Tuning Fairness with Quantities that Matter

Paper ID: 8655

Abstract

The issue of fairness in machine learning models has recently attracted a lot of attention as ensuring it will ensure continued confidence of the general public in the deployment of machine learning systems. Here, we focus on mitigating the harm incurred by a biased system that offers better outputs (e.g. loans, jobs) for certain groups than for others. We show that bias in the output can naturally be handled in probabilistic models by introducing a latent target output. This formulation has several advantages: first, it is a unified framework for several notions of fairness such as Demographic Parity and Equalized Odds; second, it is expressed as a marginalization instead of a constrained problem; and third, it allows the encoding of our knowledge of what the bias in the outputs should be. Practically, the second allows us to reuse off-the-shelf toolboxes, and the latter translates to the ability to control the level of fairness by directly varying fairness target rates such as true positive rates and positive rates. In contrast, existing approaches rely on intermediate, arguably unintuitive, control parameters such as covariance thresholds.

1 Introduction

Algorithmic assessment methods are used for predicting human outcomes such as bail decision and mortgage approval. This contributes, in theory, to a world with decreasing human biases. To achieve this, however, we need advanced machine learning models that are free of algorithmic biases (fair models), despite the fact that they are *written* by humans and trained based on historical and *biased data*.

There is no single accepted definition of algorithmic fairness for automated decision-making though several have been proposed. One definition is referred to as *statistical* or *demographic parity*. Given a binary sensitive attribute (married/unmarried) and a binary decision (yes/no to getting a mortgage), demographic parity requires equal positive rates (PR) across married and *un*married individuals, i.e. $\mathbb{P}(\text{mortgage} = \text{yes}|\text{married}) = \mathbb{P}(\text{mortgage} = \text{yes}|\text{not married})$. Another fairness criterion, *equalized odds* (Hardt et al. 2016), takes into account binary label (yes/no in making a payment), and requires equal true positive rates (TPR) and false positive rates (FPR) across married and *un*married groups, i.e. $\mathbb{P}(\text{mortgage} = \text{yes}|\text{married}, \text{payment} = \text{yes}) = \mathbb{P}(\text{mortgage} = \text{yes}|\text{not married}, \text{payment} = \text{yes})$ for equal TPR rates, and accordingly for the FPR rates.

Many models are available for enforcing demographic parity or equalized odds (Agarwal et al. 2018; Calders, Kamiran, and Pechenizkiy 2009; Kamishima et al. 2012; Zafar et al. 2017a; 2017b), however none of them give humans the control to set the *rate* of positive predictions (e.g. a PR of 0.6), or the rate of true positives (e.g. a TPR of 0.6). What is the advantage of being able to control PR/TPR/FPR rates? In this paper, we show that we can actually control the level of fairness by directly tuning those target rates. This means machine learning practitioners can trade off fairness and accuracy by directly controlling parameters that are intuitive and thereby understandable to the general public. In contrast, existing approaches to balancing accuracy and fairness rely on intermediate, unintuitive control parameters such as allowable constraint violation ϵ (e.g. 0.01) in Agarwal et al. (2018), or a covariance threshold c (e.g. 0 that is controlled by another parameters τ and μ – 0.005 and 1.2 – to trade off this threshold and accuracy) in Zafar et al. (2017a).

We propose a method for incorporating fairness into probabilistic classifiers. We assume the existence of *unbiased* output decision, which will modulate the likelihood term of the classifier. With this formulation, we can show the theoretical mutual exclusivity of demographic parity and equalized odds (Chouldechova 2017; Kleinberg, Mullainathan, and Raghavan 2016) as a by-product of the sum and product probability rules. This is in stark contrast to many existing approaches that embed fairness criteria as constraints in the optimization procedure (Donini et al. 2018; Quadrianto and Sharmanska 2017; Zafar et al. 2017a; 2017b); those methods can then violate mutual exclusivity as there is no mechanism to prevent multiple constraints being added. We instantiate our approach with a parametric logistic regression classifier and a Bayesian nonparametric Gaussian process classifier (GPC). For the latter, as our formulation is not expressed as a constrained problem, we can draw upon advancements in automated variational inference (Bonilla, Krauth, and Dez-

fouli 2016; Gardner et al. 2018; Krauth et al. 2016) for learning the fair model, and for handling large amounts of data.

