Journal of Machine Learning Research 21 (2020) 1-44 Submitted 11/19; Revised 9/20; Published 11/20

Fair Data Adaptation with Quantile Preservation

Drago Plečko DRAGO.PLECKO@STAT.MATH.ETHZ.CH
Nicolai Meinshausen MEINSHAUSEN@STAT.MATH.ETHZ.CH
Seminar für Statistik
ETH Zürich
Zürich, 8092, Switzerland

Editor: David Sontag

Abstract

Fairness of classification and regression has received much attention recently and various, partially non-compatible, criteria have been proposed. The fairness criteria can be enforced for a given classifier or, alternatively, the data can be adapted to ensure that every classifier trained on the data will adhere to desired fairness criteria. We present a practical data adaption method based on quantile preservation in causal structural equation models. The data adaptation is based on a presumed counterfactual model for the data. While the counterfactual model itself cannot be verified experimentally, we show that certain population notions of fairness are still guaranteed even if the counterfactual model is misspecified. The nature of the fulfilled observational non-causal fairness notion (such as demographic parity, separation or sufficiency) depends on the structure of the underlying causal model and the choice of resolving variables. We describe an implementation of the proposed data adaptation procedure based on Random Forests (Breiman, 2001) and demonstrate its practical use on simulated and real-world data.

Keywords: Supervised learning, Fairness in machine learning, Causality, Graphical models, Counterfactual fairness

1. Introduction

Care needs to be taken when applying machine learning techniques in socially sensitive domains, because algorithms are sometimes capable of learning societal biases we might not want them to learn. For example, women tend to be disadvantaged in credit score ratings, partially due to the fact that women are currently perceived to have lower income on average (Blau and Kahn, 2003). A credit scoring rating fair with respect to gender would be desirable. However, the precise notion of fairness one would like to achieve is often debatable. Various different notions exist and these notions can often be incompatible (Corbett-Davies and Goel, 2018).

Current approaches for building fair predictors broadly fall into three categories. Pre-processing methods focus on transforming the data in order to remove any unwanted bias (Zemel et al., 2013; Calmon et al., 2017). In-processing methods attempt to build in fairness constraints into the training step (Fish et al., 2016; Zafar et al., 2017; Donini et al., 2018). Post-processing methods focus on transforming an already constructed predictor (Hardt et al., 2016). Our work falls into the pre-processing category (although it could also be viewed as a post-processing step applied after learning the causal graph of the data).

In particular, our work

1. Provides a practical implementation of fair data adaption based on Random Forests and an underlying causal model which is assumed to be known. This allows to incorporate resolving variables (Kilbertus et al., 2017). The software is provided as an R package `fairadapt`.

2. Presumes a specific counterfactual model. We can show that a counterfactual notion of fairness is satisfied if the model is correct (unfortunately not verifiable), but that certain population fairness notions are satisfied in any case, even if the counterfactual model is wrong.

3. Combines the two existing notions of counterfactual fairness and resolving variables into a single fairness criterion that can depend on the causal graph. This might facilitate discussions about a suitable fairness notion, provided people can agree on the structure of the underlying causal graph (and a discussion about the appropriate structure might also be fruitful in itself).

We also demonstrate the empirical value of our approach.

1.1 Setup

Let random variable Y, taking values in $\mathcal{Y}$, be the outcome of interest that one would like to predict in the future in a fair way. We mostly assume binary classification so that $\mathcal{Y} = \{0, 1\}$, but we believe extending this approach to other settings is also possible, but requires some further work. The binary outcome Y represents perhaps recidivism whilst on parole or repayment of a loan. In this work, we are not considering the issues of *selective labels* (for instance not being able to observe who recidivates among people not given parole) or *wrong labels* (that is we assume there are no undiscovered cases of recidivism among people given parole). Let A be the protected attribute such as race or gender and $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ be predictor variables for the outcome of interest. We assume we have access to n i.i.d. samples (A_i, X_i, Y_i), $i = 1, \ldots, n$ coming from a distribution $F_{A,X,Y}$. For the majority of the exposition we assume that A has two levels $\{0, 1\}$. Most ingredients for extending our work to non-binary A are given in the discussion, but recent work suggests this generalization might not be straightforward (Kearns et al., 2017). The key goal is to provide a data-projection or data-adaptation

$$T : \mathbb{R}^p \times \mathcal{Y} \mapsto \mathbb{R}^p \times \mathcal{Y}.$$

The projection should be such that if we train a classifier with the adapted data

$$T\big((X_i, Y_i)\big),\ i = 1, \ldots, n$$

instead of the original data $\{(X_i, Y_i),\ i = 1, \ldots, n\}$, we want to be able to automatically guarantee appropriate fairness criteria. At the same time, we want the change induced by the data adaptation to be minimal in an appropriate sense.

1.2 Causal framework

We mainly use a standard non-parametric structural equation model (NPSEM) for $Z = (A, X, Y) \in \mathbb{R}^{p+2}$ as in Pearl (2000) and let each variable Z_k be defined as

$$Z_k = g_k(Z_{\mathrm{pa_k}}, U_k), \tag{1}$$

where $U \in \mathbb{R}^{p+2}$ is a latent variable that determines the realization of the variable Z_k and pa_k is the set of parents of the variable Z_k. We assume that the components of U are independent, that is we assume lack of confounding, also known as the Markovian assumption in (Pearl, 2009). By writing $Z(U = u)$ we refer to a specific instance of Z obtained by plugging in the latent quantiles $U = u$ into the system of equations in (1). We also use the potential outcomes notation and write $Z(A = a)$ for the potential outcome under a $\mathrm{do}(A = a)$ intervention[1]. Further, we denote by $f_k(z_k \mid z_{\mathrm{pa}_k})$ the density corresponding to Z_k. If the random variables Z_k are continuous and a density exists, without limitation of generality, we can assume that

(C1) for each fixed value of $Z_{\mathrm{pa_k}} = z$ the function

$$g_k(z, u)$$

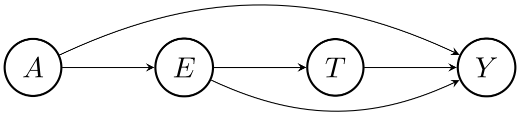
is strictly increasing in u,

(C2) each U_k follows a uniform $U[0, 1]$ distribution.

These assumptions are discussed further in Appendix J. Random variables U_k can be interpreted as the quantile of the k-th variable, conditional on the value of its parents $Z_{\mathrm{pa_k}}$. There exists a one-to-one mapping between the value of z and u. The important case of discrete random variables, where the deterministic relationship between z and u breaks down, is addressed in Section 5. Throughout the paper, we use the standard counterfactual assumptions found in (Pearl, 2009, Chapter 7.3).

1.3 Fair Twins

Consider the following example. Variable A is the protected attribute, in this case race ($A = 0$ corresponding to females, $A = 1$ to males). Let E be educational achievement (measured for example by grades achieved in school) and T the result of an admissions test for further education. Let Y be the outcome of interest (final score) upon which admission to further education is decided. Edges in the graph indicate how variables affect each other.



The main problem is that the attribute A, gender, has a causal effect on variables E, T and Y. We want to find a data projection that makes the data 'look' the same for males

[1] $Z(A = a, U = u)$ denotes a specific realization of the potential outcome variable under a $\mathrm{do}(A = a)$ intervention and latent variables $U = u$.

and females, in the sense that the conditional distribution of (E, T, Y) is identical for both levels of A. We should emphasize that this goal is just one of many possible fairness criteria one would like to achieve and we discuss different options later. We start from an

$$\text{observational distribution: } (A, E, T, Y) \sim F^{(obs)}.$$

For each individual with observed values (a, e, t, y) from the observational distribution $F^{(obs)}$, we want to transform the data and compute his/her 'fair twin' as

$$(a^{(fp)}, e^{(fp)}, t^{(fp)}, y^{(fp)}) = \text{FP}(a, e, t, y).$$

If we sample $(A, E, T, Y) \sim F^{(obs)}$, then the transformed data $\text{FP}(A, E, T, Y)$ will follow the so-called fair-projection distribution $F^{(fp)}$. The simple idea for the fair-projection is that we want all individuals to have the same protected attribute after the fair twin projection (in the example the baseline is $A^{(fp)} = 0$; all females) and the distribution of the projected $(E^{(fp)}, T^{(fp)}, Y^{(fp)})$ should match the conditional distribution of

$$(E, T, Y) \mid A = 0$$

in the observational distribution.

Subject to this, we also want to minimize the distortion in the data coming from the projection and preserve the relative achievements as much as possible. More explicitly, for a male person with education value e, we give it the transformed value $e^{(fp)}$ chosen such that

$$\mathbb{P}^{(obs)}(E \geq e \mid A = 1) = \mathbb{P}^{(obs)}(E \geq e^{(fp)} \mid A = 0).$$

The main idea is that the *relative educational achievement within the subgroup* would stay the same if we changed someone's gender. If you are male and you have a higher educational achievement than 60% of all males in the dataset, we assume you would be better than 60% of females had you been female. After computing everyone's education (in the 'female' world), we continue by computing the transformed test score values $T^{(fp)}$. The approach is again similar, but this time we condition on educational achievement. That is, a male with values $(E, T) = (e, t)$ is assigned a test score $t^{(fp)}$ such that

$$\mathbb{P}^{(obs)}(T \geq t \mid E = e) = \mathbb{P}^{(obs)}(T \geq t^{(fp)} \mid E = e^{(fp)}),$$

where the value $e^{(fp)}$ was obtained in the previous step. This way of counterfactual correction is known as *recursive substitution* (Pearl, 2009, Chapter 7). In the last step, the outcome variable Y needs to be adjusted. The adaptation is based on the values of gender, education and the test score. The transformed value $y^{(fp)}$ of $Y = y$ would satisfy

$$\mathbb{P}^{(obs)}(Y \geq y \mid E = e, T = t, A = 1) = \mathbb{P}^{(obs)}(Y \geq y^{(fp)} \mid E = e^{(fp)}, T = t^{(fp)}, A = 0). \quad (2)$$

The distribution of the transformed data is identical to the distribution under the $do(A = 0)$ intervention in the context of the structural model. We have induced a coupling between the distributions $F^{(obs)}$ and $F^{(fp)}$, since we keep the quantiles U identical for the fair-twin version of an individual. If one were not interested in the coupling, one could use the formal framework of single-world intervention graphs (SWIGs) introduced in Richardson and Robins (2013).

However, it could be argued, in the example above, that the test score T should not be adapted. In that case, we would call T a resolving variable. If so, the transformed test score value would simply equal the observed one, that is $t^{(fp)} = t$. Furthermore, the transformed value of the label y would also be different as a result. The adapted value $y^{(fp)}$ would in this case satisfy

$$\mathbb{P}^{(obs)}(Y \geq y \mid E = e, T = t, A = 1) = \mathbb{P}^{(obs)}(Y \geq y^{(fp)} \mid E = e^{(fp)}, T = t, A = 0), \qquad (3)$$

instead of Equation (2). The difference is in the conditioning on $T = t$ on the RHS of the equation, as we now keep the test score fixed.

The fair-twin projection can be easily achieved in the context of a causal structural model by a do-intervention. We set A to its baseline value and reuse the same latent quantiles as for the original data. Then, under the random value of A we obtain a draw from the empirical distribution $F^{(obs)}$, whereas by setting $do(A = 0)$ and keeping the same quantiles we obtain the fair twin of the same individual.

Definition 1 (Quantile preservation assumption (QPA)) *The conditional quantiles U_k defined by the SCM in Equation* (1) *remain unchanged under the fair-twin projection for all variables that are neither the protected attribute nor a resolver. Let $(z_k, z_{\mathrm{pa_k}})$ be the realization of $(Z_k, Z_{\mathrm{pa_k}})$. Then the realization $(z_k^{(fp)}, z_{\mathrm{pa_k}}^{(fp)})$ satisfies*

$$\mathbb{P}^{(obs)}(Z_k \leq z_k^{(fp)} \mid Z_{\mathrm{pa_k}} = z_{\mathrm{pa_k}}^{(fp)}) = \mathbb{P}^{(obs)}(Z_k \leq z_k \mid Z_{\mathrm{pa_k}} = z_{\mathrm{pa_k}}).$$

After rewriting the SCM in Equation (1) so that conditions (C1)-(C2) hold, the assumption in Definition 1 simply states that the realizations of latent quantiles $U = u$ remain unchanged under the fair-twin projection[2].

The defined counterfactual cross-world model (an assumption about the joint distribution of random variables under different interventions) is not empirically verifiable. We will try to make clear where we use single-world and where we use cross-world assumptions throughout the text. The QPA will, for example, be appealing in situations where the latent noise variables measure individual effort that is not explainable by the observed variables and there is reason to believe that the relative strength of the individual effort or achievement is something immutable for this individual. By equalizing the distributions for subgroups defined by the protected attribute A, all subgroups are treated in the same way, in the sense that any decision we make based on the transformed data will have the same distribution across all groups. The situation is somewhat more involved in the presence of resolving variables, though.

Another assumption we make is that the protected attribute A is a root node of the causal graph $\mathcal{G}$. A consequence of this is that the $do(A = a)$ intervention is equivalent to conditioning on $A = a$. This idea, which is easily shown using the 2nd rule of do-calculus (Pearl, 2009), shows up frequently in our discussion. This assumption poses a limitation, since non-root sensitive attributes (like socio-economic status or education) are not in the

[2] An alternative view would be that the fair-twin projection is obtained under a do-intervention of the form $do(A = 0, U = u^{(obs)})$, where A is set to its baseline value and the realization of the quantiles U is kept fixed (note that the latent quantiles can be determined from the observed $Z^{(obs)} = z^{(obs)}$).

scope of this paper. We believe that extending our approach to non-root A is a technical challenge which we leave open for future work.

We find the lack of hidden confounding assumption to be the biggest limitation. It is not hard to construct examples in which, with presence of confounding, the adaptation procedure we propose does not work, that is it fails to eliminate discrimination. However, if A is a root node, the effect of the $do(A = a)$ intervention is still identifiable under quite general assumptions (Tian and Pearl, 2002; Shpitser and Pearl, 2008). Therefore, there is hope for solving the problem in general, but this is beyond the scope of this manuscript.

After the adaptation procedure, the projected data can be used to construct a classifier and we will discuss which fairness conditions are guaranteed with such an approach. Generally, the projection we are describing is carried out using *Quantile Regression Forests* (Meinshausen, 2006). A full implementation of this method for a general situation is available in the `fairadapt` package on CRAN. The aim of this paper is to formalize all of the ideas above mathematically.

