When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness

Chris Russell[*]
University of Surrey
Alan Turing Institute
crussell@turing.ac.uk

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

Joshua R. Loftus
Alan Turing Institute
University of Cambridge
jloftus@turing.ac.uk

Ricardo Silva
University College London
Alan Turing Institute
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. A 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 genera, validating the causal model is impossible using observational data alone, without further assumptions. Hence, it is desirable to integrate competing causal models to provide counterfactually fair decisions, regardless of which "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 bypassing the problem of exact causal specification. We frame the goal of learning a fair classifier as an optimization problem with fairness constraints. We provide two techniques to solve the optimization problem: one via a convex program and another using Lagrange multipliers. 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[6], to predicting what you will look like when you become older [17]. 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[7]. 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, 5]. Unwittingly, this discrimination is then reflected in the predictions of machine learning algorithms. Simply being born male or female changes the predictions of a machine learning algorithm. Without taking this into account, classifiers that maximize accuracy risk perpetuating biases present in society.

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

[*] Equal contribution.

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 [8, 11].

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, 9, 19].

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 called *counterfactual fairness* [10]. In this work, the authors model how unfairness enters a dataset using techniques from causal modeling. Given such a model they state an algorithm is fair if it would give the same predictions had an individual's race, sex, or other sensitive attributes been different. They show how to formalize this notion using counterfactuals. 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 of the world, which describes how your sex changes your predictions.

Without further untestable assumptions, reconstructing the correct causal model from observational data alone is impossible [13]. Because of this, different analysts 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 [10]. We describe how uncertainty may arise over the correct causal model and some difficulties with the original counterfactual fairness definition.

2.1 Causal Modeling and Counterfactual Inference

We will use the causal framework of Pearl [14], 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 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 relationship between these variables and any unobserved latent variables, such as how motivated a student is to graduate U, using a directed acyclic graph (DAG), 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 edge specifies a set of functional relationships between the variables. For instance, one possible set of functions described by this model could be as follows:

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

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 independent of all variables from the standard normal distribution $\mathcal{N}(0, 1)$. It is also possible

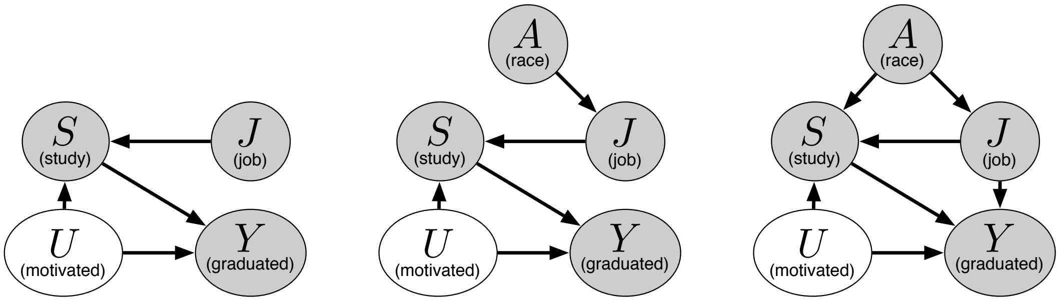

Figure 1: Dark nodes correspond to observed variables and light nodes are unobserved. (*Left*) This model predicts that both study S and motivation U directly cause graduation rate Y. However, this model does not take into account how an individual's race may affect observed variables. (*Center*) In this model we model 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. Each of which we can model easily.

to specify non-deterministic relationships:

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

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 any of the variables would have been *had certain other variables been different*. 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 [15] describes 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 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 value is an estimate of $V_{W\leftarrow w}$.

2.2 Counterfactual Fairness

In the above example, the university may wish to predict whether a student will graduate Y 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[2] 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.

