When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness

Chris Russell*
The Alan Turing Institute and
University of Surrey
crussell@turing.ac.uk

Matt J. Kusner*
The Alan Turing Institute and
University of Warwick
mkusner@turing.ac.uk

Joshua R. Loftus†
New York University
loftus@nyu.edu

Ricardo Silva
The Alan Turing Institute and
University College London
ricardo@stats.ucl.ac.uk

Abstract

Machine learning is now being used to make crucial decisions about people's lives. For nearly all of these decisions there is a risk that individuals of a certain race, gender, sexual orientation, or any other subpopulation are unfairly discriminated against. Our recent method has demonstrated how to use techniques from counterfactual inference to make predictions fair across different subpopulations. This method requires that one provides the causal model that generated the data at hand. In general, validating all causal implications of the model is not possible without further assumptions. Hence, it is desirable to integrate competing causal models to provide counterfactually fair decisions, regardless of which causal "world" is the correct one. In this paper, we show how it is possible to make predictions that are approximately fair with respect to multiple possible causal models at once, thus mitigating the problem of exact causal specification. We frame the goal of learning a fair classifier as an optimization problem with fairness constraints entailed by competing causal explanations. We show how this optimization problem can be efficiently solved using gradient-based methods. We demonstrate the flexibility of our model on two real-world fair classification problems. We show that our model can seamlessly balance fairness in multiple worlds with prediction accuracy.

1 Introduction

Machine learning algorithms can do extraordinary things with data. From generating realistic images from noise [7], to predicting what you will look like when you become older [18]. Today, governments and other organizations make use of it in criminal sentencing [4], predicting where to allocate police officers [3, 16], and to estimate an individual's risk of failing to pay back a loan [8]. However, in many of these settings, the data used to train machine learning algorithms contains biases against certain races, sexes, or other subgroups in the population [3, 6]. Unwittingly, this discrimination is then reflected in the predictions of such algorithms. Simply being born male or female can change an individual's opportunities that follow from automated decision making trained to reflect historical biases. The implication is that, without taking this into account, classifiers that maximize accuracy risk perpetuating biases present in society.

*Equal contribution.
†This work was done while JL was a Research Fellow at the Alan Turing Institute.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

For instance, consider the rise of 'predictive policing', described as "taking data from disparate sources, analyzing them, and then using the results to anticipate, prevent and respond more effectively to future crime" [16]. Today, 38% of U.S. police departments surveyed by the Police Executive Research Forum are using predictive policing and 70% plan to in the next 2 to 5 years. However, there have been significant doubts raised by researchers, journalists, and activists that if the data used by these algorithms is collected by departments that have been biased against minority groups, the predictions of these algorithms could reflect that bias [9, 12].

At the same time, fundamental mathematical results make it difficult to design fair classifiers. In criminal sentencing the COMPAS score [4] predicts if a prisoner will commit a crime upon release, and is widely used by judges to set bail and parole. While it has been shown that black and white defendants with the same COMPAS score commit a crime at similar rates after being released [1], it was also shown that black individuals were more often incorrectly predicted to commit crimes after release by COMPAS than white individuals were [2]. In fact, except for very specific cases, it is impossible to balance these measures of fairness [3, 10, 20].

The question becomes how to address the fact that the data itself may bias the learning algorithm and even addressing this is theoretically difficult. One promising avenue is a recent approach, introduced by us in [11], called *counterfactual fairness*. In this work, we model how unfairness enters a dataset using techniques from causal modeling. Given such a model, we state whether an algorithm is fair if it would give the same predictions had an individual's race, sex, or other sensitive attributes been different. We show how to formalize this notion using counterfactuals, following a rich tradition of causal modeling in the artificial intelligence literature [15], and how it can be placed into a machine learning pipeline. The big challenge in applying this work is that evaluating a counterfactual e.g., "What if I had been born a different sex?", requires a causal model which describes how your sex changes your predictions, other things being equal.

Using "world" to describe any causal model evaluated at a particular counterfactual configuration, we have dependent "worlds" within a *same* causal model that can never be jointly observed, and possibly incompatible "worlds" across *different* models. Questions requiring the joint distribution of counterfactuals are hard to answer, as they demand partially untestable "cross-world" assumptions [5, 17], and even many of the empirically testable assumptions cannot be falsified from observational data alone [14], requiring possibly infeasible randomized trials. Because of this, different experts as well as different algorithms may disagree about the right causal model. Further disputes may arise due to the conflict between accurately modeling unfair data and producing a fair result, or because some degrees of unfairness may be considered allowable while others are not.

To address these problems, we propose a method for ensuring fairness within multiple causal models. We do so by introducing continuous relaxations of counterfactual fairness. With these relaxations in hand, we frame learning a fair classifier as *an optimization problem with fairness constraints*. We give efficient algorithms for solving these optimization problems for different classes of causal models. We demonstrate on three real-world fair classification datasets how our model is able to simultaneously achieve fairness in multiple models while flexibly trading off classification accuracy.