Note that the fair projection can be used for *fair-twin inspection* - for every individual, we can compute his/hers “fair-twin” version, corresponding to the same person but with the baseline value of the protected attribute A. Fair-twin inspection can be used to help justify fair decisions on an individual level.

1.4 Structure of the paper

In Section 2 causal notions of *resolving variables* and *counterfactual fairness* are discussed. In Section 3 the adaptation procedure is introduced and we formalize its aim. Notions of resolved fairness are discussed, and in particular how they depend (or do not depend) on the assumptions used. The Section also describes the population level adaptation procedure (under no estimation error). The quantile preservation assumption that is used is also briefly discussed. We further illustrate how our desired notion of resolved fairness amounts to a single linear constraint in the simplest linear additive setting. In Section 4 the relation of our work to previously proposed methods and criteria is analyzed. Section 5 goes in depth about discussing the practical aspects of our method. Most importantly, the problem handling of discrete variables in the adaptation procedure is given much attention. After applying the data projection, there are several reasonable options for the training step and these are discussed. Two possible methodological extensions are discussed in the end of the section. In Section 6 empirical performance of our method is demonstrated both on simulated and real-world data (namely the COMPAS and Adult datasets).

2. Causal notions of fairness

We look at two counterfactual notions of fairness that play an important role in our methodology.

2.1 Counterfactual fairness

Counterfactual fairness was first introduced as a notion by Kusner et al. (2017).

Definition 2 (Counterfactual fairness, Kusner et al. (2017)) *A predictor $\widehat{Y}$ is counterfactually fair if*

$$\widehat{Y}(A = a) \mid A = a, X = x \quad \overset{d}{=} \quad \widehat{Y}(A = a') \mid A = a, X = x \;\; \forall a, a', x, \tag{4}$$

where $\widehat{Y}(a = a)$ is the potential outcome of $\widehat{Y}$ when setting A to value a.

Here and throughout we think of $A = 0$ as the *baseline value* of the protected attribute. The idea behind this notion is that if we intervene to change someone's race or gender, this should not affect the prediction they obtain.
A weaker form of counterfactual fairness would just require that the distribution of $\widehat{Y}$ under an intervention on the protected attribute remains unchanged, that is

Definition 3 (Population fairness) *A predictor $\widehat{Y}$ is said to satisfy population fairness if*

$$\widehat{Y}(A = a) \overset{d}{=} \widehat{Y} \quad \forall a.$$

The distributional equivalence from Definition 3 does not rest on cross-world assumptions and is equal to the observational criterion of demographic parity, in the case when A is a root node in the causal graph (shown later in Proposition 9). In contrast, a much stronger notion can be defined as follows

Definition 4 (Strong counterfactual fairness) *A predictor $\widehat{Y}$ is said to satisfy strong counterfactual fairness if*

$$\widehat{Y}(A = a, U = u) = \widehat{Y}(U = u) \;\; \forall a, u \; .$$

Definition 4 requires the counterfactual prediction to be identical when setting the protected attribute to any value. Note that this is an individual level fairness notion.

2.2 Resolving variables

Kilbertus et al. (2017) discuss that in some situations the protected attribute A can affect variables in a non-discriminatory way. For instance, in the Berkeley admissions dataset (Bickel et al., 1975) we observe that females often apply for departments with lower admission rates and consequently have a lower admission probability. However, we perhaps would not wish to account for this difference in the adaptation procedure if we were to argue that department choice is a choice everybody is free to make. This motivated the following definition:

Definition 5 (Resolving variables, Kilbertus et al. (2017)) *Let $\mathcal{G}$ be the causal graph of the data generating mechanism. Let the descendants of variable A be denoted by* $\mathrm{de}(A)$*. A variable R is called resolving if*

(i) $R \in \mathrm{de}(A)$

(ii) the causal effect of A on R is considered to be non-discriminatory

The idea is that the value of a resolving variable, or a resolver, R should not change under our adaptation procedure. More generally, we can consider a set of resolving variables R. The desired counterfactual fairness criteria with respect to this definition can now be stated as

$$\text{Population resolved: } \widehat{Y}(A=a, R=R^{(obs)}) \overset{d}{=} \widehat{Y}(A=a', R=R^{(obs)}) \ \forall a, a' \tag{5}$$

$$\text{Strong resolved: } \widehat{Y}(A=a, R=r^{(obs)}, U=u^{(obs)}) = \widehat{Y}(A=a', R=r^{(obs)}, U=u^{(obs)}) \ \forall a, a', u^{(obs)} \tag{6}$$

where $r^{(obs)}$ is the value naturally attained by R under quantiles $U = u^{(obs)}$.

In the presence of resolving variables, we additionally keep the resolving variables to the value they have attained naturally, by writing $do(R = R^{(obs)})$ (or $do(R = r^{(obs)})$ if we refer to a specific individual with a realized value $r^{(obs)}$). The strong notion requires that the counterfactual predictions remain unchanged under a do-intervention on the protected attribute. Note that by using resolving variables we allow for some additional flexibility in the exact fairness criterion. This notion, however, is not flexible enough to treat *path-specific effects*. More about this is discussed in Section 5.5.

It is not immediately clear which variables should be considered as resolving. It can even happen that the same variable can be resolving or non-resolving in different applications. For instance, when recruiting students for an athletics training programme, we perhaps do not wish to give males an advantage based on physical ability. In this case, physical strength is not a resolving variable. However, if we are hiring workers for a physical job, we might want to consider physical strength as a resolving variable.

3. Adaptation

The main goal of this paper is to combine the two causal notions given in Sections 2.1 and 2.2 to describe a preprocessing procedure which gives a fair representation of the data. After this, any method can be used to construct a fair classifier $\widehat{Y}$. A more detailed discussion of the training step is given in Section 5.4.

Adaptation aim. We want to find a projection $\text{FP}(Z) = \text{FP}(A, X, Y)$ such that using the projected data automatically guarantees certain fairness notions. In absence of resolving variables, we discussed already that for an individual z drawn from the observational distribution $F^{(obs)}$, we define $z^{(fp)} = \text{FP}(z)$ as the value obtained under the causal structural model if setting A to its baseline value 0 and keeping the values of the latent quantiles $U = u^{(obs)}$ constant.

Suppose now that the features are decomposed into non-resolvers N and resolvers R, $X = (N,\ R)$. Let $FP_S(S)$ be the transformed values of a subset S. With slight abuse of notation, we drop the subscript S in the notation and write only $\text{FP}(S)$. The data projection works as follows. For the protected attribute, we have that $\text{FP}(a) = 0$, since we set it to the baseline value $A = 0$. The resolving variables R are unaffected by the projection, that is $\text{FP}(r) = r$. We also evaluate the quantiles u and keep them constant. The non-resolvers N and target Y are then obtained by evaluating the SCM while setting (A, R, U) equal to (a, r, u). That means we: (i) keep the latent quantile variables U identical[3]; (ii) set the

[3] A different view on this would be that the latent quantiles U are in fact resolving variables.

protected attribute A to its baseline value; (iii) keep the value of the resolvers equal to their naturally attained values R under no intervention. Note that under this construction, the fair projection satisfies

$$\mathrm{FP}(\mathrm{FP}(z)) = \mathrm{FP}(z),$$

that is the projection is idempotent. The idempotent property guarantees the strong counterfactual notion of fairness (6). If $\widehat{Y} = g \circ \mathrm{FP}(a, x)$ for any real-valued function g, we have that $\widehat{Y}(a, x) = \widehat{Y}(\mathrm{FP}(a, x))$ for all (a, x). We call this a counterfactual property because it compares an individual z with its counterfactual fair-projected value $\mathrm{FP}(z)$. Even if the counterfactual model from Definition 1 is misspecified, we still achieve the population fairness notion. We can hence summarize the goal as in the following table.

	population fairness	strong counterfactual fairness
counterfactual model true	✔	✔
counterfactual model false	✔	✘

An interesting special case is when there are no resolvers, $R = \emptyset$. In this case $\mathrm{FP}(X) \perp\!\!\!\perp A$ and $\widehat{Y} = g \circ \mathrm{FP}(A, X)$ guarantees that *demographic parity* (also known as *separation*) holds, $\widehat{Y} \perp\!\!\!\perp A$.

The above formulation requires that the quantiles U can be identified from the observed data. This is possible in the case of continuous random variables that permit a density. We get back to the need for randomization in the case of discrete random variables.

3.1 Population level adaptation

The input of our procedure is the causal graph $\mathcal{G}$, a choice of resolving variables R and data $(A_k, X_k, Y_k)_{k=1:n} = (A, X, Y)$. Even though we are describing the procedure on population level (meaning we are ignoring finite sample estimation errors) we still work with data samples to emphasize our counterfactual construction. We also assume that the densities $f_k(x_k \mid \mathrm{pa}(x_k))$ of the SCM in Equation (1) are known. The output of our procedure is the projected data $\mathrm{FP}(X, Y)$. In Section 5.3 we explain how the procedure is carried out non-parametrically on sample level. We note that Algorithms 1 and 2 (given later) produce the same result for any valid topological ordering, due to the Markov property of directed acyclic graphs (DAGs). We show that the procedure in Algorithm 1 satisfies the desired fairness criteria in the following theorem:

Theorem 6 (Population and strong resolved fairness) *Let $FP(\cdot)$ be the projection from Algorithm 1. Suppose f is any classifier built based on the transformed data $FP(X, Y)$. Then we have that the classifier $\widehat{Y} = f \circ FP(\cdot)$ satisfies population resolved fairness, that is*

$$\widehat{Y}(A = a, R = R^{(obs)}) \quad \overset{d}{=} \quad \widehat{Y}(A = a', R = R^{(obs)}) \quad \forall a, a'.$$

Algorithm 1: Population Fairness Adaptation

```
Input: causal graph $\mathcal{G}$, density of the data generating mechanism
       $f(x_1, ..., x_k) = \prod f(x_i \mid \text{pa}(x_i))$, choice of resolving variables $R$, data
       $(A_k, X_k, Y_k)_{k=1:n} = (A, X, Y)$
Output: adapted data $\text{FP}(X, Y)$
1 $\text{FP}(A_k) \leftarrow 0$ for each $k$
2 $\text{FP}(R_k) \leftarrow R_k$ for each $k$
3 for $V \in \text{de}(A) \setminus R$ in topological order do
4     using the density $f(v \mid \text{pa}(v))$ obtain the inverse quantile function $g_V$ of $V$, such
        that $V = g_V(U^{(V)}, \text{pa}(V))$
5     obtain the latent quantile $U_k^{(V)}$ of $V_k$ for each $k$
6     obtain the transformed value $\text{FP}(V_k)$ using the transformed values of its parents
        (obtained in previous steps), by setting $\text{FP}(V_k) \leftarrow g_V(U_k^{(V)},\ \text{FP}(\text{pa}(V_k)))$
7 return $FP(X, Y)$
```

Further, under the quantile preservation assumption (QPA) $\widehat{Y}$ satisfies strong resolved fairness,

$$\widehat{Y}(A = a, R = r^{(obs)}, U = u^{(obs)}) \quad = \quad \widehat{Y}(A = a', R = r^{(obs)}, U = u^{(obs)}) \quad \forall a, a', u^{(obs)}.$$

The main idea of the proof is to show that the $\text{FP}(\cdot)$ projection is equivalent to the $do(A = 0, R = R^{(obs)})$ intervention and to use the counterfactual construction to show that the stronger fairness notion is satisfied as well. The full proof is given in Appendix A. Note that our procedure also outputs the adapted labels $\text{FP}(Y)$ that are to be used in the training step. The reader might wonder if original labels Y could also be used. This is discussed in Appendix B.

Obtaining the (inverse) quantile function and latent quantiles. In line 4 of Algorithm 1, we obtain the inverse quantile function g_V of V. More precisely, we first obtain the quantile function $Q_V(V; \text{pa}(V))$ using quantile regression forests (Meinshausen, 2006), after which g_V is obtained by inverting $Q_V(V; \text{pa}(V))$. The latent quantiles in line 5 are also obtained using quantile regression forests. Even though tree ensemble methods might perform worse in presence of heteroscedastic noise, we focus on this option because of its computational tractability. Alternative options for the quantile step are linear methods (Koenker, 2013) and neural network approaches (Cannon, 2018). All of the mentioned methods are available within the `fairadapt` package. An empirical comparison of these methods can be found in Appendix D.

Discussion of the assumptions. The quantile preservation assumption (QPA) from Definition 1 is used for constructing a joint cross-world distribution. The assumption is equivalent to the equal noise assumption used in the NPSEM-IE framework of Pearl (2009) in cases where the noise is additive in the structural equations. This assumption has been much debated in the causal community in its different forms. See, for instance, (Dawid,

2000) or (Pearl, 2000). The assumption is indeed not testable, not even in principle. We offer some thoughts on why QPA might be sensible in the fairness context:

(a) If we consider two continuous distributions with cumulative functions F_{X_0} and F_{X_1} and $p \geq 1$, the Wasserstein distance $W_p(F_{X_0}, F_{X_1})$ is minimized[4] by the optimal transport map $F_{X_1}^{-1} \circ F_{X_0}$, as shown by Cuesta-Albertos et al. (1993). This mapping precisely represents quantile matching. We can see how QPA arises naturally as minimizing the distance between counterfactual worlds.

(b) The quantile preservation assumption ensures that we retain the original ordering of the values. Namely, if for a variable V two individuals have equal values for all $\text{an}(V)$, then QPA guarantees their counterfactual values $V(A = 0)$ will retain the original ordering.

(c) In mathematical modeling, we often use noise to describe variations which are not explained by the data. This does not necessarily mean these are completely random, but could be a result of certain unobserved variables, which possibly cannot even be measured in practice. For instance, in the hypothetical intervention "What would have happened had this female been born a male?", there are a number of genetic, parenting and societal factors that would have remained the same.

For these reasons, it seems that if we consider individuals with high quantiles U (very successful individuals), it would be hard to argue how it could be fair that these quantiles do not stay the same in the counterfactual world, as this would change the relative ranking within a group (where the group is defined by levels of the protected attribute A).

Linear additive case. The following theorem is intended to provide intuition about what the resolved fairness condition ensures in the simplest linear additive case.