A recent approach called *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 [10]). *A predictor $f(\mathcal{X}, A)$ of Y is* ***counterfactually fair*** *given the sensitive attribute $A=a$ and any remaining observed variables $\mathcal{X}$ if, for a deterministic causal system we have that:*

$$f(\mathcal{X}_{A\leftarrow a}, a) = f(\mathcal{X}_{A\leftarrow a'}, a') \tag{1}$$

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

for all $a' \neq a$. For a non-deterministic causal system, f is ***counterfactually fair*** *if:*

$$\mathbb{P}(f(\mathcal{X}_{A\leftarrow a}, a)=y) = \mathbb{P}(f(\mathcal{X}_{A\leftarrow a'}, a')=y) \tag{2}$$

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

The probabilities in eq. (1) are given by the posterior distribution over the unobserved variables $\mathbb{P}(\mathcal{U}|\mathcal{X})$. 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). The authors give an efficient algorithm for designing a predictor f 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 know the true causal relationships between variables. In general different causal models will create different fair predictors f. But there are several reasons it may be unrealistic to assume 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 an edge in the graph, based on competing notions of what is fair. For instance, for the university graduation model one may ask if differences in study are due only to differences in employment, or instead is there is some other direct effect of A on study? 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. 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 functions along an edge; 3. Preventing certain paths from propagating counterfactual values.

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 is Unsatisfiable

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 the causal models generate different values for counterfactual quantities then it will be impossible to satisfy counterfactual fairness with any classifier except a constant classifier. This is because if we observe $A=a$ and $\mathcal{X}=\mathbf{x}$ then the 'consistency rule' of counterfactuals [15] states that $\mathcal{X}_{A\leftarrow a}=\mathbf{x}$ (i.e., the value of the counterfactual of $\mathcal{X}$ when we set A to its observed value a is just the vector of observed values $\mathbf{x}$). Denote the value of $\mathcal{X}$ for counterfactual $A=a'$ in world i as $\mathcal{X}_{A^i\leftarrow a'}$. Then, counterfactual fairness will dictate that the counterfactuals in different worlds must have the same prediction: $f(\mathcal{X}_{A^1\leftarrow a'}, a')=f(\mathcal{X}_{A^2\leftarrow a'}, a')$. If the system of equations that describes the causal model in each world is not identical (or reducible to one another) these predictions will be different. A similar argument follows for non-deterministic counterfactual fairness. Thus, if we hope to achieve fairness across different worlds postulated by different worlds, we need an alternative to counterfactual fairness.