2 Background

We begin by describing aspects causal modeling and counterfactual inference relevant for modeling fairness in data. We then briefly review counterfactual fairness [11], but we recommend that the interested reader should read the original paper in full. We describe how uncertainty may arise over the correct causal model and some difficulties with the original counterfactual fairness definition. We will use A to denote the set of protected attributes, a scalar in all of our examples but which without loss of generality can take the form of a set. Likewise, we denote as Y the outcome of interest that needs to be predicted using a predictor $\hat{Y}$. Finally, we will use $\mathcal{X}$ to denote the set of observed variables other than A and Y, and $\mathcal{U}$ to denote a set of hidden variables, which without loss of generality can be assumed to have no observable causes in a corresponding causal model.

2.1 Causal Modeling and Counterfactual Inference

We will use the causal framework of Pearl [15], which we describe using a simple example. Imagine we have a dataset of university students and we would like to model the causal relationships that

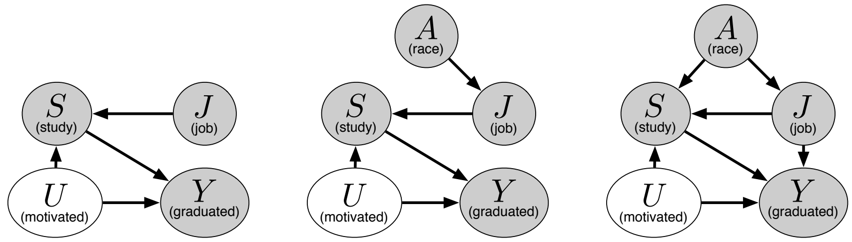

Figure 1: Dark nodes correspond to observed variables and light nodes are unobserved. (*Left*) This model predicts that both study S and motivation U directly cause graduation rate Y. However, this model does not take into account how an individual's race may affect observed variables. (*Center*) In this model, we encode how an individual's race may affect whether they need to have a job J while attending university. (*Right*) We may wonder if there are further biases in society to expect different rates of study for different races. We may also suspect that having a job may influence one's graduation likelihood, independent of study.

lead up to whether a student graduates on time. In our dataset, we have information about whether a student holds a job J, the number of hours they study per week S, and whether they graduate Y. Because we are interested in modeling any unfairness in our data, we also have information about a student's race A. Pearl's framework allows us to model causal relationships between these variables and any postulated unobserved latent variables, such as some U quantifying how motivated a student is to graduate. This uses a directed acyclic graph (DAG) with causal semantics, called a causal diagram. We show a possible causal diagram for this example in Figure 1, (*Left*). Each node corresponds to a variable and each set of edges into a node corresponds to a generative model specifying how the "parents" of that node causally generated it. In its most specific description, this generative model is a functional relationship deterministically generating its output given a set of observed and latent variables. For instance, one possible set of functions described by this model could be as follows:

$$S = g(J, U) + \epsilon \quad Y = \mathbb{I}[\phi(h(S, U)) \geq 0.5] \tag{1}$$

where g, h are arbitrary functions and $\mathbb{I}$ is the indicator function that evaluates to 1 if the condition holds and 0 otherwise. Additionally, ϕ is the logistic function $\phi(a) = 1/(1 + \exp(-a))$ and ϵ is drawn independently of all variables from the standard normal distribution $\mathcal{N}(0, 1)$. It is also possible to specify non-deterministic relationships:

$$U \sim \mathcal{N}(0, 1) \quad S \sim \mathcal{N}(g(J, U), \sigma_S) \quad Y \sim \text{Bernoulli}(\phi(h(S, U)) \tag{2}$$

where σ_S is a model parameter. The power of this causal modeling framework is that, given a fully-specified set of equations, we can compute what (the distribution of) any of the variables would have been *had certain other variables been different, other things being equal*. For instance, given the causal model we can ask "Would individual i have graduated ($Y = 1$) if they hadn't had a job?", even if they did not actually graduate in the dataset. Questions of this type are called *counterfactuals*.

For any observed variables V, W we denote the value of the counterfactual "What would V have been if W had been equal to w?" as $V_{W \leftarrow w}$. Pearl et al. [15] describe how to compute these counterfactuals (or, for non-deterministic models, how to compute their distribution) using three steps: 1. **Abduction**: Given the set of observed variables $\mathcal{X} = \{X_1, \ldots, X_d\}$ compute the values of the set of unobserved variables $\mathcal{U} = \{U_1, \ldots, U_p\}$ given the model (for non-deterministic models, we compute the posterior distribution $\mathbb{P}(\mathcal{U}|\mathcal{X})$); 2. **Action**: Replace all occurrences of the variable W with value w in the model equations; 3. **Prediction**: Using the new model equations, and $\mathcal{U}$ (or $\mathbb{P}(\mathcal{U}|\mathcal{X})$) compute the value of V (or $P(V|\mathcal{X})$). This final step provides the value or distribution of $V_{W \leftarrow w}$ given the observed, factual, variables.