Theorem 7 (Strong resolved fairness for linear additive SEMs) *Assume that we have an additive, linear structural equation model for variables $(A, X_1, ..., X_k, Y)$ and that $A \in \{0, 1\}$ is a root node in the causal graph*

$$X^{(i)} \leftarrow \sum_{V \in \text{pa}_i} \beta_V^{(i)} V + n_i(U_i),$$

$$Y \leftarrow \sum_{V \in \text{pa}(Y)} \beta_V^Y V + n_Y(U_Y).$$

where n_i, n_Y are any monotonic functions. The noise terms $n_i(U_i)$ are assumed to be independent and U are the latent quantiles. Let R be the set of the resolving variables. If $\widehat{Y} = \alpha_A A + \sum_{i=1}^{k} \alpha_i X^{(i)}$ is a linear predictor for Y then the strong resolved fairness condition (6) *implies that*

$$\sum_{i=1}^{k} \alpha_i \times \left(\sum_{\substack{\textit{paths } A \to X_j \\ \textit{disjoint from } R}} \prod_{m \in \textit{ path}} \beta_m \right) + \alpha_A = 0. \tag{7}$$

[4] We note that this minimization is achieved uniquely in the case where $p > 1$, since the cost function is strictly convex.

Proof In the proof we suppress the notation $U = u$ to indicate that the quantiles are unchanged under the $do(A = a, R = R^{(obs)})$ intervention. We can expand

$$X^{(i)}(A = 1, R = R^{(obs)}) - X^{(i)}(A = 0, R = R^{(obs)}) =$$
$$\sum_{V \in \text{pa}_i} \beta_V^{(i)} (V(A = 1, R = R^{(obs)}) - V(A = 0, R = R^{(obs)}))$$

By recursively expanding the sum on the RHS we obtain that

$$X^{(i)}(A = 1, R = R^{(obs)}) - X^{(i)}(A = 0, R = R^{(obs)}) = \sum_{\substack{\text{paths } A \to X_j \\ \text{disjoint from } R}} \prod_{k \in \text{ path}} \beta_k. \tag{8}$$

Finally, we know that

$$\begin{aligned}
&\widehat{Y}(A = 1, R = R^{(obs)}) - \widehat{Y}(A = 0, R = R^{(obs)}) = \\
&\alpha_A + \sum_{i=1}^{k} \alpha_i \Big[X^{(i)}(A = 1, R = R^{(obs)}) - X^{(i)}(A = 0, R = R^{(obs)}) \Big] = \\
&\alpha_A + \sum_{i=1}^{k} \alpha_i \times \Bigg(\sum_{\substack{\text{paths } A \to X_j \\ \text{disjoint from } R}} \prod_{m \in \text{ path}} \beta_m \Bigg)
\end{aligned}$$

which should be equal to 0 by condition (6). Therefore, the constraint (7) holds. ■

Notice that resolved fairness in the linear additive case amounts to a single linear constraint on the coefficients of $\widehat{Y}$.

Lastly, we summarize the main advantages of our method:

(i) it does not throw away information contained in de(A) which is potentially useful for prediction, as proposed in some previous works (Kusner et al., 2017),

(ii) it takes the causal perspective into account and therefore ensures that fairness criteria are not satisfied spuriously,

(iii) it allows us to relax demographic parity (achieved when $R = \emptyset$) by introducing resolving variables, since demographic parity could be a prohibitively strong notion in certain applications. In Section 6 we discuss how enlarging the set of resolving variables improves the calibration of the constructed predictor.

4. Relation to existing work

In this section we discuss the relation of fair adaptation to previous work on fairness.

4.1 Observational notions of fairness

For sake of brevity, we do not mention all the definitions of fairness proposed so far. We only review the most important observational notions. By *observational notions* we refer to all notions that only focus on the observational distribution of the data, without taking the generating causal mechanism into account.

(i) One of the first observational notions, called *demographic parity*, goes all the way back to Darlington (1971).

Definition 8 (Demographic parity) *A predictor $\widehat{Y}$ satisfies demographic parity if*

$$\widehat{Y} \perp\!\!\!\perp A. \tag{9}$$

In the special context of binary predicted labels, $\widehat{Y} \in \{0, 1\}$, demographic parity is equivalent to $\mathbb{P}(\widehat{Y} = 1 \mid A = 0) = \mathbb{P}(\widehat{Y} = 1 \mid A = 1)$. In words, this definition requires that our prediction is independent of the protected attribute. We now show that our population fairness criterion (5) is equivalent to demographic parity in the case when A is a root node in the causal graph:

Proposition 9 *Suppose that the protected attribute A is a root node in the causal graph $\mathcal{G}$. If $\widehat{Y}$ is a binary predictor for the outcome Y, then we have that*

$$\widehat{Y} \perp\!\!\!\perp A \quad \Longleftrightarrow \quad \widehat{Y}(A = a) \overset{d}{=} \widehat{Y} \ \forall a. \tag{10}$$

In words, if A is a root node, then population fairness is equivalent to demographic parity.

Proof By applying the Action/Observation exchange rule (2nd rule of do-calculus), found in (Pearl, 2009) we have

$$\widehat{Y}(A = a) \quad \overset{d}{=} \quad \widehat{Y} \mid A = a$$

Therefore, if $\widehat{Y}(A = a) \overset{d}{=} \widehat{Y} \ \forall a$, then for any a, a',

$$\widehat{Y} \mid A = a' \quad \overset{d}{=} \quad \widehat{Y}(A = a') \quad \overset{d}{=} \quad \widehat{Y}(A = a) \quad \overset{d}{=} \quad \widehat{Y} \mid A = a,$$

implying demographic parity. The reverse implication works analogously. ■

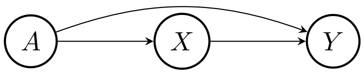
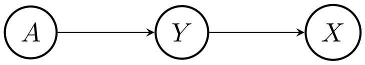
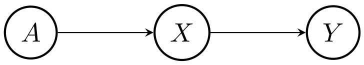
(ii) Another population definition of fairness is *equality of odds*, first proposed by Hardt et al. (2016).

Definition 10 (Equality of odds) *A predictor $\widehat{Y}$ satisfies equality of odds if*

$$\widehat{Y} \perp\!\!\!\perp A \mid Y.$$

For binary response Y and prediction $\widehat{Y}$ (the original context in which it was proposed), equality of odds is equivalent to $\mathbb{P}(\widehat{Y} = 1 \mid Y = y, \ A = 0) = \mathbb{P}(\widehat{Y} = 1 \mid Y = y, \ A = 1)$ for $y \in \{0, 1\}$. Only taking the equality above for $y = 1$ gives equality of opportunity. In words, this definition requires our prediction to be independent of the protected attribute, given the true outcome.

Table 1: Examples that describe the intrinsic relation of observational criteria to counterfactual fairness.

Example	Causal graph	Observational criterion achieved
(a)	 X not resolving	$\widehat{Y} \perp\!\!\!\perp A$
(b)	 Y resolving	$\widehat{Y} \perp\!\!\!\perp A \mid Y$
(c)	 X resolving	$Y \perp\!\!\!\perp A \mid \widehat{Y}$

(iii) The last observational notion we wish to mention is *calibration*, originally discussed by Chouldechova (2017) and Pleiss et al. (2017). Here, it is assumed that $\widehat{Y}$ is an estimator of the true conditional probability of $Y = 1$. Calibration is defined as follows.

Definition 11 (Calibration) *A prediction $\widehat{Y}$ satisfies calibration if*

$$Y \perp\!\!\!\perp A \mid \widehat{Y}.$$

For binary outcomes calibration is equivalent to $\mathbb{P}(Y = 1 \mid \widehat{Y} = y,\ A = 0) = \mathbb{P}(Y = 1 \mid \widehat{Y} = y,\ A = 1)$ for $y \in [0, 1]$. Calibration states that, given our prediction, the protected attribute should not provide us with additional information about the true outcome.

4.2 Adaptation and observational criteria

We discuss the relation between the observational criteria and our adaptation method. Consider the examples given in Table 1, in which we consider a classifier $\widehat{Y}$ to be a function of the adapted data $\text{FP}(X)$. For understanding the examples, it suffices to think of a non-resolving variable as adapted to contain no effect of A. In the table we discuss the possibility of Y being a resolving variable, which might seem confusing. By Y being resolving we simply mean that the true outcome is considered fair as it is. In the given examples we consider X to be a single feature, although the conclusions remain the same for multiple features. We first provide a formal statement about Table 1.

Theorem 12 (Fairness criteria) *Assume that for examples (a) and (b) from Table 1 we are building a classifier $\widehat{Y}$ based on appropriately adapted data $FP(X, Y)$ which satisfies the condition (5) for the choice of resolvers R given in the table. In example (c) we are building a predictor of the positive outcome probability $S(x) = \mathbb{P}(Y = 1 \mid X = x)$. For the given examples, in the population level case (infinite samples), we have the following:*

(a) for $\widehat{Y}$ built based on FP(X, Y) we have $\widehat{Y} \perp\!\!\!\perp A$

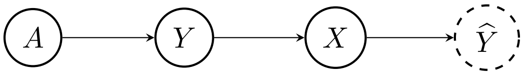
(b) for $\widehat{Y}$ built based on FP$(X, Y) = (X, Y)$ we have $\widehat{Y} \perp\!\!\!\perp A \mid Y$

(c) for $\widehat{Y} = \mathbb{E}[Y \mid X = x]$ built based on FP$(X, Y) = (X, Y)$ we have that $Y \perp\!\!\!\perp A \mid \widehat{Y}$.

Proof Consider the following:

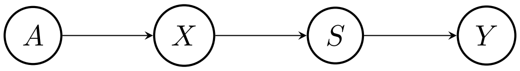
(a) The values of X and Y are transformed to $\mathrm{FP}(X, Y)$ which are independent of A, by condition (5). Any predictor $\widehat{Y}$ which is a function of $\mathrm{FP}(X)$ must also be independent of A.

(b) Consider the following graph where the predictor $\widehat{Y}$ is included in the causal graph



where the parent set $\mathrm{pa}(\widehat{Y}) = \{X\}$ indicates that only X is used for the predictor (in particular, A is not fed into the predictor). In this case, we have that Y d-separates A and $\widehat{Y}$ and the conclusion follows.

(c) In this example we can view the causal representation to be expanded as follows:



where $S(x) = \mathbb{P}(Y = 1 \mid X = x)$ is the true probability of positive outcome. For the infinite sample case, we have that $\widehat{Y} = S$, which implies that $\widehat{Y}$ d-separates A and Y.

■

We can now clarify the core ideas of observational notions discussed in this section. Note that in the toy examples from Table 1 we have that:

(a) Demographic parity is achieved when X or Y are considered to be non-resolving. We can see that in some sense demographic parity is a criterion that requires us to treat all subpopulations as exactly the same, regardless of what is observed in the data.

(b) Equality of odds is achieved when Y is considered to be resolving. In this case the adaptation procedure does not change the values of X, Y. The idea that the true outcomes Y are fair is in the heart of this notion.

(c) Calibration can be[5] achieved when X is considered to be resolving. In this case our adaptation procedure does not change the values of X, Y. Calibration is a criterion that ensures we do not discriminate any subpopulation beyond the differences observed in the data. Calibration should often come as a result of a good unconstrained predictor optimized for accuracy.

[5] It might be valuable to note that calibration does not necessarily have to arise in this case. The criterion is still dependent on how we build our classifier.

For examples (b) and (c), the fair projection FP is the identity. However, the examples illustrate that starting from the resolved fairness criterion (5) and the causal graph, all three observational notions can arise in these simple cases.

4.3 Mediation and path-specific methods

Fairness methods based on mediation analysis (Zhang and Bareinboim, 2018a,b) build on some ideas that are related to our work. However, the two most similar methods to ours are the path specific counterfactual methods (Nabi and Shpitser, 2018; Chiappa, 2019), which we now discuss.

Chiappa (2019) aims to eliminate the mean effect of the protected attribute along the causal pathways which are considered as unfair. We write

$$Y(A = a, X(A \overset{\text{f.p.}}{=} a'))$$

for the potential outcome variable obtained by setting the protected attribute to a along the unfair pathways and setting it to a' along the fair ones (for a precise clarification we refer the reader to the original paper). The path-specific effect is then defined as

$$PSE = \mathbb{E}[Y(A = a, X(A \overset{\text{f.p.}}{=} a'))] - \mathbb{E}[Y(A = a')]$$

and should be removed, as it is considered unfair. This is achieved by introducing a variational autoencoder for each variable whose counterfactual value we wish to compute. We refer to this approach as PSCF. The two main advantages of `fairadapt` compared to PSCF are computational speed and stability, together with absence of tuning parameters. The main advantage of PSCF, however, is the ability to remove the dependence of the encoder's latent projection and the protected attribute A, which can come as a result of hidden confounding. However, this flexibility comes at the expense of an additional term in the autoencoder objective which has an associated tuning parameter. There is unfortunately no canonical way for choosing this parameter and the authors emphasize that the additional term can cause instability and loss of information. Additionally, the PSCF method is slightly more flexible, since it can remove path-specific effects. We provide an implementation of the PSCF method and make an empirical comparison to `fairadapt` in Appendix I.

Further, Nabi and Shpitser (2018) start with the joint distribution $p(X, Y)$ and define discrimination as $\phi(p(X, Y))$, where ϕ is some functional of the distribution. After that, their goal is to find another distribution $p^\star(X, Y)$ which is close to the original $p(X, Y)$ and satisfies $|\phi(p^\star(X, Y))| \leq \epsilon$. One possible choice of ϕ they work with is the natural direct effect (NDE) which is defined as

$$NDE = E\big[Y(A = a, R = R(a')) - Y(A = a')\big].$$

The NDE can be interpreted as the total causal effect of the protected attribute on the outcome that does not go through resolving variables. If we use the transformed distribution $\text{FP}(X, Y)$ as the $p^\star(X, Y)$, we can relate their approach to our method as follows. If the condition (5) holds, that is if

$$\widehat{Y}(A = a, R = R^{(obs)}) \quad \overset{d}{=} \quad \widehat{Y}(A = a', R = R^{(obs)}) \;\; \forall a, a',$$

then it follows that

$$E\big[\widehat{Y}(A = a, R = R(a')) - \widehat{Y}(A = a')\big] = 0.$$

The short proof of this fact is omitted. We conclude that condition (5), which our method achieves, is sufficient for the NDE to vanish for the transformed distribution. However, it is not necessary, since the NDE is defined using the mean of the difference whereas the fair adaptation places a requirement on the whole distribution.

5. Practical aspects and extensions

After explaining the main ideas and fairness criteria our method achieves, we turn to discussing the practical aspects and extensions of our method.

5.1 Categorical (and discrete) variables

An important practical aspect of our method is dealing with variables that take values on a discrete domain. There is an immediate problem we encounter in this case. If we think about the mapping $u \to V(U = u)$, we can see that different values of u can correspond to the same value of $V(U = u) = v$, that is the mapping is no longer injective (as opposed to the continuous case). In short, the conditional distribution $U \mid V = v$ is deterministic in the continuous case, and non-deterministic in the discrete case.

