithms. In *Proceedings of the 2018 ACM Conference on Economics and Computation* (Ithaca, NY, USA). Association for Computing Machinery, New York, NY, USA, 705. https://doi.org/10.1145/3219166.3277556

[13] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS 2012)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[15] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD ï£¡15)*. Association for Computing Machinery, New York, NY, USA, 259ï£¡268. https://doi.org/10.1145/2783258.2783311

[16] Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. (09 2016).

[17] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-Discriminatory Machine Learning Through Convex Fairness Criteria. (2018). https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16476

[18] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS-16)*. Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[19] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33 (2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[20] F. Kamiran, A. Karim, and X. Zhang. 2012. Decision Theory for Discrimination-Aware Classification. In *2012 IEEE 12th International Conference on Data Mining*. 924–929. https://doi.org/10.1109/ICDM.2012.45

[21] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*, Peter A. Flach, Tijl De Bie, and Nello Cristianini (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 35–50.

[22] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA). Curran Associates Inc., Red Hook, NY, USA, 656–666.

[23] Jon Kleinberg. 2018. Inherent Trade-Offs in Algorithmic Fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems* (Irvine, CA, USA) *(SIGMETRICS-18)*. Association for Computing Machinery, New York, NY, USA, 40. https://doi.org/10.1145/3219617.3219634

[24] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4066–4076. http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf

[25] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. 2018. Causal Reasoning for Algorithmic Fairness. arXiv:cs.AI/1805.05859

[26] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2016.00960.x

[27] P Manisha and S Gujar. 2018. A Neural Network Framework for Fair classifier. *arXiv preprint arXiv:1811.00247* (2018).

[28] Carina Mood. 2009. Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review* 26, 1 (03 2009), 67–82. https://doi.org/10.1093/esr/jcp006 arXiv:https://academic.oup.com/esr/article-pdf/26/1/67/1439459/jcp006.pdf

[29] Razieh Nabi and Ilya Shpitser. 2018. Fair Inference on Outcomes. (2018). https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16683

[30] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA.

[31] Judea Pearl. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (Seattle, Washington). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 411–420.

[32] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press, USA.

[33] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS-17)*. Curran Associates Inc., Red Hook, NY, USA, 5684–5693.

[34] Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (04 1983), 41–55. https://doi.org/10.1093/biomet/70.1.41 arXiv:http://oup.prod.sis.lan/biomet/article-pdf/70/1/41/662954/70-1-41.pdf

[35] Paul R. Rosenbaum and Donald B. Rubin. 1985. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician* 39, 1 (1985), 33–38. http://www.jstor.org/stable/2683903

[36] Harini Suresh and John V. Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002* (2019).

[37] Tyler VanderWeele and Stijn Vansteelandt. 2009. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface* 2 (2009), 457–468.

[38] Stijn Vansteelandt, Maarten Bekaert, and Gerda Claeskens. 2010. On Model Selection and Model Misspecification in Causal Inference. *Statistical methods in medical research* 21 (11 2010), 7–30. https://doi.org/10.1177/0962280210387717

[39] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2019. Achieving Causal Fairness through Generative Adversarial Networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 1452–1458. https://doi.org/10.24963/ijcai.2019/201

[40] Tal Zarsky. 2012. Automated Prediction: Perception, Law, and Policy. *Commun. ACM* 55 (2012), 33–35. https://ssrn.com/abstract=2149518

[41] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28* (Atlanta, GA, USA) *(ICML 13)*. JMLR.org, 325–333.

[42] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES 18)*. Association for Computing Machinery, New York, NY, USA, 335–340. https://doi.org/10.1145/3278721.3278779

[43] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 3929–3935. https://doi.org/10.24963/ijcai.2017/549

REPRODUCIBILITY NOTES

Supplementary materials are available in the online version of this paper.

Scope

In this document, we wish to include a few details that should help the reader reproduce our results. In particular, we will be focused on the results for the synthetic dataset and the UCI adult dataset (Figures 3 and 4, respectively, of the main text). The credit-risk results will not be reproducible due to the private nature of the dataset.