2.2 Counterfactual Fairness

In the above example, the university may wish to predict Y, whether a student will graduate, in order to determine if they should admit them into an honors program. While the university prefers to admit students who will graduate on time, it is willing to give a chance to some students without a confident graduation prediction in order to remedy unfairness associated with race in the honors program. The university believes that whether a student needs a job J may be influenced by their race. As evidence they cite the National Center for Education Statistics, which reported[3] that fewer (25%) Asian-American students were employed while attending university as full-time students relative to students of other races (at least 35%). We show the corresponding casual diagram for this in Figure 1 (*Center*). As having a job J affects study which affects graduation likelihood Y this may mean different races take longer to graduate and thus unfairly have a harder time getting into the honors program.

Counterfactual fairness aims to correct predictions of a label variable Y that are unfairly altered by an individual's sensitive attribute A (race in this case). Fairness is defined in terms of counterfactuals:

Definition 1 (Counterfactual Fairness [11]). *A predictor $\hat{Y}$ of Y is **counterfactually fair** given the sensitive attribute $A = a$ and any observed variables $\mathcal{X}$ if*

$$\mathbb{P}(\hat{Y}_{A\leftarrow a} = y \mid \mathcal{X} = \mathbf{x}, A = a) = \mathbb{P}(\hat{Y}_{A\leftarrow a'} = y \mid \mathcal{X} = \mathbf{x}, A = a) \tag{3}$$

for all y and $a' \neq a$.

In what follows, we will also refer to $\hat{Y}$ as a function $f(\mathbf{x}, a)$ of hidden variables $\mathcal{U}$, of (usually a subset of) an instantiation $\mathbf{x}$ of $\mathcal{X}$, and of protected attribute A. We leave $\mathcal{U}$ implicit in this notation since, as we will see, this set might differ across different competing models. The notation implies

$$\hat{Y}_{A\leftarrow a} = f(\mathbf{x}_{A\leftarrow a}, a). \tag{4}$$

Notice that if counterfactual fairness holds exactly for $\hat{Y}$, then this predictor can only be a non-trivial function of $\mathcal{X}$ for those elements $X \in \mathcal{X}$ such that $X_{A\leftarrow a} = X_{A\leftarrow a'}$. Moreover, by construction $\mathcal{U}_{A\leftarrow a} = \mathcal{U}_{\mathcal{A}\leftarrow a'}$, as each element of $\mathcal{U}$ is defined to have no causes in $A \cup \mathcal{X}$.

The probabilities in eq. (3) are given by the posterior distribution over the unobserved variables $\mathbb{P}(\mathcal{U} \mid \mathcal{X} = \mathbf{x}, A = a)$. Hence, a counterfactual $\hat{Y}_{A\leftarrow a}$ may be *deterministic* if this distribution is degenerate, that is, if $\mathcal{U}$ is a deterministic function of $\mathcal{X}$ and A. One nice property of this definition is that it is easy to interpret: a decision is fair if it would have been the same had a person had a different A (e.g., a different race[4]), other things being equal. In [11], we give an efficient algorithm for designing a predictor that is counterfactually fair. In the university graduation example, a predictor constructed from the unobserved motivation variable U is counterfactually fair.

One difficulty of the definition of counterfactual fairness is it requires one to postulate causal relationships between variables, including latent variables that may be impractical to measure directly. In general, different causal models will create different fair predictors $\hat{Y}$. But there are several reasons why it may be unrealistic to assume that any single, fixed causal model will be appropriate. There may not be a consensus among experts or previous literature about the existence, functional form, direction, or magnitude of a particular causal effect, and it may be impossible to determine these from the available data without untestable assumptions. And given the sensitive, even political nature of problems involving fairness, it is also possible that disputes may arise over the presence of a feature of the causal model, based on competing notions of dependencies and latent variables. Consider the following example, formulated as a dispute over the presence of edges. For the university graduation model, one may ask if differences in study are due only to differences in employment, or whether instead there is some other direct effect of A on study levels. Also, having a job may directly affect graduation likelihood. We show these changes to the model in Figure 1 (*Right*). There is also potential for disagreement over whether some causal paths from A to graduation should be excluded from the definition of fairness. For example, an adherent to strict meritocracy may argue the numbers of hours a student has studied should not be given a counterfactual value. This could be incorporated in a separate model by omitting chosen edges when propagating counterfactual information through the graph in the **Prediction** step of counterfactual inference[5]. To summarize, there may be disagreements about the right causal model due to: 1. Changing the structure of the DAG, e.g. adding an edge; 2. Changing the latent variables, e.g. changing the function generating a vertex to have a different signal vs. noise decomposition; 3. Preventing certain paths from propagating counterfactual values.

[3] `https://nces.ed.gov/programs/coe/indicator_ssa.asp`

[4] At the same time, the notion of a "counterfactual race," sex, etc. often raises debate. See [11] for our take on this.

[5] In the Supplementary Material of [11], we explain how counterfactual fairness can be restricted to particular paths from A to Y, as opposed to all paths.