Ordered categorical and discrete variables. As a starting point, we describe our method for a binary variable $V \in \{0, 1\}$. Consider the probabilities $p_0 := \mathbb{P}(V = 0 \mid \text{pa}(V),\ A = 0)$ and $p_0' := \mathbb{P}(V = 0 \mid \text{pa}(V),\ A = 1)$. Assume without losing generality that $V = 0$. Then we compute the transformed value $\text{FP}(V)$ as:

- if $p_0' \leq p_0$ then $\text{FP}(V) = 0$
- if $p_0' > p_0$ then

$$\text{FP}(V) = \begin{cases} 0 & \text{with probability } \frac{p_0}{p_0'} \\ 1 & \text{with probability } \frac{p_0' - p_0}{p_0'} \end{cases}$$

We need to generalize this approach to non-binary, discrete variables V. Suppose now that V takes values in $\{1, ..., m\}$. Similarly as above, define $p = (p_1, ..., p_m)$ and $p' = (p_1', ..., p_m')$ where:

$$p_i := \mathbb{P}(V = i \mid \text{pa}(V),\ A = 0) \tag{11}$$
$$p_i' := \mathbb{P}(V = i \mid \text{pa}(V),\ A = 1) \tag{12}$$

These probabilities can, for example, be estimated using probability random forests (Malley et al., 2012). Motivated by the quantile matching assumption, which arises as a solution that induces minimal change in the counterfactual world, we want to find a joint density for $p,\ p'$ that minimizes some transport cost. This can be done by solving the following optimization problem:

$$\begin{aligned} &\min_{\Pi \in \mathbb{R}^{m \times m}} \quad \mathrm{Tr}(\Pi C) \\ \text{s.t.} \quad &\sum_{j=1}^{m} \Pi_{ij} = p_i \quad \forall i \in \{1, ..., m\} \\ &\sum_{i=1}^{m} \Pi_{ij} = p'_j \quad \forall j \in \{1, ..., m\} \end{aligned} \tag{13}$$

where the cost matrix C has entries $C_{ij} = |i - j|^p$. The exact value of p does not really matter, since any $p > 1$ will give the same (unique) solution. When $V = i$, we sample $\mathrm{FP}(V)$ from the distribution given by $\widehat{\Pi}_i$, the i^{th} row of the optimal transport matrix. In particular $\widehat{\Pi}_i$ needs to be normalized, and we let $F_{\widehat{\Pi}_i}$ be the corresponding cumulative distribution function. We then have

$$\mathrm{FP}(V) = F^{-1}_{\widehat{\Pi}_i}(U), \text{ where } U \sim U[0,1] \tag{14}$$

Notice that $\mathrm{FP}(V)$ is not necessarily deterministic. It can happen that $\widehat{\Pi}_i$ has multiple non-zero entries, meaning that the value $V = i$ is coupled with multiple counterfactual outcomes. The reason for this was already mentioned, namely the fact that the conditional distribution $U \mid V = v$ is non-deterministic in the discrete case.

Unordered categorical variables. Let V be categorical and unordered. We first obtain an ordering for it. Suppose V takes values $C_1, ..., C_l$. We then find a bijection $\sigma : \{C_1, ..., C_l\} \to \{1, ..., l\}$ such that

$$\sigma(C_i) \leq \sigma(C_j) \implies \mathbb{P}(Y = 1 \mid V = C_i,\ A = 0) \leq \mathbb{P}(Y = 1 \mid V = C_j,\ A = 0) \tag{15}$$

Then simply define $V' = \sigma(V)$ and use it as a replacement for V. Note that the condition (15) implies that the marginal probability $\mathbb{P}(Y = 1 \mid V' = v,\ A = 0)$ is increasing in v. Implicitly, we assume that the same holds for $A = 1$. That is, we assume that $\mathbb{P}(Y = 1 \mid V' = v,\ A = 1)$ is also increasing in v. Then we can again apply the approach used for discrete variables.

If there is no meaningful ordering, or we have reason to believe that imposing an ordering does not make sense, a slightly different approach is needed. We define $p,\ p'$ the same way as above, with $p_i = \mathbb{P}(V = C_i \mid \mathrm{pa}(V),\ A = 0)$ and $p'_i := \mathbb{P}(V = C_i \mid \mathrm{pa}(V), A = 1)$. We again solve the optimization problem (13), but with a different cost matrix C, namely $C_{ij} = \mathbb{1}(i \neq j)$. When $V = C_i$, the distribution of $\mathrm{FP}(V)$ is given by the (appropriately normalized) i^{th} column of $\widehat{\Pi}$.

5.2 Inherent limitation of the discrete case

In Section 3.1 we gave an optimal transport interpretation of the quantile preservation assumption (QPA). In particular, we state that for two random variables X, Y with distribution functions F_X, F_Y the Wasserstein distance $W_p(X, Y)$ is minimized by matching the quantiles, that is using the optimal transport map given by $F_Y^{-1} \circ F_X$. This map is the optimal transport map for every $p \geq 1$ and also a unique optimal transport map for $p > 1$, since the cost function then becomes strictly convex.

The quantile matching is the *greedy solution.* Note that this approach also extends to the discrete case - a greedy solution[6] is optimal whenever the cost function is strictly convex (Santambrogio, 2015, chap. 2). However, there is a major difference between the continuous and the discrete case.

In the continuous case, using the quantile preservation assumption, we are able to compute the counterfactual values exactly. Richness of the ambient space allows for the optimal transport map to be deterministic, whereas in the discrete setting this is never the case. The solution of the problem (13) gives us a non-deterministic distribution over the counterfactual outcomes. The reason for this is that it is impossible to distinguish individuals which have the same value of a variable V - in some sense, the information coming from the quantiles is compressed.

A possible solution which first comes to mind is to perhaps take the expectation over this randomness. But even if we consider the simplest example, we run into a problem. Consider using a single binary predictor $X \in \{0, 1\}$ distributed as $\mathbb{P}(X = 1 \mid A = 0) = 0.5$ and $\mathbb{P}(X = 1 \mid A = 1) = 0.4$. Suppose that the outcome Y simply equals X. After solving the optimal transport problem, all individuals with $A = 1, X = 0$ would have the counterfactual distribution

$$\mathbb{P}(X^{(fp)} = 1 \mid A = 1, X = 0) = 1 - \mathbb{P}(X^{(fp)} = 0 \mid A = 1, X = 0) = \frac{1}{6}$$

and all other individuals would retain the values they have. But when taking the expectation over this randomness, we have that

$$\mathbb{E}[X^{(fp)} \mid A = 1, X = 0] = \frac{1}{6}$$

meaning we get no additional information to distinguish between individuals with $A = 1, X = 0$. To treat everyone equally, we would have to either assign everyone $X = 1$ or $X = 0$, neither of which options is desirable. Therefore, we use *randomization*, which in this case chooses a "lucky" $1/6$ of the individuals with $A = 1, X = 0$ and sets their counterfactual values $X^{(fp)}$ to 1. For some regression applications integrating outcomes over different counterfactual worlds is meaningful. In that case, taking expectation over the assignment randomness might be sensible. For classification, where labels are either 0 or 1, this might fail, as shown above.

Further, consider two variables $X_1 \sim N(0, 1)$ and $X_2 = \mathbb{1}(X_1 > 0)$. If we only have the variable X_2 available, then it is impossible to distinguish between individuals that have $X_2 = 1$. However, if we use X_1 instead, then no two individuals will be the same - we will always be able to distinguish them. When going from X_1 to X_2, we see that *quantiles are compressed* and they can only be determined up to an interval. This causes the counterfactual value to be non-deterministic. [7]

Finally, we clarify the difference in approach for different types of variables taking values on discrete domains. There are two cases we consider:

[6] A reader familiar with optimal transport will recognize that here we are talking about solutions satisfying the c-monotonicity property.

[7] The result of this is that if we have discrete variables, the strong notion (6) is no longer fulfilled, but the slightly weaker notion holds instead, namely that $\widehat{Y}(A = a, R = R^{(obs)}) \mid A = a, X = x \overset{d}{=} \widehat{Y}(A = a', R = R^{(obs)}) \mid A = a, X = x$ for all a, x .

(a) Discrete and ordered categorical variables:

We solve the optimization problem (13) with the cost matrix $C_{ij} = |i - j|^p$ corresponding to ℓ_p-loss. The cost matrix reflects the fact that, due to an inherent ordering of the values, we wish to penalize larger changes more. For any $p > 1$ the greedy solution is the unique optimal solution.

(b) Unordered categorical variables:

We also solve the optimization problem (13), but with the cost matrix $C_{ij} = \mathbb{1}(i \neq j)$. This cost matrix corresponds to ℓ_0-loss. Since in this case we do not have an inherent ordering structure of the values, we penalize all changes equally. The optimal solution in this case is not unique.

5.3 Sample-level adaptation

Let $\mathcal{G}$ be the causal graph and let R be a choice of resolving variables. Further, let $f(V \mid \text{pa}(V))$ be the density corresponding to variable V. Let $g(\text{pa}(V),\ U^{(V)})$ represent the inverse quantile function of the distribution of V, that is $V = g(\text{pa}(V),\ U^{(V)})$. Sample level fair adaptation is given in Algorithm 2. Notice that our procedure treats the response Y separately. The only reason for this is that Y is unavailable on the test set. The quantile regression step can be done either using random forests (Meinshausen, 2006) or by using an optimal transport approach (Carlier et al., 2016).

5.4 About the training step

We mentioned in Section 3.1 the possibility of using the original labels Y together with the transformed data $\text{FP}(X)$ in the training step. Appendix B discusses this possibility in depth and introduces an additional criterion, called the parity gap condition, given in Equation (17) (this criterion is, however, optional). Here we describe two training options we recommend for the training step.

(A) train the classifier on the original data $(A_k,\ X_k,\ Y_k)_{k=1:n}^{\text{train}}$

(B) train with the adapted data and the adapted labels $\text{FP}(A_k,\ X_k,\ Y_k)_{k=1:n}^{\text{train}}$

For both methods, all of the features of X are used and also the attribute A (leaving out features at training time could cause a violation of condition (17)). Note that, however, for method (B) the attribute $\text{FP}(A_k) = 0$ for everyone, so this feature is not useful. The adapted test data $\text{FP}(A_k,\ X_k,\ Y_k)_{k=1:n}^{\text{test}}$ should always be used to produce the predictions for the test set. A proof showing that methods (A) and (B) satisfy the parity gap condition (17) is given in Appendix B.

An empirical comparison of the two methods on the synthetic examples introduced in Section 6 is given in Appendix G. We note that the two methods perform similarly over a range of different examples and class imbalance settings. Therefore, the better of the two should be chosen via cross-validation as the one with better fairness-accuracy trade-off. For experimental results in Section 6 we by default use training method (B).

Algorithm 2: Fairness Adaptation

```
Input: Data $(A_k, X_k, Y_k)_{k=1:n}$, causal graph $\mathcal{G}$, choice of resolving variables $R$.
Output: Adapted data $\mathrm{FP}(A_k,\ X_k,\ Y_k)_{k=1:n}$
for $V \in \mathrm{de}(A) \setminus R$ in topological order do
    if $V$ continuous then
        estimate the quantiles $(\widehat{U}_k^{(V)})_{k=1:n}$ of $V$ in the distribution $f(V \mid \mathrm{pa}(V))$
          using quantile regression on the data $(V_k,\ \mathrm{pa}(V_k))_{k=1:n}$
        using $(V_k,\ \mathrm{pa}(V_k), \widehat{U}_k^{(V)})_{k=1:n}$ obtain an estimator $\widehat{g}(\mathrm{pa}(V),\ U^{(V)})$ of
          $g(\mathrm{pa}(V),\ U^{(V)})$
    else
        case 1. $V$ discrete and $V \neq Y$ do
            estimate the probability distributions $\widehat{p}(\mathrm{pa}(V_k))_{k=1:n}$ as in Equations
              (11)-(12)
            obtain the transformed probability distributions $\widehat{p}(\mathrm{FP}(\mathrm{pa}(V_k)))_{k=1:n}$
            $\forall k$ solve the optimal transport problem (13) between $\widehat{p}(\mathrm{pa}(V_k))$ and
              $\widehat{p}(\mathrm{FP}(\mathrm{pa}(V_k)))$ with $\ell_p$-loss to get $(\widehat{\Pi}^k)_{k=1:n}$
        case 2. $V = Y$ do
            perform case 1. restricted to the training set
    for all $k$ with $A_k = 1$ and $V_k$ known do
        if $V$ continuous then
            $\mathrm{FP}(V_k) \leftarrow \widehat{g}(\mathrm{FP}(\mathrm{pa}(V_k)),\ \widehat{U}_k^{(V)})$
        else
            $\mathrm{FP}(V_k) \leftarrow$ sample from the distribution $\widehat{\Pi}_{V_k}^k$ as in Equation (14)
return $FP(A_k,\ X_k,\ Y_k)_{k=1:n}$
```

5.5 Method extensions

There are two methodological extensions of our approach that we briefly discuss, leaving out some of the detail.

Is there really a baseline? So far we have considered the subpopulation $A = 0$ to be the baseline. This choice is somewhat arbitrary. We briefly comment on the implications of choosing a baseline.

Firstly, the choice of the baseline can influence the number of positive outcomes predicted by $\widehat{Y}$. Imagine that we are trying to predict recidivism on parole, with race being the protected attribute. If we adapt the data using the white subpopulation as baseline, then our predictor would predict fewer recidivism outcomes (compared to using the non-white population as baseline), since the currently measured base rates of recidivism are unequal between the groups in favor of the white population.

Secondly, if we have discrete variables, then our procedure will include some randomization. There is no randomization for the baseline population, but there is for the rest. If the baseline is the advantaged group, then randomization can serve as *positive discrimination* and might be seen as acceptable. However, we might want to consider an approach in which both subpopulations are randomized equally. We briefly discuss how we might split the burden of randomization between the subpopulations.

A non-baseline approach. We previously discussed adapting the data to the $A = 0$ baseline using Algorithm 2, which gives us the pre-processed version of the data, which we here label $X^{(fp),A=0}$. Of course, the same procedure can be applied to obtain the version corresponding to the $A = 1$ baseline, which we label $X^{(fp),A=1}$. Then we can use the following approach:

1. Obtain $(X^{(fp),A=0}, Y^{(fp),A=0})$ and $(X^{(fp),A=1}, Y^{(fp),A=1})$ using Algorithm 2.

2. Concatenate the two versions to obtain
$$X^\star = (X^{(fp),A=0}, X^{(fp),A=1}).$$

3. Build predictors $\widehat{\pi}^{A=0}(x^\star), \widehat{\pi}^{A=1}(x^\star)$ that estimate the probabilities $\mathbb{P}(Y^{(fp),A=0} = 1 \mid X^\star = x^\star)$, $\mathbb{P}(Y^{(fp),A=1} = 1 \mid X^\star = x^\star)$ respectively.