The method presented in this paper is closely related to a number of previous works, e.g. Kamiran and Calders; Calders and Verwer (2012; 2010). Proper comparisons with them requires knowledge of our approach. We will thus explain our approach in the subsequent sections, and defer detailed comparisons to Section 4 Related Work.

2 Target labels for handling label bias

In order to motivate the notion of optimising for target labels, let us consider the recidivism prediction problem used to inform bail and parole decisions. Here, the goal is to predict whether a person, if released, will re-offense within a fixed period of time. We, however, do not have data on who commits crimes (target labels), we only have data on who is arrested (proxy labels). Arrest data could be a poor proxy for actual crimes committed if over-policing in a certain group of individuals happens.

Suppose then we want to enforce the fairness criterion demographic parity. In demographic parity, we demand that the overall probability of being assigned a positive prediction ($\hat{y} = 1$) is the same for all demographic groups s (here with $s \in \{0, 1\}$): $\mathbb{P}(\hat{y} = 1|s = 0) = \mathbb{P}(\hat{y} = 1|s = 1)$. The above equation does not (in general) hold for the labels in the training set. Enforcing demographic parity can be understood as learning from a dataset with "incorrect" or proxy labels (Chouldechova and Roth 2018; Tolan 2019). The fair classifier — with respect to demographic parity — makes predictions that are distributed differently than the labels in the dataset. From the perspective of the fair classifier, the training labels are "wrong", because they are biased.

Furthermore, we argue that for *classification*, it is not as useful to consider any bias in the features and we will therefore only consider the bias in the labels. This is because when building a fair classifier it is not necessary to explicitly construct a fair version of the features, all that matters in the end is the fair prediction. If the classifier learns from *unbiased* labels, it will find an implicit representation of the features that can predict these unbiased labels. There are, however, situations where an explicit fair representation of the features is necessary; e.g. when wanting to sell the data (Madras et al. 2018).

Inspired by this point of view, we introduce a new variable to represent the unbiased labels (or "true" labels): $\bar{y}$. We call these the *target labels*. The target labels are unknown but using the prior knowledge about what the distribution of the target labels, we can establish a relationship between the training labels and the target labels.

Our framework can be used with any likelihood-based model, such as Logistic Regression of Gaussian Process models. The loss for these models typically contains a term for the negative log-likelihood, e.g. the negative Bernoulli log-likelihood for binary classification problems, or the negative Categorical log-likelihood for the multi-class setting. These aspects of the loss remain the same, but the computation of the likelihood changes. The goal is for the model to learn to predict the target labels instead of the training labels. The likelihood with respect to the target label $\mathbb{P}(\bar{y}|x, \theta)$, where θ represents the model parameters, can be readily computed in the usual way, but the concrete $\bar{y}$ is unknown, so we have to express it in terms of y.

Focusing on the binary case, the training label likelihood is $\mathbb{P}(y = 1|x, s, \theta)$ for the fairness-aware case, where the prediction is allowed to depend on s. We expand this likelihood in terms of $\bar{y}$:

$$\begin{aligned}\mathbb{P}(y = 1|x, s, \theta) &= \sum_{\bar{y}\in\{0,1\}} \mathbb{P}(y = 1, \bar{y}|x, s, \theta) \\ &= \sum_{\bar{y}\in\{0,1\}} \mathbb{P}(y = 1|\bar{y}, x, s, \theta)\, \mathbb{P}(\bar{y}|x, s, \theta)\ . \end{aligned} \tag{1}$$

The overall likelihood is still computed with respect to the training labels, but the model itself is asked to predict the likelihood of $\bar{y}$. Tying this together is the conditional probability $\mathbb{P}(y = 1|\bar{y}, x, s, \theta)$, which captures the relationship between $\bar{y}$ and y.

By considering the relations between all the variables more closely, we can make some simplifications to the stated probabilities. In particular, we are assuming that the corruption of $\bar{y}$ does not depend on x. This is a relatively strong assumption and we can consider a relaxation. In that case, we need a probability distribution over the high-dimensional space in which x lives, that tells us how likely $y = 1$ is, if $\bar{y} = 0$ or $\bar{y} = 1$. This distribution has to be prior knowledge as it cannot be learned from the data which inputs x experienced the highest discrimination. This prior knowledge relates to the weakly meritocratic notion of *individual fairness* (Joseph et al. 2016) which relies on the unknown labels of an individual as a measure of merit, and requires that individuals who have a higher probability of really having a positive label should have only a higher probability of being classified as positive. As we do not have this prior knowledge, we have to make the assumption that the corruption is independent of x.