3 Fairness under Causal Uncertainty

In this section, we describe a technique for learning a fair predictor without knowing the true casual model. We first describe why in general counterfactual fairness will often not hold in multiple different models. We then describe a relaxation of the definition of counterfactual fairness for both deterministic and non-deterministic models. Finally we show an efficient method for learning classifiers that are simultaneously accurate and fair in multiple worlds. In all that follows we denote sets in calligraphic script $\mathcal{X}$, random variables in uppercase X, scalars in lowercase x, matrices in bold uppercase $\mathbf{X}$, and vectors in bold lowercase $\mathbf{x}$.

3.1 Exact Counterfactual Fairness Across Worlds

We can imagine extending the definition of counterfactual fairness so that it holds for every plausible causal world. To see why this is inherently difficult consider the setting of deterministic causal models. If each causal model of the world generates different counterfactuals then each additional model induces a new set of constraints that the classifier must satisfy, and in the limit the only classifiers that are fair across all possible worlds are constant classifiers. For non-deterministic counterfactuals, these issues are magnified. To guarantee counterfactual fairness, Kusner et al. [11] assumed access to latent variables that hold the same value in an original datapoint and in its corresponding counterfactuals. While the latent variables of one world can remain constant under the generation of counterfactuals from its corresponding model, there is no guarantee that they remain constant under the counterfactuals generated from different models. Even in a two model case, if the P.D.F. of one model's counterfactual has non-zero density everywhere (as is the case under Gaussian noise assumptions) it may be the case that the only classifiers that satisfy counterfactual fairness for both worlds are the constant classifiers. If we are to achieve some measure of fairness from informative classifiers, and over a family of different worlds, we need a more robust alternative to counterfactual fairness.

3.2 Approximate Counterfactual Fairness

We define two approximations to counterfactual fairness to solve the problem of learning a fair classifier across multiple causal worlds.

Definition 2 ((ϵ, δ)-Approximate Counterfactual Fairness)**.** *A predictor $f(\mathcal{X}, A)$ satisfies $(\epsilon, 0)$-**approximate counterfactual fairness** ($(\epsilon, 0)$-ACF) if, given the sensitive attribute $A = a$ and any instantiation $\mathbf{x}$ of the other observed variables $\mathcal{X}$, we have that:*

$$\left| f(\mathbf{x}_{A \leftarrow a}, a) - f(\mathbf{x}_{A \leftarrow a'}, a') \right| \leq \epsilon \tag{5}$$

*for all $a' \neq a$ if the system deterministically implies the counterfactual values of $\mathcal{X}$. For a non-deterministic causal system, f satisfies (ϵ, δ)-**approximate counterfactual fairness**, ((ϵ, δ)-ACF) if:*

$$\mathbb{P}(\left| f(\mathcal{X}_{A \leftarrow a}, a) - f(\mathcal{X}_{A \leftarrow a'}, a') \right| \leq \epsilon \mid \mathcal{X} = \mathbf{x}, A = a) \geq 1 - \delta \tag{6}$$

for all $a' \neq a$.

Both definitions must hold uniformly over the sample space of $\mathcal{X} \times A$. The probability measures used are with respect to the conditional distribution of background latent variables $\mathcal{U}$ given the observations. We leave a discussion of the statistical asymptotic properties of such plug-in estimator for future work. These definitions relax counterfactual fairness to ensure that, for deterministic systems, predictions f change by at most ϵ when an input is replaced by its counterfactual. For non-deterministic systems, the condition in (6) means that this ϵ change must occur with high probability, where the probability is again given by the posterior distribution $\mathbb{P}(\mathcal{U}|\mathcal{X})$ computed in the **Abduction** step of counterfactual inference. If $\epsilon = 0$, the deterministic definitions eq. (5) is equivalent to the original counterfactual fairness definition. If also $\delta = 0$ the non-deterministic definition eq. (6) is actually a stronger condition than the counterfactual fairness definition eq. (3) as it guarantees equality in probability instead of equality in distribution[6].

[6] In the Supplementary Material of [11], we describe in more detail the implications of the stronger condition.

Algorithm 1 Multi-World Fairness

Input: features $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, labels $\mathbf{y} = [y_1, \ldots, y_n]$, sensitive attributes $\mathbf{a} = [a_1, \ldots, a_n]$, privacy parameters (ϵ, δ), trade-off parameters $\mathcal{L} = [\lambda_1, \ldots, \lambda_l]$.
Fit causal models: $\mathbf{M}_1, \ldots, \mathbf{M}_m$ using $\mathbf{X}, \mathbf{a}$ (and possibly $\mathbf{y}$).
Sample counterfactuals: $\mathcal{X}_{A^1 \leftarrow a'}, \ldots, \mathcal{X}_{A^m \leftarrow a'}$ for all unobserved values a'.
for $\lambda \in \mathcal{L}$ **do**
  Initialize classifier f_λ.
  while loop until convergence **do**
    Select random batches $\mathbf{X}_b$ of inputs and batch of counterfactuals $\mathbf{X}_{A^1 \leftarrow a'}, \ldots, \mathbf{X}_{A^m \leftarrow a'}$.
    Compute the gradient of equation (7).
    Update f_λ using any stochastic gradient optimization method.
  end while
end for
Select model f_λ: For deterministic models select the smallest λ such that equation (5) using f_λ holds. For non-deterministic models select the λ that corresponds to δ given f_λ.