4. For any test observation with $X^\star_{\text{test}} = x^\star_{\text{test}}$ return the predicted probability of
$$\widehat{\pi}(x^\star_{\text{test}}) = \frac{\widehat{\pi}^{A=0}(x^\star_{\text{test}}) + \widehat{\pi}^{A=1}(x^\star_{\text{test}})}{2}.$$

We offer an interpretation of the approach above. First we combine the information from the two worlds in which $A = 0$ and $A = 1$. We then use the joint information to predict probabilities of positive outcomes in both of these worlds. In the final step, we combine the probabilities from the two worlds by simply taking the mean probability. In this way, we obtain probability estimates for positive outcomes, which can then be used to construct a classifier by thresholding.

Edge specific extension. We quickly mention another possible extension of our method. So far we discussed resolving variables, on which the effect of A is deemed fair. Sometimes deciding if a variable is resolving might not be straightforward. For example, we might see the causal effect of several causal parents as being fair, but of several as being unfair. In some sense, our method so far was focused on the nodes in the causal graph $\mathcal{G}$. An extension of this, which focuses more on specific edges in $\mathcal{G}$ is possible. The edge-specific case could be seen as a special case of path-specific discrimination removal, discussed in (Chiappa, 2019). In this context, already known algorithms for identifying path-specific effects could also be useful (Shpitser, 2013). A short discussion of the edge-specific case, motivated by an example, is given in Appendix E.

6. Experimental results

An implementation of our method, which uses tree ensembles for the quantile learning step, is available as the `fairadapt` package on CRAN. Our experimental results consist of two parts. In the first part, we look at synthetic examples which demonstrate how our methodology offers flexibility compared to possibly prohibitively strong demographic parity. In the second part, we look at the method performance on two real world datasets, comparing it to several different baseline methods.

6.1 Measures of fairness and performance

Before displaying the experimental results, we discuss all the measures that are used for assessment of our classifiers. For measuring performance, we report on accuracy in the simple classification tasks. It is, however, sometimes desirable to work with probability predictions, in which case we report the area under the receiver operator characteristic (AUC).

Fairness measures. There are two fairness measures we use. To assess demographic parity, we use the *parity gap*, defined as $\mathbb{P}(\widehat{Y} = 1 \mid A = 0) - \mathbb{P}(\widehat{Y} = 1 \mid A = 1)$. When dealing with probability predictions, we simply report the parity gap at the 0.5 threshold.

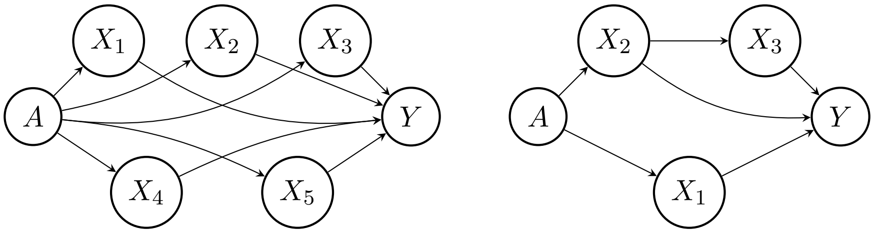
In order to assess calibration, we need a measure for it. We introduce the *k-level inter-group calibration measure*. This score is in spirit very similar to expected calibration error (Naeini et al., 2015), but is not strictly a calibration measure. Suppose we have the predicted positive probabilities $\mathbb{P}(\widehat{Y} = 1 \mid X = x)$ and the true labels Y. We start by splitting the individuals with $A = 0$ into k groups, based on the predicted probability of $\mathbb{P}(\widehat{Y} = 1 \mid X = x)$. In particular, if $\mathbb{P}(\widehat{Y} = 1 \mid X = x) \in [\frac{i}{k}, \frac{i+1}{k})$, then the individual is assigned to group G_i. In each group we compute the mean of the true outcomes Y for that group, $\mathbb{E}[Y \mid G_i]$, which is simply the proportion of positive outcomes in the group G_i. Assume that the vector $\boldsymbol{c}^{A=0}$ contains these proportions for each group. We compute $\boldsymbol{c}^{A=1}$ for the $A = 1$ population in the same way. The k-level inter-group calibration measure is defined as

$$||\boldsymbol{c}^{A=0} - \boldsymbol{c}^{A=1}||_1.$$

Note that for a well-calibrated score, this measure should be small.

Synthetic A	Synthetic B
$A \leftarrow \text{Bernoulli}(0.5)$	$A \leftarrow \text{Bernoulli}(0.5)$
$X_i \leftarrow -\frac{A}{4} + \frac{1}{8} + \epsilon_i \quad \text{for } i \in \{1, ..., 5\}$	$X_i \leftarrow -\frac{A}{4} + \frac{1}{8} + \epsilon_i \quad \text{for } i \in \{1, 2\}$
$Y \leftarrow \text{Bernoulli}(\text{expit}(\sum_{i=1}^{5} X_i))$	$X_3 \leftarrow \frac{1}{4} X_2 + \epsilon_3$
	$Y \leftarrow \text{Bernoulli}(\text{expit}(\sum_{i=1}^{3} X_i))$

Table 2: Structural equation models for the two synthetic examples A and B. All the noise variables ϵ_i are independent and $\text{expit}(x) = \frac{e^x}{1+e^x}$.



(A) (B)

Figure 1: A graphical model representation of the SEMs used for Synthetic examples A and B.

6.2 From parity to calibration

Earlier in the text we claimed that our method can offer suitable relaxations of demographic parity. Namely, Theorem 12 shows that in a simple case demographic parity is achieved when none of the variables are resolving, and that calibration can be achieved if all of the variables are resolving. Depending on the choice of the resolving variables, we can interpolate between these two notions of fairness. Roughly speaking, the larger the resolving set is, the larger the effect of A is in the data. In that case, the predictor $\widehat{Y}$ is closer to the unconstrained maximum accuracy $\widehat{Y}^{\text{max-acc}}$ predictor (which we assume is calibrated), meaning that we are closer to satisfying calibration. The smaller the resolving set is, the smaller the effect of A is in the data, meaning that we are closer to demographic parity.

We demonstrate this by looking at two synthetic examples, with their structural equation models given in Table 2. All the noise terms ϵ_i are independent $N(0, 1)$ variables and $\text{expit}(x) = \frac{e^x}{1+e^x}$. In words, Y follows a logistic regression model based on the X_i's. The causal graphs of the two synthetic examples are given in Figure 1. In both examples, we analyze the AUC-parity gap and parity gap-calibration score trade-offs via the resolving variables. Two baseline methods were implemented, to compare our results. These are *reweighing* (Kamiran and Calders, 2012) and *fair reductions* (Agarwal et al., 2018). Both of these methods aim to achieve demographic parity. Therefore, we should compare them to our method with no resolving variables. More details on comparison methods are given shortly in Section 6.3. The fair reductions approach performs poorly on both of these task

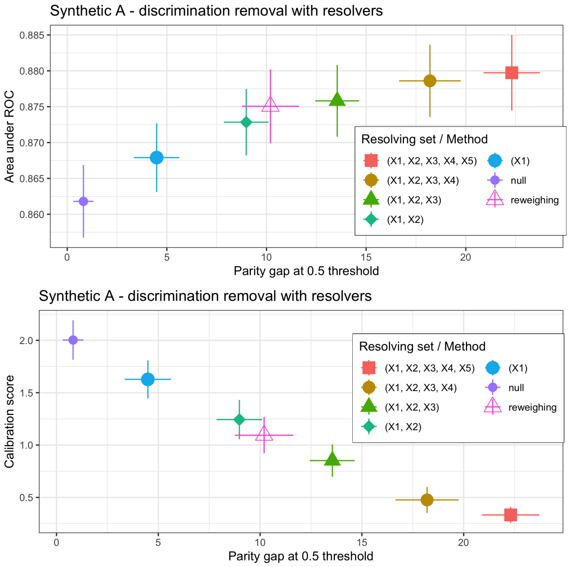


Figure 2: AUC-parity and parity-calibration score trade-off for example A. Vertical bars represent standard deviations obtained from 10 repeats.

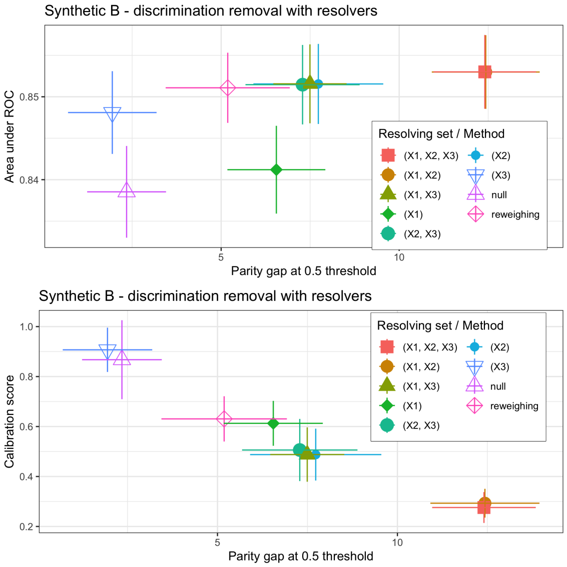


Figure 3: AUC-parity and parity-calibration score trade-off for example B. Vertical bars represent standard deviations obtained from 10 repeats.

(for a range of parameter values ϵ of the method), so it is not included in the final analysis of the results (full plots including the fair reductions approach are given in Appendix H).

In example A we enlarge the set of resolving variables stepwise, including X_i at step i. For example B, we try out all possible subsets of resolving variables. We run our method ten times, with 5000 training and test samples generated from the given SCMs. A logistic regression classifier is used after applying `fairadapt`. On each repeat we measure the AUC, parity gap and the calibration score. The results are shown in Figures 2 and 3. Vertical error bars in the figures represent the standard deviations of respective measures obtained from the ten repeats.

Example A. In Figure 2 the AUC and the parity gap are increasing, whereas the calibration score is becoming smaller, as we enlarge the resolving set. The case $R = \{\emptyset\}$ corresponds to demographic parity. It also shows that the population level notion in Equation (5) is satisfied, even if one does not believe in using the quantile preservation assumption (QPA). The case $R = \{X_1, ..., X_5\}$ corresponds to calibration (since the underlying classifier is consistent). By varying the choice of resolving variables, our method can offer a range of fairness-accuracy trade-offs between the two notions. Finally, we note that the baseline method of reweighing obtains better accuracy than `fairadapt` with $R = \{\emptyset\}$, but fails to eliminate discrimination fully.

Example B. Note that setting X_2 to be resolving implicitly sets X_3 to be resolving, since X_2 is the only parent of X_3. Therefore, if X_2 does not change in the adaptation procedure, neither will X_2. Setting $R = \{X_1, X_2\}$ has almost the same performance as setting $R = \{X_1, X_2, X_3\}$. Similarly, resolving sets $\{X_2, X_3\}$ and $\{X_2\}$ show very similar results, as expected. This example helps us to illustrate that the adapted value of a variable V also depends on whether some of its parents are resolving. In particular, if all variables in $\text{pa}(V)$ are resolving, this will implicitly set V to be resolving as well.

6.3 Real data experiments

We next look at real data experiments. We summarize all the baseline methods against which we benchmark our results. In the real data comparisons, we only consider the case of demographic parity (meaning no resolving variables) as other comparisons methods are designed to achieve precisely this notion.

Baseline methods. The comparison methods that we look at are:

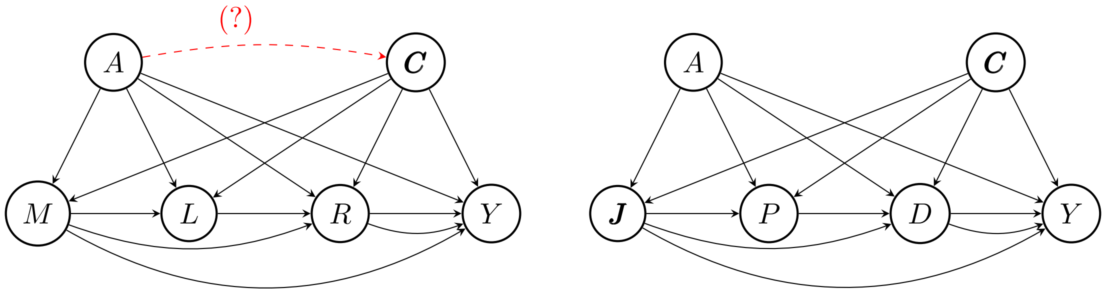
- standard implementation of random forests (Wright and Ziegler, 2015), serving as a fairness-ignorant baseline
- *fairness through unawareness* - RF applied to the data after excluding the protected attribute
- the *reweighing* preprocessing method (Kamiran and Calders, 2012), which learns specific weights for the combinations of the class label and the protected attribute which are then used for building a classifier (in this case we use the logistic regression classifier and the implementation from the IBM toolkit (Bellamy et al., 2018))
- the *reductions* approach (Agarwal et al., 2018) casts the fairness problem in a linear programming (LP) form in order to find a sample-weighted classifier which satisfies the desired fairness constraint (we again use logistic regression for our classifier that allows sample-weighting and vary the fairness constraint violation parameter $\epsilon \in \{0.1, 0.01, 0.001\}$)

UCI Adult. The Adult dataset from the UCI machine learning repository (Lichman et al., 2013) contains information on 48842 individuals and the outcome to be predicted is whether an individual has a yearly income of more than 50 thousand dollars. The data comprises of the following features[8]:

- gender, labeled A, which we consider to be the protected attribute
- demographic information C - including age, race and nationality
- marital status M and years of education L
- work related information R - job occupation, hours of work per week and work class
- a binary outcome Y representing whether a person's income exceeds 50000 dollars a year

The UCI Adult dataset has been previously analyzed as an application of different fairness procedures, for instance in (Nabi and Shpitser, 2018) and (Chiappa, 2019). The proposed causal graph for the dataset is presented in Figure 4(a). While we do agree that this causal

[8] The original dataset contains a few more features, but we focus on those that have been used in previous fairness applications.



(a)

(b)

Figure 4: (a) the causal graph (black edges) claimed to correspond to the UCI Adult dataset. The additional red, dashed edge corresponds to a sampling bias in the data; (b) causal graph of the COMPAS dataset.