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 $(\epsilon, \delta)$, 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 $f_\lambda$.
  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 (5).
    Update $f_\lambda$ using any stochastic gradient optimization method.
  end while
end for
Select model $f_\lambda$: For deterministic models select the smallest $\lambda$ such that equation (3) using $f_\lambda$ holds. For non-deterministic models select the $\lambda$ that corresponds to $\delta$ given $f_\lambda$.
```

3.2 An Approximation

We introduce an approximation 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** if given the sensitive attribute $A = a$ and any other observed variables $\mathcal{X}$, for a deterministic causal system we have that:*

$$\left| f(\mathcal{X}_{A \leftarrow a}, a) - f(\mathcal{X}_{A \leftarrow a'}, a') \right| \leq \epsilon \tag{3}$$

*for all $a' \neq a$. For a non-deterministic causal system, f satisfies (ϵ, δ)-**approximate counterfactual fairness** if:*

$$\mathbb{P}(\left| f(\mathcal{X}_{A \leftarrow a}, a) - f(\mathcal{X}_{A \leftarrow a'}, a') \right| \leq \epsilon) \geq 1 - \delta \tag{4}$$

for all $a' \neq a$.

These definitions relax counterfactual fairness to ensure that for deterministic systems predictions of f change by at most ϵ when an input is replaced by its counterfactual. For non-deterministic systems 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 definition eq. (3) is equivalent to the counterfactual fairness definition eq. (1). If also $\delta = 0$ the non-deterministic definition eq. (4) is actually a stronger condition than the counterfactual fairness definition eq. (2) as it guarantees equality in probability instead of equality in distribution[3].

Bayesian alternatives and their shortcomings. One may argue that a more direct alternative is to provide probabilities associated with each world and to return the corresponding weighted prediction. We argue this is undesirable for two reasons: first, there is no guarantee that the averaged prediction for any particular individual violates (1) or (2) by an undesirable margin for one, more or even *all* provided 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.

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 $\mu_j(f, \mathbf{x}, a)$ for each causal model j to reduce the unfairness in f. We

[3] The authors of counterfactual fairness [10] actually describe a technique to learn a classifier that also satisfies this stronger condition.

formulate this as a constrained optimization problem:

$$\min_{f} \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i, a_i), y_i) + \lambda \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_j(f, \mathbf{x}_i, a_i) \tag{5}$$

where λ smoothly 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 satisfy $(\epsilon, 0)$-approximate counterfactual fairness a natural definition of unfairness is when the condition does not hold:

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

Unfortunately, the indicator function $\mathbb{I}$ is not continuous and so is very difficult to optimize. Instead, we propose to minimize a convex relaxation of the indicator function:

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

Note that when $(\epsilon, 0)$-approximate counterfactual fairness is not satisfied μ_j will be non-zero and thus the optimization problem will penalize f for this unfairness. If the definition is satisfied μ_j will evaluate to 0 and it will not affect the objective. Note that $(\epsilon, 0)$-approximate counterfactual fairness will only hold for every observation $i = \{1, \ldots, n\}$ for certain values of λ. 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. Note that we can write a Monte-Carlo approximation to eq. (4) as follows: $\frac{1}{s}\sum_{k=1}^{s} \mathbb{I}(\big|f(\mathbf{x}^k_{A^j \leftarrow a_i}, a_i) - f(\mathbf{x}^k_{A^j \leftarrow a'}, a')\big| \leq \epsilon) \geq 1 - \delta$ where $\mathbf{x}^k$ is sampled from the posterior distribution $\mathbb{P}(\mathcal{U}|\mathcal{X})$. Thus a natural unfairness function for (ϵ, δ)-approximate counterfactual fairness is:

$$\mu_j(f, \mathbf{x}_i, a_i) := \frac{1}{s} \sum_{k=1}^{s} \mathbb{I}[\big|f(\mathbf{x}_{i,A^j \leftarrow a_i}, a_i) - f(\mathbf{x}_{i,A^j \leftarrow a'}, a')\big| \geq \epsilon] \leq \delta$$

We can make the same convex relaxation to the indicator function here,

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

and notice that every possible choice of δ corresponds to a specific setting of λ. Specifically, if λ is increased it corresponds to decreasing δ. In both cases, when $\lambda = 0$ we recover the traditional supervised learning objective. 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.