3.3 Learning a Fair Classifier

Assume we are given a dataset of n observations $\mathbf{a} = [a_1, \ldots, a_n]$ of the sensitive attribute A and of other features $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ drawn from $\mathcal{X}$. We wish to accurately predict a label Y given observations $\mathbf{y} = [y_1, \ldots, y_n]$ while also satisfying (ϵ, δ)-approximate counterfactual fairness. We learn a classifier $f(\mathbf{x}, a)$ by minimizing a loss function $\ell(f(\mathbf{x}, a), y)$. At the same time, we incorporate an unfairness term $\mu_j(f, \mathbf{x}, a, a')$ for each causal model j to reduce the unfairness in f. We formulate this as a penalized optimization problem:

$$\min_f \frac{1}{n} \sum_{i=1}^n \ell(f(\mathbf{x}_i, a_i), y_i) + \lambda \sum_{j=1}^m \frac{1}{n} \sum_{i=1}^n \sum_{a' \neq a_i} \mu_j(f, \mathbf{x}_i, a_i, a') \tag{7}$$

where λ trades-off classification accuracy for multi-world fair predictions. We show how to naturally define the unfairness function μ_j for deterministic and non-deterministic counterfactuals.

Deterministic counterfactuals. To enforce $(\epsilon, 0)$-approximate counterfactual fairness a natural penalty for unfairness is an indicator function which is one whenever $(\epsilon, 0)$-ACF does not hold, and zero otherwise:

$$\mu_j(f, \mathbf{x}_i, a_i, a') := \mathbb{I}[\big|f(\mathbf{x}_{i,A^j \leftarrow a_i}, a_i) - f(\mathbf{x}_{i,A^j \leftarrow a'}, a')\big| \geq \epsilon] \tag{8}$$

Unfortunately, the indicator function $\mathbb{I}$ is non-convex, discontinuous and difficult to optimize. Instead, we propose to use the tightest convex relaxation to the indicator function:

$$\mu_j(f, \mathbf{x}_i, a_i, a') := \max\{0, \big|f(\mathbf{x}_{i,A^j \leftarrow a_i}, a_i) - f(\mathbf{x}_{i,A^j \leftarrow a'}, a')\big| - \epsilon\} \tag{9}$$

Note that when $(\epsilon, 0)$-approximate counterfactual fairness is not satisfied μ_j is non-zero and thus the optimization problem will penalize f for this unfairness. Where $(\epsilon, 0)$-approximate counterfactual fairness is satisfied μ_j evaluates to 0 and it does not affect the objective. For sufficiently large λ, the value of μ_j will dominate the training loss $\frac{1}{n}\sum_{i=1}^n \ell(f(\mathbf{x}_i, a_i), y_i)$ and any solution will satisfy $(\epsilon, 0)$-approximate counterfactual fairness. However, an overly large choice of λ causes numeric instability, and will decrease the accuracy of the classifier found. Thus, to find the most accurate classifier that satisfies the fairness condition one can simply perform a grid or binary search for the smallest λ such that the condition holds.

Non-deterministic counterfactuals. For non-deterministic counterfactuals we begin by writing a Monte-Carlo approximation to (ϵ, δ)-ACF, eq. (6) as follows:

$$\frac{1}{S} \sum_{s=1}^S \mathbb{I}(\big|f(\mathbf{x}^s_{A^j \leftarrow a_i}, a_i) - f(\mathbf{x}^s_{A^j \leftarrow a'}, a')\big| \geq \epsilon) \leq \delta \tag{10}$$

where $\mathbf{x}^k$ is sampled from the posterior distribution $\mathbb{P}(\mathcal{U}|\mathcal{X})$. We can again form the tightest convex relaxation of the left-hand side of the expression to yield our unfairness function:

$$\mu_j(f, \mathbf{x}_i, a_i, a') := \frac{1}{S} \sum_{s=1}^S \max\{0, \big|f(\mathbf{x}^s_{i,A^j \leftarrow a_i}, a_i) - f(\mathbf{x}^s_{i,A^j \leftarrow a'}, a')\big| - \epsilon\} \tag{11}$$

Note that different choices of λ in eq. (7) correspond to different values of δ. Indeed, by choosing $\lambda = 0$ we have the (ϵ, δ)-fair classifier corresponding to an unfair classifier[7]. While a sufficiently large, but finite, λ will correspond to a $(\epsilon, 0)$ approximately counterfactually fair classifier. By varying λ between these two extremes, we induce classifiers that satisfy (ϵ, δ)-ACF for different values of δ.