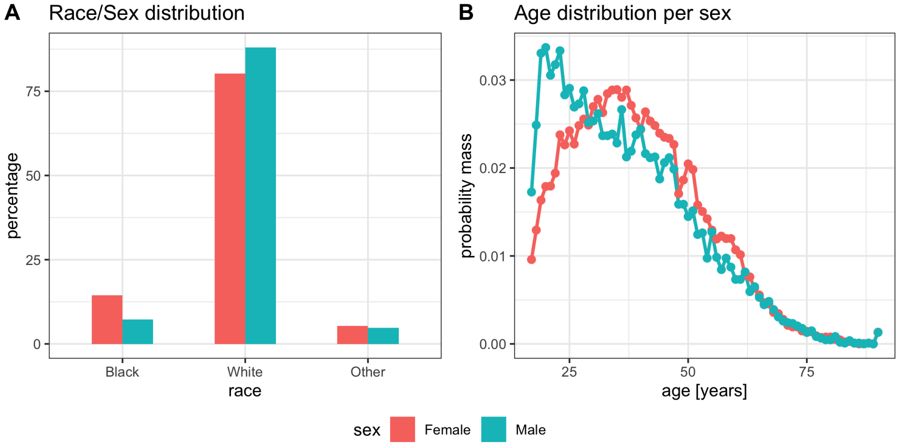


Figure 5: Plots of influence of sex on race (**A**) and age (**B**) in the UCI Adult dataset.

graph makes sense intuitively, care needs to be taken because the sampling bias can induce dependencies that have no explanation in reality. This is precisely the case with the UCI Adult dataset, which we can observe by inspecting the relation between two features in $\boldsymbol{C}$ (age and race) and the protected attribute A. From the plots in Figure 5 we see that in the dataset gender is not independent of age and race, as the causal graph would imply. To solve the problem (in order to satisfy our assumption that A is a root node), we subsample the dataset in order to mitigate the sampling bias. Details about how we pre-processed the dataset are given in Appendix A.

COMPAS dataset. The second real dataset we analyze is the COMPAS dataset (Larson et al., 2016) which contains the following features:

- outcome Y is recidivism whilst on parole within a two year period
- protected attribute A in this case is race (White vs. Non-White)

- demographic information $\boldsymbol{C}$
- juvenile offense counts $\boldsymbol{J}$, count of prior offenses P and degree of the charge D

The causal graph that we propose is given in Figure 4(b). The reader here might disagree that this example falls into the class of Markovian models (that is, there could be some latent confounding).

We select several individuals in the dataset which are Non-White, male and of age 30. We look at their values of juvenile counts and prior counts (J_1, J_2, J_3, P) before and after applying `fairadapt`

```
> compas[id, offense.count]
          juv_fel_count  juv_other_count priors_count
241                   0                0            4
646                   0                0            8
807                   0                0           17
1425                  2                0           20
1470                  1                2           15
> adapted_compas[id, offense.count]
          juv_fel_count juv_other_count priors_count
241                   0               0             3
646                   0               0             5
807                   0               0            13
1425                  0               0            11
1470                  0               2             9
```

We can notice that fair adaptation reduces the number of offenses for these individuals in the adapted version, since in the dataset the baseline population (white) has fewer offenses on average. Notice how the transformed values allow *fair-twin inspection.* For instance, individual 1470 had values $(1, 2, 15)$ for (J_1, J_3, P). His fair-twin (in a fair world) would have attained values $(0, 2, 9)$ instead. Hypothetical statements like "if you were white, your juvenile offense counts would have been J_1, J_2, J_3, in turn resulting in prior count of P" now become possible. This part of our method, however, rests on the assumption from Definition 1.

Results. For both UCI Adult and COMPAS, we split the dataset into 75% training and 25% testing randomly 20 times. Each time, we apply all the baseline methods and our `fairadapt` method, measuring accuracy and the parity gap each classifier achieves. Figures 6 and 7 summarize the obtained results. For the Adult dataset, no method is better than `fairadapt` on both criteria. For the COMPAS dataset, the reweighing method performs slightly better and `fairadapt` is also contained within the confidence interval fair reductions ($\epsilon = 0.1$). We note our method has very satisfying performance. We can also see that our method is able to exploit the information in the descendants of the attribute A (as opposed to some previous approaches (Kusner et al., 2017)), which is very important in practice. On top of this, we mention that our method has the ability to relax the fairness criterion via resolving variables, has a causal interpretation and allows fair-twin inspection (under the QPA). Even if the QPA is not true (which is untestable), the results still show that the

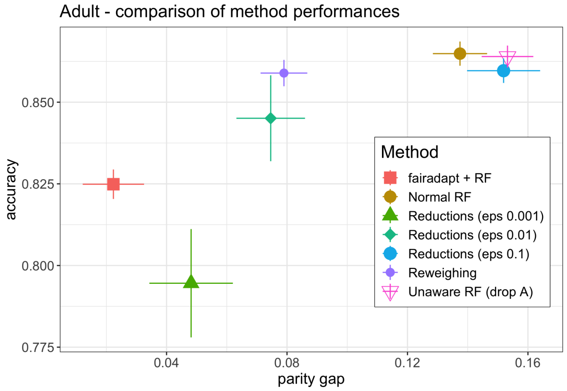


Figure 6: Comparison of the performance of different fairness methods on the UCI Adult dataset. Vertical bars represent standard deviations obtained from 20 repeats.

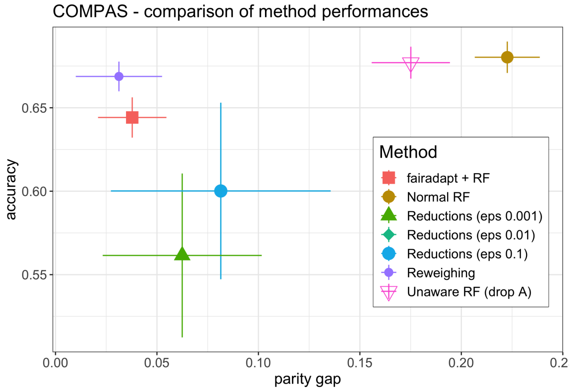


Figure 7: Comparison of the performance of different fairness methods on the COMPAS dataset. Vertical bars represent standard deviations obtained from 20 repeats.

condition (5) is achieved. Achieving this condition does not rely on the QPA, as we discuss in Section 3.

Finally, we take a look at how `fairadapt` affects the distribution of the positive outcome probabilities. We plot the densities of $\mathbb{P}(\widehat{Y} = 1 \mid A = a)$ for both levels of A for the two cases of not applying and applying `fairadapt`. The results are shown in Figures 8 and 9. Note that the densities are much closer when applying `fairadapt`, indicating a clear reduction in discrimination.

7. Conclusion

In the final section we revisit some of the ideas discussed previously and conclude our argument.

About observational criteria. Causal and observational notions of fairness have an inherent link. If the protected attribute is a root node, the intervention on A is equivalent to conditioning on A. Causality is necessary, not just to provide new criteria, but to give meaning to the existing observational criteria used.

About fair data adaptation. We conclude that `fairadapt` shows competitive performance compared to other baseline methods in the case of demographic parity. It also gives a causal and interpretable perspective on the data projection that is carried out. Further, it offers various relaxations of demographic parity, all the way to the case of calibration, which is achieved when all the variables are considered to be resolving. The output of fair data adaptation also allows us to see which values individuals were assigned in the projection procedure. This helps justify and interpret why a certain individual was given his prediction.

About the current datasets and methods. We emphasize that it would be beneficial for the advancement of fairness if there were established real world datasets with agreed-

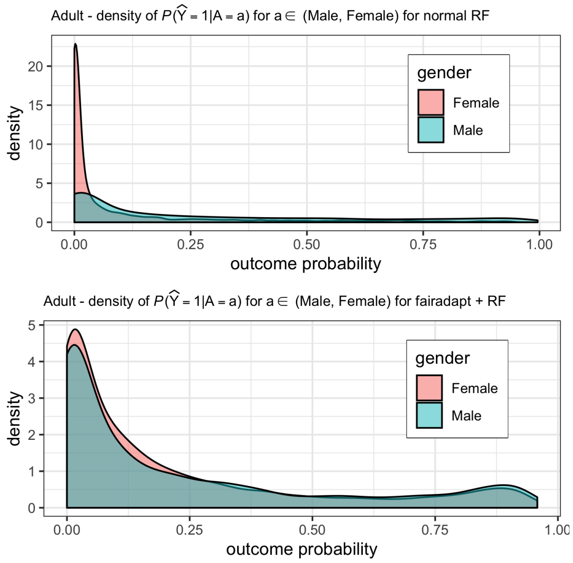


Figure 8: Change in the positive outcome probability density due to applying `fairadapt` to UCI Adult.

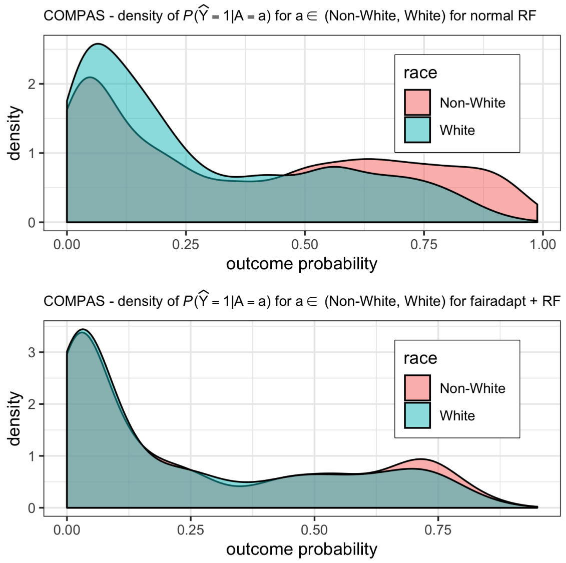


Figure 9: Change in the positive outcome probability density due to applying `fairadapt` to COMPAS.

upon causal graphs. This would allow different authors to compare their methods in a meaningful way, demonstrating the performance and measuring different fairness criteria. Having a benchmark for algorithm performance and fairness criteria achieved could also help us understand how and why different methods yield different results on the same datasets. We feel like this is not yet the case and that much more could be done on this front.

About future work. We have discussed a method which achieves certain fairness criteria and shown how it can be used in practice. However, this is only the very first step of fairness. A big component of the whole problem are the temporal implications of fairness criteria on the well-being of different groups. Some interesting work on this topic we are currently aware of includes (Liu et al., 2018; Kannan et al., 2019; Milli et al., 2019; Hu and Chen, 2018). Although many of the fairness criteria make intuitive sense and perhaps have some philosophical backing, we have no reason to convince ourselves that they are necessarily doing the right thing in terms of their long-term effect. This is a serious question that requires much further consideration.

Acknowledgments

We would like to thank Domagoj Ćevid, Yuansi Chen and Federico Glaudo for useful discussions and suggestions that greatly improved our work.

Appendix A. Proof of Theorem 6

Proof We prove that the projection FP$(\cdot)$ as defined in Algorithm 1 satisfies

$$\mathrm{FP}(X(U=u)) \quad = \quad (N(A=0, R=r, U=u),\ r)$$

for all u, r such that $R(U=u) = r$. Under the $do(A=0, R=R^{(obs)})$ intervention and the realization of quantiles $U=u$ the assignment equations of A and R change to

$$\begin{aligned} A &\leftarrow 0, \\ R &\leftarrow r. \end{aligned}$$

For any V non-descendant of A or R we have that

$$\mathrm{FP}(V(U=u)) = V(A=0, R=r, U=u).$$

We proceed inductively. Let $U^{(V)}$ be the component of $U=u$ corresponding to variable V. In the first step, for any $V \in \mathrm{ch}(A) \setminus R$ we can show that

$$\begin{aligned} V(A=0, R=r, U=u) &= g(\mathrm{pa}(V)(A=0, R=r, U=u), U^{(V)}) \qquad (16) \\ &= g(\mathrm{FP}(\mathrm{pa}(V)), U^{(V)}) \\ &= \mathrm{FP}(V(U=u)) \end{aligned}$$

where the first equality holds by definition of the intervention and the quantile preservation assumption (QPA, Definition 1), the second because we showed $\mathrm{FP}(V(U=u)) = V(A=0, R=r, U=u)$ for all $V \in \mathrm{nde}(A)$ (here $\mathrm{nde}(A)$ are non-descendants of A), the third from the definition of Algorithm 1. Using the fact that Algorithm 1 goes through variables V in topological order, inductively we can show $\mathrm{FP}(V(U=u)) = V(A=0, R=r, U=u)$ for any V in $\mathrm{ch}(\mathrm{ch}(A))\backslash R$ and so on (sometimes called *recursive substitution*). This shows that strong resolved fairness holds under the QPA. If the QPA is not used, then the equality (16) does not hold anymore. However, even without QPA it still holds that $V(A=0, R=R^{(obs)}) \overset{d}{=} g(\mathrm{pa}(V)(A=0, R=R^{(obs)}), U^{(V)})$ where $U^{(V)}$ is a uniform $\mathrm{U}[0,1]$ random variable independent of the quantiles U. This is enough to guarantee that $\mathrm{FP}(N(A=a, R=R^{(obs)})) \overset{d}{=} \mathrm{FP}(N(A=a', R=R^{(obs)}))\ \forall a, a'$. From this it follows that for any classifier $\widehat{Y} = f \circ \mathrm{FP}$ we have that

$$\widehat{Y}(A=a, R=R^{(obs)}) \overset{d}{=} \widehat{Y}(A=a', R=R^{(obs)}).$$

■

Appendix B. Resolver-induced parity gap

The reader might wonder if adapting the labels Y is necessary in our procedure. Sometimes it is possible to obtain better performance when using the original Y labels. To discuss this, we look at a simple example of a linear, additive regression model. The details of it

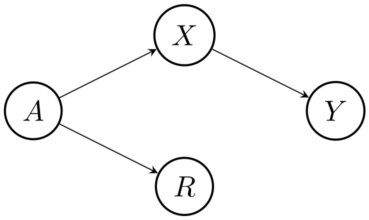
Graphical model representation	Generating mechanism (SCM)
	$A \leftarrow \text{Bernoulli}(0.5)$ $X \leftarrow \frac{1}{2}\mathbb{1}(A=0) + \epsilon_X$ $R \leftarrow \frac{3}{4}\mathbb{1}(A=0) + \epsilon_R$ $Y \leftarrow \frac{1}{2}X + \epsilon$

Table 3: A full description of the example discussed in the text.

are given in Table 3. Suppose we had data coming from this model. Assume that we want the variable R to be resolving. Then our data adaptation would change the value of X so that

$$\text{FP}(X) = X - \frac{1}{2}\mathbb{1}(A=0)$$

in order to remove the effect of A from X. Suppose that, after this, we want to use the transformed values $\text{FP}(X)$ and R to construct a predictor for Y. Since Y can in fact be written as

$$Y = \frac{1}{2}X + \epsilon = \frac{1}{2}\text{FP}(X) + \frac{1}{4}\mathbb{1}(A=0) + \epsilon$$

we can notice that Y and R are correlated. Furthermore, $Y \not\perp\!\!\!\perp R \mid \text{FP}(X)$. Therefore, if we linearly regress Y onto $\{\text{FP}(X), R\}$ and obtain $\widehat{Y}$, then R will have a non-zero coefficient. In fact, by using R, $\widehat{Y}$ will predict higher values for the $A=0$ population. However, variable R has no causal effect on Y, yet we are discriminating based on it.