Technology

We used python v. 3.6, with the main packages employed being scikit-learn v. 0.21, pandas v. 0.24, numpy v. 1.16 and xgboost v. 0.82.

Computation of gradients and diagonal hessians

In this section we will give insights on how to compute gradients and diagonal hessians w.r.t. the probability scores of the model evaluated in each training example. Specifically, given the model scores on the training set $\tilde{Y}_i$, we want therefore to compute $\partial \mathcal{R}_f^{\text{CDE}} / \partial \tilde{Y}_i$ and $\partial^2 \mathcal{R}_f^{\text{CDE}} / \partial \tilde{Y}_i^2$ for all the examples i in the training set. Using these derivatives, XGBoost models can be trained directly[8], and any parametric model can be trained evaluating $\nabla_\theta \mathcal{R}_f^{\text{CDE}}$ through the chain rule as usual. Since these computations are quite trivial for the loss of Eq. (9), we will only focus on the loss of Eq. (18).

The approach we used was to compute the regression coefficients of both Eqs. (13) and (14) using OLS regression. Given the coefficients of Eq. (13), the loss of Eq. (9) can be seen as a function of the coefficients of Eq. (14), that we'll collectively denote ζ:

$$\zeta = [\tilde{\alpha}_0, ..., \tilde{\alpha}_{\max(N_1, N_2)}, \tilde{\beta}_0, ..., \tilde{\beta}_{N_2}]. \tag{22}$$

We want to express ζ as a function of the $\tilde{Y}_i$, so that computing the desired derivatives will be straightforward to the reader.

We define a regression matrix Γ whose rows Γ_i for each example in the training set are given by:

$$\Gamma_i = [1, b_i, b_i^2, ..., b^{\max(N_1, N_2)}, Z_i, bZ_i, ..., b_i^{N_2} Z_i], \tag{23}$$

where b_i and Z_i are the balancing scores and protected attribute for the i-th example. The coefficients ζ in OLS regression are then easily found to be:

$$\zeta(\tilde{\mathbf{Y}}) = (\Gamma \Gamma^T)^{-1} \Gamma^T \tilde{\mathbf{Y}} \tag{24}$$

where $\tilde{\mathbf{Y}}$ is the vector of the scores evaluated in each training example.

Data splits, pre-processing and hyperparameters

For the UCI adult dataset, we used the official train/test split and used one-hot encoding for the categorical variables. For the synthetic dataset, we sampled 10^5 points and used 33% for testing. We used scikit-learn's StandardScaler to preprocess all the datasets.

For our logistic regression models we did not use any hyperparameter apart from λ. For our XGBoost models, with reference to the python API, we used the default parameters with the exception of reg_lambda = 10, learning_rate = 0.1 and max_depth = 2.

Propensity Scores

Propensity scores were computed using scikit-learn's LogisticRegression and we used GridSearchCv from the same library to search for the inverse L_1 penalty term $C \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$. We used 5-Fold cross-validation scored by accuracy.

Training procedures

We trained all our models using a double early stopping procedure as follows. For logistic regression, we initialized the weights of every model to the ones of a logistic regression trained with scikit-learn with the default parameters (we used the "liblinear" solver) and subsequently used gradient descent using our modified loss. We also defined an early stopping set, which was the whole training set in the logistic regression case and a separate holdout set for XGBoost. The holdout set was derived using 33% of the training set (we used a numpy.random seed fixed at 123 throughout).

Then, for each value of λ, we used the following procedure:

(1) Define $\lambda^* = \min(\lambda, 0.3)$
(2) Train using the loss $\mathcal{L}_f = (1 - \lambda^*)\mathcal{L}_o + \lambda^* \mathcal{R}_f$ until $\mathcal{L}_o$ does not improve on the early stopping set for 5 steps
(3) Increase λ^* linearly over 50 steps from $\min(\lambda, 0.3)$ to λ
(4) Train using the loss $\mathcal{L}_f = (1 - \lambda)\mathcal{L}_o + \lambda \mathcal{R}_f$ until $\mathcal{L}_f$ does not improve on the early stopping set for 20 steps