With these unfairness functions we have a differentiable optimization problem eq. (7) which can be solved with gradient-based methods. Thus, our method allows practitioners to smoothly trade-off accuracy with multi-world fairness. We call our method *Multi-World Fairness* (MWF). We give a complete method for learning a MWF classifier in Algorithm 1.

For both deterministic and non-deterministic models, this convex approximation essentially describes an expected unfairness that is allowed by the classifier:

Definition 3 (Expected ϵ-Unfairness)**.** *For any counterfactual $a' \neq a$, the Expected ϵ-Unfairness of a classifier f, or $\mathbb{E}_\epsilon[f]$, is*

$$\mathbb{E}\Big[\max\{0, \big|f(\mathcal{X}_{A\leftarrow a}, a) - f(\mathcal{X}_{A\leftarrow a'}, a')\big| - \epsilon\} \mid \mathcal{X} = \mathbf{x}, A = a\Big] \tag{12}$$

where the expectation is over any unobserved U (and is degenerate for deterministic counterfactuals). We note that the term $\max\{0, \big|f(\mathcal{X}_{A\leftarrow a}, a) - f(\mathcal{X}_{A\leftarrow a'}, a')\big| - \epsilon\}$ is strictly non-negative and therefore the expected ϵ-unfairness is zero if and only if f satisfies $(\epsilon, 0)$-approximate counterfactual fairness almost everywhere.

Linear Classifiers and Convexity Although we have presented these results in their most general form, it is worth noting that for linear classifiers, convexity guarantees are preserved. The family of linear classifiers we consider is relatively broad, and consists those linear in their learned weights w, as such it includes both SVMs and a variety of regression methods used in conjuncture with kernels or finite polynomial bases.

Consider any classifier whose output is linear in the learned parameters, i.e., the family of classifiers f all have the form $f(\mathcal{X}, A) = \sum_l w_l g_l(\mathcal{X}, a)$, for a set of fixed kernels g_l. Then the expected ϵ-unfairness is a linear function of w taking the form:

$$\begin{aligned} &\mathbb{E}\Big[\max\{0, \big|f(\mathcal{X}_{A\leftarrow a}, a) - f(\mathcal{X}_{A\leftarrow a'}, a')\big| - \epsilon\}\Big] \qquad (13) \\ =\ &\mathbb{E}\Big[\max\{0, \big|\sum_l w_l(g_l(\mathcal{X}_{A\leftarrow a}, a) - g_l(\mathcal{X}_{A\leftarrow a'}, a'))\big|\}\Big] \end{aligned}$$

This expression is linear in w and therefore, if the classification loss is also convex (as is the case for most regression tasks), a global optima can be ready found via convex programming. In particular, globally optimal linear classifiers satisfying $(\epsilon, 0)$-ACF or (ϵ, δ)-ACF, can be found efficiently.

Bayesian alternatives and their shortcomings. One may argue that a more direct alternative is to provide probabilities associated with each world and to marginalize set of the optimal counterfactually fair classifiers over all possible worlds. We argue this is undesirable for two reasons: first, the averaged prediction for any particular individual may violate (3) by an undesirable margin for one, more or even *all* considered worlds; second, a practitioner may be restricted by regulations to show that, to the best of their knowledge, the worst-case violation is bounded across all viable worlds with high probability. However, if the number of possible models is extremely large (for example if the causal structure of the world is known, but the associated parameters are not) and we have a probability associated with each world, then one natural extension is to adapt Expected ϵ-Unfairness eq. (3) to marginalize over the space of possible worlds. However, we leave this extension to future work.

4 Experiments

We demonstrate the flexibility of our method on two real-world fair classification problems: 1. fair predictions of student performance in law schools; and 2. predicting whether criminals will re-offend upon being released. For each dataset we begin by giving details of the fair prediction problem. We then introduce multiple causal models that each possibly describe how unfairness plays a role in the data. Finally, we give results of *Multi-World Fairness* (MWF) and show how it changes for different settings of the fairness parameters (ϵ, δ).

[7] In the worst case, δ may equal 1.

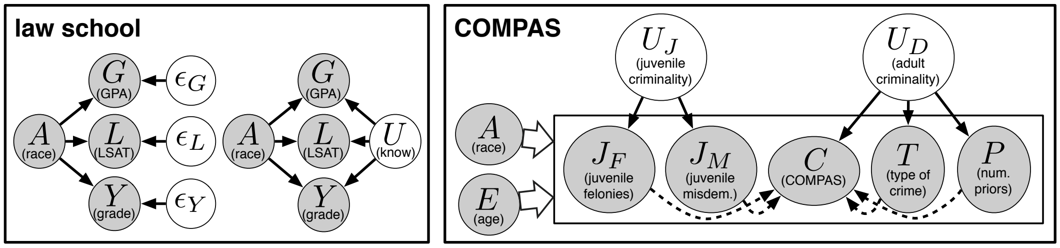

Figure 2: Causal models for the law school and COMPAS datasets. Shaded nodes are observed an unshaded nodes are unobserved. For each dataset we consider two possible causal worlds. The first law school model is a deterministic causal model with additive unobserved variables $\epsilon_G, \epsilon_L, \epsilon_Y$. The second is a non-deterministic causal model with a latent variable U. For COMPAS, the first causal model omits the dotted lines, and the second includes them. Both models are non-deterministic models with latent variables U_J, U_D. The large white arrows signify that variables A, E are connected to every variable contained in the box they point to. The law school model equations are given in eq. (14) and COMPAS model equations are shown in eq. (15).