We believe it is good practice to bound the maximum parity gap that can occur in the presence of the resolving variables. The difference between subpopulations should be at most the difference resulting from the causal effect of the resolvers (although the reader might not think this is strictly necessary). We introduce the following definition.

Definition 13 (Resolver-induced parity gap) *We say that a predictor $\widehat{Y}$ for Y satisfies the resolver-induced parity gap with respect to a set of resolving variables R if*

$$\mathbb{E}\big[\widehat{Y}(A=0) - \widehat{Y}(A=1)\big] \leq \mathbb{E}\big[Y(A=0) - Y(A=0, R=R(A=1))\big]. \tag{17}$$

The quantity on the LHS of criterion (17) is the parity gap of our predictor $\widehat{Y}$. The quantity on the RHS measures the causal effect of the resolvers R on the outcome Y. We believe this should be the maximum parity gap we allow for $\widehat{Y}$. The example above shows that using the original labels Y can cause a violation of criterion (17).

We next show that by using the transformed labels $\text{FP}(Y)$ we will not violate criterion (17). If we allow $\widehat{Y}$ to be a probability predictor (instead of a $\{0,1\}$ classifier), then

$$f^\star(\text{FP}(X)) = \mathbb{E}\big[\text{FP}(Y) \mid \text{FP}(X)\big]$$

satisfies

$$\begin{aligned}\mathbb{E}\big[f^\star(\mathrm{FP}(X))\big] &= \mathbb{E}\big[\mathbb{E}[\mathrm{FP}(Y) \mid \mathrm{FP}(X)]\big]\\ &= \mathbb{E}\big[\mathrm{FP}(Y)\big]\\ &= \mathbb{E}\big[Y(A=0, R=R^{(obs)})\big],\end{aligned}$$

where the last equality comes from Theorem 6. Therefore, $\widehat{Y} = f^\star \circ FT$ satisfies

$$\mathbb{E}\big[\widehat{Y}(A=a)\big] = \mathbb{E}\big[Y(A=0, R=R^{(obs)})\big]$$

from which it follows that this $\widehat{Y}$ satisfies the condition (17).
Similar reasoning can be used for training method (A), that is if $f^\star$ is such that

$$f^\star(X) = \mathbb{E}\big[Y \mid X\big]$$

then for $\widehat{Y} = f^\star \circ FT$ we have that

$$\begin{aligned}\mathbb{E}\big[\widehat{Y}(A=0, R=R^{(obs)})\big] &= \mathbb{E}\big[f^\star(X(A=0, R=R^{(obs)}))\big]\\ &= \mathbb{E}\big[\mathbb{E}\big[Y(A=0, R=R^{(obs)}) \mid X(A=0, R=R^{(obs)})\big]\big]\\ &= \mathbb{E}\big[Y(A=0, R=R^{(obs)})\big]\end{aligned}$$

from which it follows that this $\widehat{Y}$ also satisfies the condition (17). Note that consistent classifiers f will converge to the population optimal prediction $f^\star$. For small sample sizes, the parity gap-condition is only fulfilled modulo sampling noise. The reader might wonder why we work with probability predictions instead of $\{0, 1\}$ predictions. This is discussed in depth in Section 5. A brief discussion related to the above argument is also given in Appendix C.

To summarize: if the reader believes that the amount of discrimination coming from resolving variables R should be in line with the causal of effect of R on Y, then criterion (17) should hold and the transformed labels $\mathrm{FP}(Y)$ should be used. However, if the reader does not think the discrimination level must be explained by the causal effect, both the adapted and the unadapted labels can be used and the criterion (5) will still hold.

Appendix C. Probability predictions satisfying resolver-induced parity gap

Take the following simple example

$$\begin{aligned}A &\leftarrow \text{Bernoulli}(0.5)\\ X_1 &\leftarrow \frac{1}{2}\mathbb{1}(A=0) + \epsilon_1\\ X_2 &\leftarrow \frac{2}{3}(\mathbb{1}(A=0) - \frac{1}{2}) + \epsilon_2\\ Y &\leftarrow \text{Bernoulli}(\text{expit}(X_1 + X_2))\end{aligned}$$

where ϵ_1, ϵ_2 are both $N(0, \sigma^2)$ variables with $\sigma^2 = 0.05$. Variable A represents gender, with $A = 0$ being the male population. Suppose that X_2 is resolving and X_1 is not.

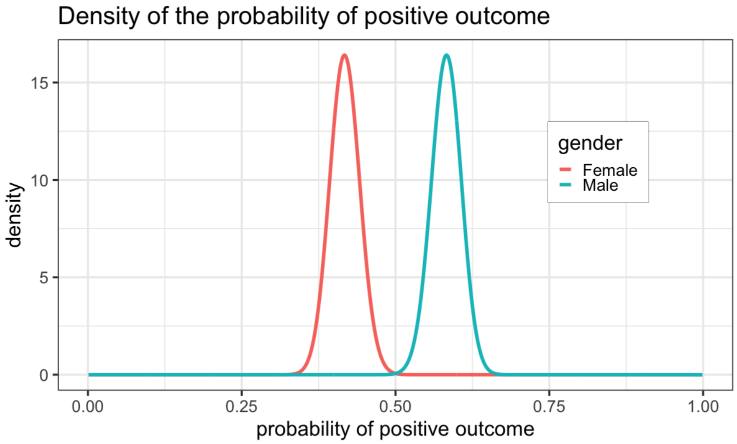


Figure 10: Density of the probability of positive outcome $\mathbb{P}(\text{FP}(Y) = 1)$.

After adaptation (assuming no estimation error) we have that $\text{FP}(X_1) \leftarrow \epsilon_1$ and $\text{FP}(Y) \leftarrow$ Bernoulli(expit($\text{FP}(X_1) + X_2$)). Plot of the density of the probability of a positive outcome $\mathbb{P}(\text{FP}(Y) = 1 \mid A = a)$ are shown in Figure 10. Note that an optimal probability predictor $\widehat{Y} = \mathbb{E}\big[\text{FP}(Y) \mid \text{FP}(X)\big]$ would have

$$\mathbb{E}\big[\widehat{Y}(A=0) - \widehat{Y}(A=1)\big] = \mathbb{E}\big[\text{FP}(Y) \mid A=0\big] - \mathbb{E}\big[\text{FP}(Y) \mid A=1\big] \approx 0.164.$$

However, an optimal $\{0, 1\}$ classifier $\widetilde{Y}$ trying to minimize (for example) the L_2-loss would simply be constructed as $\widetilde{Y} = \mathbb{1}(\widehat{Y} \geq \frac{1}{2})$. Note that (referring to Figure 10) for this $Y^{(fp)}$ we have that

$$\mathbb{E}\big[\widetilde{Y}(A=0) - \widetilde{Y}(A=1)\big] \approx 1.$$

Due to examples like this, the criterion (17) was defined for probability and not class predictions. A much more involved, general discussion of this problem is given in Section 5.

Appendix D. Empirical comparison of quantile regression methods

The `fairadapt` package allows for three different quantile regression methods: quantile regression forests, linear quantile regression and monotone composite quantile regression neural network. All three methods are applied to Synthetic B (given in Section 6) and Synthetic C examples. We use 5000 training and testing samples and no resolving variables. Each experiment is repeated 10 times. Synthetic C example is constructed to demonstrate how a more complex, non-parametric case could be handled with `fairadapt`. The causal graph is shown in Figure 11 and the SCM is given in Table 4. The performances of different methods are given in Figures 12 and 13, respectively.

Example B shows that if the underlying generating mechanism of the data is linear, the linear quantile regression approach will perform well. We recommend using it for cases where linear approximation works well. This approach is also computationally efficient. The neural network based approached will not perform much worse (in a linear setting), but is

Synthetic C
$A \leftarrow \text{Bernoulli}(0.5)$
$X_i \leftarrow \mathbb{1}(A=0) + u_i \quad \text{for } i \in \{1,2\}$
$X_3 \leftarrow X_1 X_2 + X_1^2 + \epsilon_3$
$X_4 \leftarrow \frac{1}{4} X_3^2 + 2X_2^2 + \epsilon_4$
$Y \leftarrow \text{Bernoulli}(\text{expit}(\frac{1}{2}(\sum_{i=1}^{4} X_i - 7)))$

Table 4: Structural equation model for the two synthetic C example. Noise variables u_1, u_2 are uniform U$[0,1]$ and ϵ_3, ϵ_4 are N$(0,1)$. All noise variables are independent and $\text{expit}(x) = \frac{e^x}{1+e^x}$.

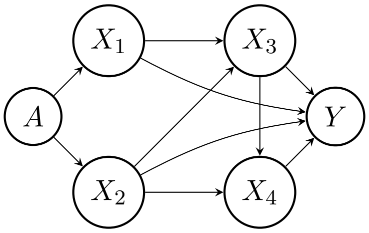


Figure 11: A graphical model representation of the SCM for Synthetic C example.

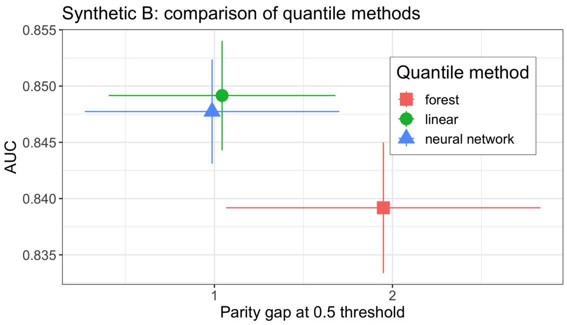


Figure 12: AUC and parity at 0.5 threshold for applying `fairadapt` with different quantile regression methods on Synthetic B example. Vertical bars represent standard deviations obtained from 10 repeats.

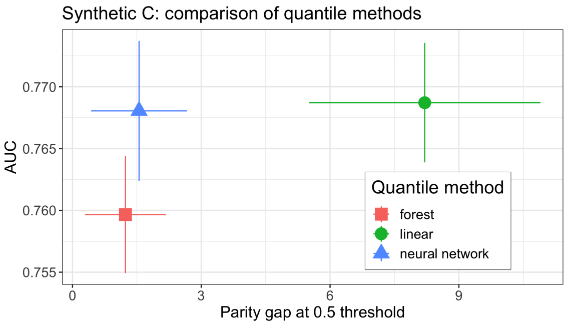


Figure 13: AUC and parity at 0.5 threshold for applying `fairadapt` with different quantile regression methods on Synthetic C example. Vertical bars represent standard deviations obtained from 10 repeats.

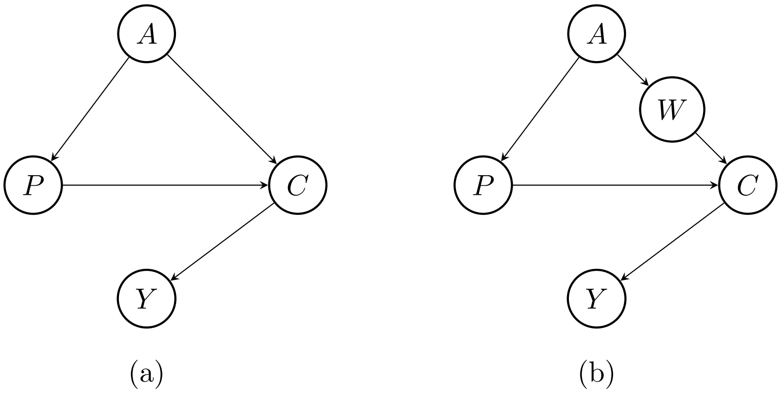


Figure 14: (a) example that motivates the edge extension of the idea of resolving variables; (b) example where an edge specific extension might arise naturally.

much more computationally intensive. Example C shows that for a non-linear setting, the linear method has trouble fully eliminating discrimination. The non-parametric methods still perform well, though. Generally, we recommend using the forest based approach, because of the non-parametric nature and computational speed. However, we note that for smaller sample sizes, the neural network approach might in fact be the best option.

Appendix E. Edge specific extension

Consider a dataset consisting of the following features[9]:

- protected attribute A, in this case race
- information about amount of policing the person experiences, P (given explicitly or perhaps implicitly through a ZIP code)
- information about prior convictions C
- recidivism outcome Y when the person is released on parole

A possible causal graph for this dataset is given in Figure 14(a). One approach could be to treat the variable C as resolving. If, however, information about policing is available, we might want to account for this. Suppose that the difference in prior convictions between the black and white population was partly due to the fact that black people experience more policing. We would, in this case, consider this effect unfair. Therefore, we need to find a way to remove the $A \to P \to C$ effect, but keep the direct $A \to C$ effect (removal of path-specific effects as this is discussed in Nabi and Shpitser (2018)). This example demonstrates that sometimes we perhaps want to have *partially resolving variables.*

We argue that sometimes it is hard to choose if a variable is simply resolving or non-resolving. Going back to the case of policing from Figure 14, it would be difficult to determine whether the prior convictions variable C is resolving or non-resolving. In some

[9] This example, not surprisingly, is motivated by COMPAS.

sense, both choices would be wrong. We therefore think the approach of choosing which edges to remove allows for some additional flexibility with modeling.

Another way in which the edge extension might arise naturally is the following. Imagine that the path $A \to C$ was actually going through some unmeasured variable W, as shown in Figure 14(b). If we considered W as resolving, keeping the effect of A on C captures what we want to achieve with our adaptation.

For every variable V we need to define its *adaptation parent set*, $\text{aps}(V) \subset \text{pa}(V)$, which is the subset of parents of V that mediate unwanted bias coming from A. For a resolving variable R, $\text{aps}(R) = \emptyset$. All the effect of A to R going through $\text{pa}(R)$ is considered to be ok. For a non-resolving variable X we have that $\text{aps}(X) = \text{pa}(X)$, that is all the effect of A on X going through $\text{pa}(X)$ is seen as unfair. In the example from Figure 14(a) we have would have $\text{aps}(C) = P$. This would mean, as described above, that we wish to remove the $A \to P \to C$ effect (since policing is a form of bias), while at the same time keeping the direct $A \to C$ effect (which might be seen as permissible).