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).

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 [18] . It contains information on over 20,000 students including their race A, 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

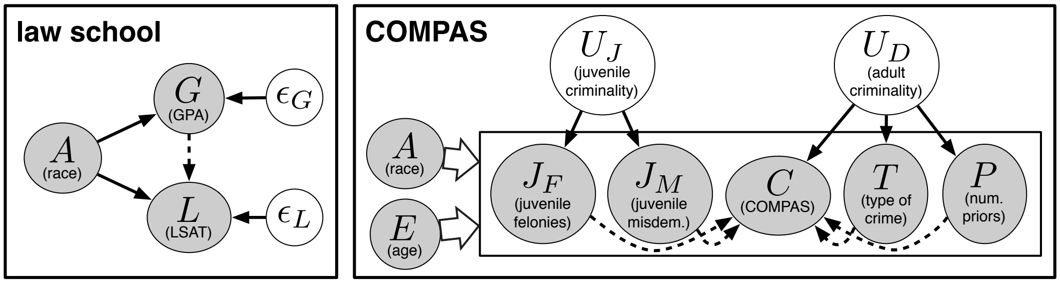

Figure 2: Causal models for the law school and COMPAS datasets. As before shaded nodes are observed an unshaded nodes are unobserved. For each dataset we consider two possible causal worlds: the first model is given by omitting dotted arrows and the second includes them. The law school model is a deterministic causal model with additive unobserved variables ϵ_G, ϵ_L. For COMPAS, the large white arrows signify that variables A, E are connected to every variable contained in the box they point to. The COMPAS model is non-deterministic and the equations for each model are given in eq. (6) and eq. (7).

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 graph in Figure 2 (*Left*). In this graph we also model the fact that G, L may have been affected by other unobserved quantities which we model with ϵ_G, ϵ_L. However, we may be uncertain whether G causes L, thus we indicate a possible connection between G and L. Thus we propose to model this dataset with two worlds, one that omits the connection between G and L and one that contains it. The corresponding equations for these two worlds are as follows:

$$G = b_G + w_G^A A + \epsilon_G \qquad G = b_G + w_G^A A + \epsilon_G \tag{6}$$

$$L = b_L + w_L^A A + \epsilon_L \qquad L = b_L + w_L^A A + w_L^G G + \epsilon_L \tag{7}$$

$$\epsilon_G, \epsilon_G \sim \mathcal{N}(0,1)$$

where $b_G, w_G^A, b_L, w_L^A, w_L^G$ are parameters of the causal model.

Results. Figure 3 shows the result of learning the MWF classifier on the deterministic law school models. We split the law school data in to a random 80/20 train/test split and we fit our causal model and classifier 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 for different fairness levels ϵ. For each ϵ, we selected the smallest λ such that the constraint in eq. (3) held across 99.9% of the inputs. We see that MWF is able to reliably sacrifice accuracy for fairness as ϵ is reduced.

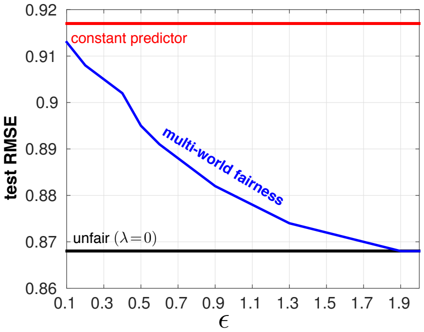

Figure 3: Test prediction results for different ϵ on the law school dataset.

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 [12] 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, their age E, their number of juvenile felonies J_F, juvenile misdemeanors J_M, the type of crime they committed T, and the number of prior offenses they have P. There is also a proprietary COMPAS score [12] C designed to indicate the likelihood a prisoner recidivates. We model this dataset with a

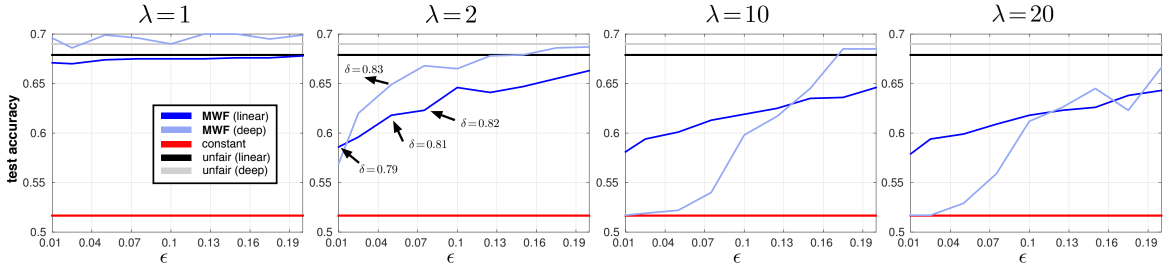

Figure 4: Test prediction results for different ϵ and λ on the COMPAS dataset.

non-deterministic causal model:

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

where the first causal model includes the blue text and the second does not.

Results. Figure 4 shows how classification accuracy using both logistic regression (linear) and a 3-layer neural network (deep) changes with both ϵ and λ (as a proxy for δ) changes. We note that for small λ MWD is able to match the accuracies of the unfair models. However as λ increases, indicating stricter fairness requirements, the accuracy is reduced.

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. Both the conceptual simplicity of our method in 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 drop-in and many worlds fair replacement for other existing classifiers.

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] 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

[6] 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

[7] 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

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

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

[10] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017. 2, 3, 5

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

[12] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 2016. 7

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

[14] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000. 2

[15] J. Pearl, M. Glymour, and N. Jewell. *Causal Inference in Statistics: a Primer*. Wiley, 2016. 3, 4

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

[17] 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

[18] Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998. 6

[19] 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