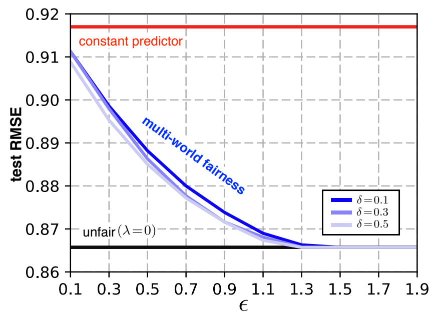

Figure 3: Test prediction results for different ϵ on the law school dataset.

4.1 Fairly predicting law grades

We begin by investigating a dataset of survey results across 163 U.S. law schools conducted by the Law School Admission Council [19] . It contains information on over 20,000 students including their race A (here we look at just black and white students as this difference had the largest effect in counterfactuals in [11]), their grade-point average G obtained prior to law school, law school entrance exam scores L, and their first year average grade Y. Consider that law schools may be interested in predicting Y for all applicants to law school using G and L in order to decide whether to accept or deny them entrance. However, due to societal inequalities, an individual's race may have affected their access to educational opportunities, and thus affected G and L. Accordingly, we model this possibility using the causal graphs in Figure 2 (*Left*). In this graph we also model the fact that G, L may have been affected by other unobserved quantities. However, we may be uncertain whether what the right way to model these unobserved quantities is. Thus we propose to model this dataset with the two worlds described in Figure 2 (*Left*). Note that these are the same models as used in Kusner et al. [11] (except here we consider race as the sensitive variable). The corresponding equations for these two worlds are as follows:

$$\begin{aligned}
G &= b_G + w_G^A A + \epsilon_G & \quad G &\sim \mathcal{N}(b_G + w_G^A A + w_G^U U, \sigma_G) \\
L &= b_L + w_L^A A + \epsilon_L & \quad L &\sim \text{Poisson}(\exp(b_L + w_L^A A + w_L^U U)) \\
Y &= b_Y + w_Y^A A + \epsilon_Y & \quad Y &\sim \mathcal{N}(w_Y^A A + w_Y^U U, 1) \\
\epsilon_G, \epsilon_L, \epsilon_Y &\sim \mathcal{N}(0,1) & \quad U &\sim \mathcal{N}(0,1)
\end{aligned} \tag{14}$$

where variables b, w are parameters of the causal model.

Results. Figure 3 shows the result of learning a linear MWF classifier on the deterministic law school models. We split the law school data into a random 80/20 train/test split and we fit casual models and classifiers on the training set and evaluate performance on the test set. We plot the test RMSE of the constant predictor satisfying counterfactual fairness in red, the unfair predictor with $\lambda = 0$, and MWF, averaged across 5 runs. Here as we have one deterministic and one non-deterministic model we will evaluate MWF for different ϵ and δ (with the knowledge that the only change in the MWF classifier for different δ is due to the non-deterministic model). For each ϵ, δ, we selected the smallest λ across a grid ($\lambda \in \{10^{-5} 10^{-4}, \ldots, 10^{10}\}$) such that the constraint in eq. (6) held across 95% of the individuals in both models. We see that MWF is able to reliably sacrifice accuracy for fairness as ϵ is reduced. Note that as we change δ we can further alter the accuracy/fairness trade-off.

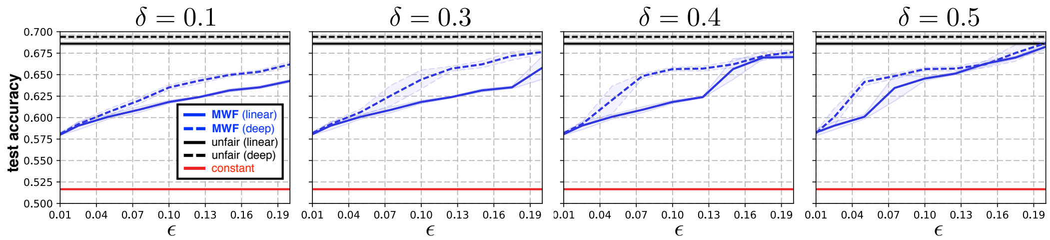

Figure 4: Test prediction results for different ϵ and δ on the COMPAS dataset.

4.2 Fair recidivism prediction (COMPAS)

We next turn our attention to predicting whether a criminal will re-offend, or 'recidivate' after being released from prison. ProPublica [13] released data on prisoners in Broward County, Florida who were awaiting a sentencing hearing. For each of the prisoners we have information on their race A (as above we only consider black versus white individuals), their age E, their number of juvenile felonies J_F, juvenile misdemeanors J_M, the type of crime they committed T, the number of prior offenses they have P, and whether they recidivated Y. There is also a proprietary COMPAS score [13] C designed to indicate the likelihood a prisoner recidivates.