The main difference from the original version is in line 6 of Algorithm 1, in which we assign the transformed value as

$$\text{FP}(V_k) \leftarrow g_V(U_k,\ \text{FP}(\text{pa}(V_k))). \tag{18}$$

In the edge specific case, instead of using transformed values of all the parents $\text{pa}(V)$ in the assignment (18), we use the original values of parents in $\text{pa}(V) \setminus \text{aps}(V)$ and the transformed values $\text{FP}(\text{aps}(V))$ of the parents in $\text{aps}(V)$.

Appendix F. UCI Adult dataset

We give more details about how we preprocessed the UCI Adult dataset. The preliminary cleaning of the dataset is similar to that of Zhu (2016). In particular, the following operations on the features are performed:

- variables “relationship”, “final weight”, “education” (categorical), “capital gain” and “capital loss” were removed
- levels of variable “work class” were merged, so that we obtain four different levels - Government, Self-Employed, Private and Other/Unknown
- levels of the variable “marital status” were merged so that we obtain two levels - Married and Not-Married
- levels of variable “native country” were merged so that we obtain two levels - US and Non-US

Categorical variables that are descendants of gender A were given an ordering, so that the probability of success $\mathbb{P}(Y = 1 \mid F = f)$ is marginally increasing in levels of F. This is described more precisely in Section 5.1.

From Figure 5 we see that females in the dataset are much more likely to be in their early twenties and are also more likely to be black than males. Since we do believe that additional edges between A and C are present only due to sampling, we propose a subsampling method to resolve the problem and obtain a dataset for which the causal graph in Figure 4(a) is

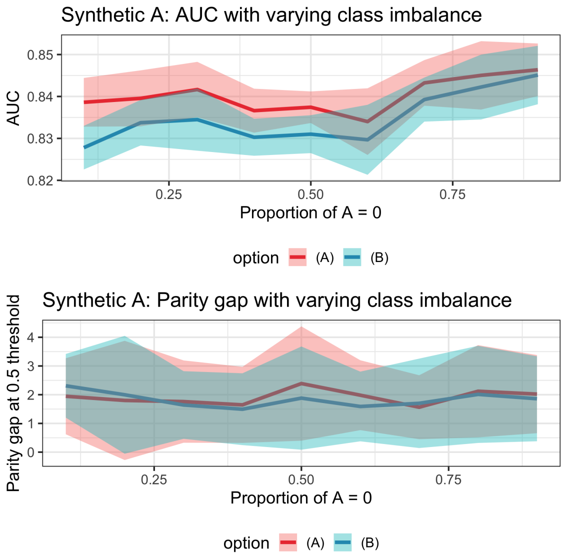


Figure 15: AUC and parity gap at the 0.5 threshold for example A and training options (A) and (B) with varying class imbalance. Confidence bounds represent standard deviations of values obtained from 10 repeats.

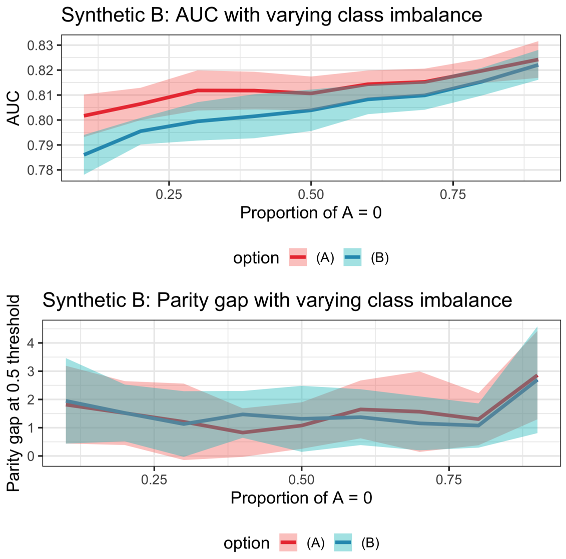


Figure 16: AUC and parity gap at the 0.5 threshold for example B and training options (A) and (B) with varying class imbalance. Confidence bounds represent standard deviations of values obtained from 10 repeats.

valid. In particular, we take only the white subpopulation. Since there are strictly more males than females for every age value, we subsample the males randomly so that we achieve exact matching in the age distributions between genders. In this way, we avoid the problem of biased sampling. The dataset still consists of 26052 individuals, which is a sufficient amount of data.

Appendix G. Empirical comparison of training options (A) and (B)

We show the empirical performance of training methods (A) and (B) (described in Section 5.4) on the two synthetic examples used in Section 6. We focus only on the case of no resolving variables ($R = \emptyset$) and we vary the class imbalance (that is we vary the proportion p_0 of the $A = 0$ instances in the data, $p_0 \in \{0.1, 0.2, \dots, 0.9\}$). For each value of p_0 we generate 5000 training and 5000 testing samples and run both methods (A) and (B) (this process is repeated 10 times). We measure the AUC and the parity gap at the 0.5 threshold. The results are shown in Figures 15 and 16. The confidence bounds represent the standard deviations of the values obtained from the 10 repeats. We note that both training methods in this case (and a variety of other cases) exhibit fairly similar performance. Therefore, we do not explicitly recommend using one or the other. For a specific problem, the better of the two can be chosen via cross-validation.

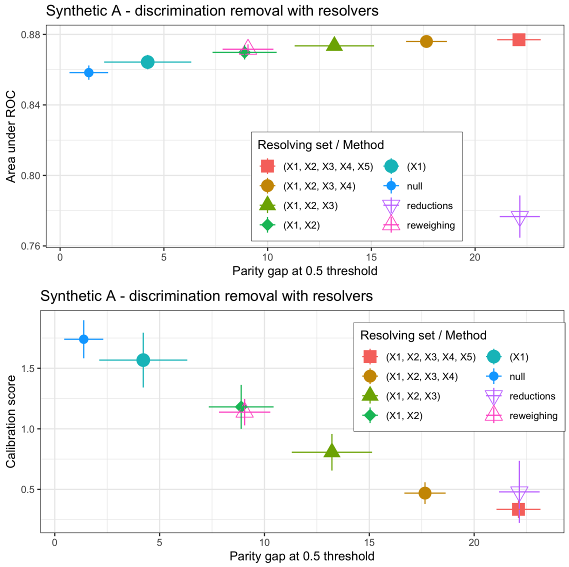


Figure 17: AUC-parity and parity-calibration score trade-off for example A. Vertical bars represent standard deviations obtained from 10 repeats.

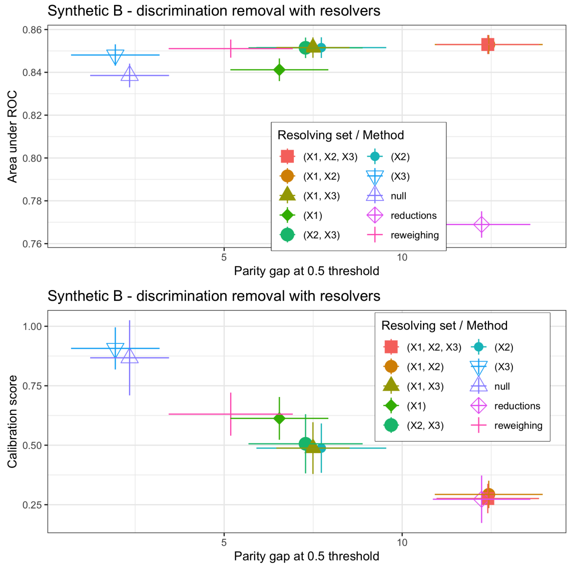


Figure 18: AUC-parity and parity-calibration score trade-off for example B. Vertical bars represent standard deviations obtained from 10 repeats.

Appendix H. Complete performance plots for the synthetic examples

In Section 6, for the synthetic examples, the fair reductions comparison method was removed to improve the readability of the plots. We provide the full plots in Figures 17 and 18.

Appendix I. Fairadapt and Path-Specific Counterfactual Fairness

We empirically compare `fairadapt` to path-specific counterfactual fairness (PSCF) of Chiappa (2019) on UCI Adult and COMPAS datasets. Dataset descriptions are given in Section 6.3 and the respective causal graphs are given in Figure 4. We provide a `PyTorch` implementation of the PSCF method, which can be found in our Github repository. We follow the implementation description of the authors as closely as possible. The PSCF method has an additional tuning parameter β which determines the degree of independence between the protected attribute A and the latent encoder embedding (for details we refer the reader to the original paper). We note that in our comparison, the PSCF method gives similar results over a range of parameters β (we note we are using a subsampled version of the UCI Adult dataset) and for large values of β the behavior becomes unstable.

For the UCI Adult dataset, we set the work-related variables R to be resolving (in the PSCF view, every $A \to Y$ path through R is fair and every path not through R is unfair). The accuracy and parity gap of different methods are given in Figure 19. Additionally, we include the normal Random Forest applied to the original data. The runtime of `fairadapt` on a single 2.8GHz CPU is 23 seconds, compared to 396 seconds for a single value of β for PSCF (excluding any hyperparameter search).

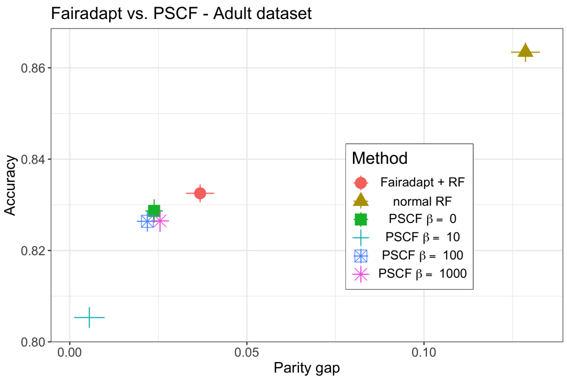


Figure 19: Accuracy and parity gap of different methods for the Adult dataset with R resolving. Vertical bars represent standard deviations obtained from 10 bootstrap samples of the testing set.

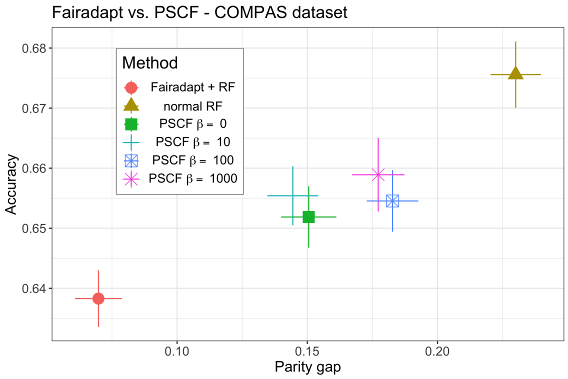


Figure 20: Accuracy and parity gap of different methods for the COMPAS dataset with D resolving. Vertical bars represent standard deviations obtained from 10 bootstrap samples of the testing set.

For the COMPAS dataset, we set the charge degree D to be resolving. The accuracy and parity gap of different methods are given in Figure 20. The runtime of `fairadapt` on a single 2.8GHz CPU is 8 seconds, compared to 102 seconds for a single value of β for PSCF.

In terms of performance metrics, neither of the methods outperforms the other in both accuracy and fairness. Both methods eliminate discrimination at the small expense of accuracy and provide slightly different, but competitive results.

Appendix J. NPSEM transformation

Suppose we start with a non-parametric structural equation model (NPSEM) defined as

$$Z_k = g_k(Z_{\mathrm{pa_k}}, U_k), \tag{19}$$

Fix the value of $Z_{\mathrm{pa_k}} = z$. Define $g^\star$ as $g^\star(u) = g_k(z, u)$. Let U be a random variable which has the same distribution as U_k. Cumulative distribution of $g^\star$ is defined as:

$$F(x) = \mathbb{P}(g^\star(U)) \leq x).$$

The function $F : \mathbb{R} \to [0, 1]$ is invertible since the variables Z_k are assumed to be continuous. Therefore, we can write

$$g^\star(U) = F^{-1}(F(g^\star(U))).$$

The argument $F(g^\star(U))$ within the function F^{-1} satisfies:

$$\begin{aligned}\mathbb{P}(F(g^\star(U)) \leq u) &= \mathbb{P}(g^\star(U) \leq F^{-1}(u)) \\ &= F(F^{-1}(u)) = u,\end{aligned}$$

showing that $\widetilde{U} = F(g^\star(U))$ has a uniform $U[0, 1]$ distribution. We further define $\widetilde{g}(u) = F^{-1}(u)$. Note that writing $Z_k = \widetilde{g}(\widetilde{U})$ gives the same assignment equation as (19) for

$Z_{\mathrm{pa_k}} = z$. The only difference is that multiple values (but always a set of measure 0) of U_k can correspond to a single value of $\widetilde{U}$.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404, 1975.

Francine D Blau and Lawrence M Kahn. Understanding international differences in the gender pay gap. *Journal of Labor economics*, 21(1):106–144, 2003.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

Alex J Cannon. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic environmental research and risk assessment*, 32(11):3207–3225, 2018.

Guillaume Carlier, Victor Chernozhukov, Alfred Galichon, et al. Vector quantile regression: an optimal transport approach. *The Annals of Statistics*, 44(3):1165–1192, 2016.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

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

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Juan A Cuesta-Albertos, L Ruschendorf, and Araceli Tuerodiaz. Optimal coupling of multivariate distributions and stochastic processes. *Journal of Multivariate Analysis*, 46(2): 335–361, 1993.

Richard B Darlington. Another look at "cultural fairness" 1. *Journal of Educational Measurement*, 8(2):71–82, 1971.

A Philip Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, 2000.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248, 2019.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

Roger Koenker. quantreg: Quantile regression. r package version 5.05. *R Foundation for Statistical Computing: Vienna) Available at: http://CRAN. R-project. org/package= quantreg*, 2013.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

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

Moshe Lichman et al. Uci machine learning repository. `https://archive.ics.uci.edu/ml`, 2013.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.

James D Malley, Jochen Kruppa, Abhijit Dasgupta, Karen G Malley, and Andreas Ziegler. Probability machines. *Methods of information in medicine*, 51(01):74–81, 2012.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Judea Pearl. The logic of counterfactuals in causal inference. *Journal of the American Statistical Association*, 95(450):428–435, 2000.

Judea Pearl. *Causality.* Cambridge University press, 2009.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128 (30):2013, 2013.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55: 58–63, 2015.

Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science*, 37(6):1011–1035, 2013.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002.

Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, pages 3675–3685, 2018a.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018b.

Haojun Zhu. Predicting earning potential using the adult dataset. `https://rpubs.com/H_Zhu/235617`, 2016.