We model this dataset with two different non-deterministic causal models, shown in Figure 2 (*Right*). The first model includes the dotted edges, the second omits them. In both models we believe that two unobserved latent factors juvenile criminality U_J and adult criminality U_D also contribute to J_F, J_M, C, T, P. We show the equations for both of our casual models below, where the first causal model includes the blue terms and the second does not:

$$
\begin{aligned}
T &\sim \text{Bernoulli}(\phi(b_T + w_C^{U_D} U_D + w_C^E E + w_C^A A) \\
C &\sim \mathcal{N}(b_C + w_C^{U_D} U_D + w_C^E E + w_C^A A + w_C^T T + w_C^P P + w_C^{J_F} J_F + w_C^{J_M} J_M, \sigma_C) \\
P &\sim \text{Poisson}(\exp(b_P + w_P^{U_D} U_D + w_P^E E + w_P^A A)) \\
J_F &\sim \text{Poisson}(\exp(b_{J_F} + w_{J_F}^{U_J} + w_{J_F}^E E + w_{J_F}^A A)) \\
J_M &\sim \text{Poisson}(\exp(b_{J_M} + w_{J_M}^{U_J} + w_{J_M}^E E + w_{J_M}^A A)) \\
[U_J, U_D] &\sim \mathcal{N}(0, \Sigma)
\end{aligned} \tag{15}
$$

Results. Figure 4 shows how classification accuracy using both logistic regression (linear) and a 3-layer neural network (deep) changes as both ϵ and δ change. We split the COMPAS dataset randomly into an 80/20 train/test split, and report all results on the test set. As in the law school experiment we grid-search over λ to find the smallest value such that for any ϵ and δ the (ϵ, δ)-ACF) constraint in eq. (6) is satisfied for at least 95% of the individuals in the dataset, across both worlds. We average all results except the constant classifier over 5 runs and plot the mean and standard deviations. We see that for small δ (high fairness) both linear and deep MWF classifiers significantly outperform the constant classifier and begin to approach the accuracy of the unfair classifier as ϵ increases. As we increase δ (lowered fairness) the deep classifier is better able to learn a decision boundary that trades-off accuracy for fairness. But if ϵ, δ is increased enough (e.g., $\epsilon \geq 0.13, \delta = 0.5$), the linear MWF classifier matches the performance of the deep classifier.

5 Conclusion

This paper has presented a natural extension to counterfactual fairness that allows us to guarantee fair properties of algorithms, even when we are unsure of the causal model that describes the world.

As the use of machine learning becomes widespread across many domains, it becomes more important to take algorithmic fairness out of the hands of experts and make it available to everybody. The conceptual simplicity of our method, our robust use of counterfactuals, and the ease of implementing our method mean that it can be directly applied to many interesting problems. A further benefit of our approach over previous work on counterfactual fairness is that our approach only requires the estimation of counterfactuals at training time, and no knowledge of latent variables during testing. As such, our classifiers offer a fair drop-in replacement for other existing classifiers.

6 Acknowledgments

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. CR acknowledges additional support under the EPSRC Platform Grant EP/P022529/1.

References

[1] Compas risk scales: Demonstrating accuracy equity and predictive parity performance of the compas risk scales in broward county, 2016. 2

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, 2016. Accessed: Fri 19 May 2017. 2

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207*, 2017. 1, 2

[4] Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40, 2009. 1, 2

[5] A. P. Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, pages 407–448, 2000. 2

[6] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012. 1

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 1

[8] Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010. 1

[9] Keith Kirkpatrick. It's not the algorithm, it's the data. *Communications of the ACM*, 60(2):21–23, 2017. 2

[10] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016. 2

[11] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 31, 2017. 2, 4, 5, 8

[12] Moish Kutnowski. The ethical dangers and merits of predictive policing. *Journal of Community Safety and Well-Being*, 2(1):13–17, 2017. 2

[13] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 2016. 9

[14] David Lopez-Paz. From dependence to causation. *arXiv preprint arXiv:1607.03300*, 2016. 2

[15] J. Pearl, M. Glymour, and N. Jewell. *Causal Inference in Statistics: a Primer*. Wiley, 2016. 2, 3

[16] Beth Pearsall. Predictive policing: The future of law enforcement. *National Institute of Justice Journal*, 266(1):16–19, 2010. 1, 2

[17] T.S. Richardson and J. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Working Paper Number 128, Center for Statistics and the Social Sciences, University of Washington*, 2013. 2

[18] Paul Upchurch, Jacob Gardner, Kavita Bala, Robert Pless, Noah Snavely, and Kilian Weinberger. Deep feature interpolation for image content changes. *arXiv preprint arXiv:1611.05507*, 2016. 1

[19] Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998. 8

[20] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *arXiv preprint arXiv:1610.08452*, 2016. 2