We arrive at the following:

$$\mathbb{P}(y = 1|x, s, \theta) = \sum_{\bar{y}\in\{0,1\}} \mathbb{P}(y = 1|\bar{y}, s)\, \mathbb{P}(\bar{y}|x, \theta)\ . \tag{2}$$

At test time, we simply use $\mathbb{P}(\bar{y}|x, \theta)$ to make predictions, which does not depend on s. This is important in order to avoid *direct discrimination* (Barocas and Selbst 2016).

We refer to the parameters expressed by $\mathbb{P}(y = 1|\bar{y}, s)$ as *debiasing parameters*. We give an intuition for the meaning of these parameters and how to set their values in Section 3. For a binary sensitive attribute s (and binary label y), there are 4 debiasing parameters (see Algorithm 1 where $d^{s=j}_{\bar{y}=i} := \mathbb{P}(y = 1|\bar{y} = i, s = j)$):

$$\mathbb{P}(y = 1|\bar{y} = 0, s = 0), \quad \mathbb{P}(y = 1|\bar{y} = 1, s = 0) \tag{3}$$

$$\mathbb{P}(y = 1|\bar{y} = 0, s = 1), \quad \mathbb{P}(y = 1|\bar{y} = 1, s = 1)\ . \tag{4}$$

The above derivation applies to binary classification but can easily be extended to the multi-class case.

3 Realization of concrete fairness constraints

This section focuses on how to set values of the debiasing parameters for tuning a variety of fairness target rates.

Algorithm 1 Training loop with Target Labels

Input: Training set $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^N$, debiasing parameters $d_{\bar{y}=0}^{s=0}$, $d_{\bar{y}=1}^{s=0}$, $d_{\bar{y}=0}^{s=1}$, $d_{\bar{y}=1}^{s=1}$
Output: fair model parameters θ

Initialize θ (randomly)
for all x_i, y_i, s_i **do**
  $P_{\bar{y}=1} \leftarrow \bar{c}(x_i, \theta)$ (e.g. logistic($\langle x, \theta \rangle$))
  $P_{\bar{y}=0} \leftarrow 1 - P_{\bar{y}=1}$
  if $s_i = 0$ **then**
    $P_{y=1} \leftarrow d_{\bar{y}=0}^{s=0} \cdot P_{\bar{y}=0} + d_{\bar{y}=1}^{s=0} \cdot P_{\bar{y}=1}$
  else
    $P_{y=1} \leftarrow d_{\bar{y}=0}^{s=1} \cdot P_{\bar{y}=0} + d_{\bar{y}=1}^{s=1} \cdot P_{\bar{y}=1}$
  end if
  $\ell \leftarrow y_i \cdot P_{y=1} + (1 - y_i) \cdot (1 - P_{y=1})$
  update θ to maximize likelihood ℓ
end for

3.1 Meaning of the parameters

Before we consider concrete values, we give some intuition for the debiasing parameters. Let $s = 0$ refer to the disadvantaged group. For this group, we want to make more positive predictions than the dataset labels indicate. Variable $\bar{y}$ is supposed to be our fair label. Thus, in order to make more positive predictions, some of the $y = 0$ labels should be associated with $\bar{y} = 1$. However, we do not know which. So, if our model predicts $\bar{y} = 1$ (high $\mathbb{P}(\bar{y} = 1|x, \theta)$) while the dataset label is $y = 0$, then we allow for the possibility that this is actually correct. That is, $\mathbb{P}(y = 0|\bar{y} = 1, s = 0)$ is not 0. If we choose, for example, $\mathbb{P}(y = 0|\bar{y} = 1, s = 0) = 0.3$ then that means that 30% of positive virtual labels $\bar{y} = 1$ may correspond to negative dataset labels $y = 0$. This way we can have more $\bar{y} = 1$ than $y = 1$, overall. On the other hand, predicting $\bar{y} = 0$ when $y = 1$ holds, will always be deemed incorrect: $\mathbb{P}(y = 1|\bar{y} = 0, s = 0) = 0$; this is because we do not want any additional negative labels.

For the advantaged group $s = 1$, we have the exact opposite situation. If anything, we have too many positive labels. So, if our model predicts $\bar{y} = 0$ (high $\mathbb{P}(\bar{y} = 0|x, \theta)$) while the dataset label is $y = 1$, then we should again allow for the possibility that this is actually correct. That is, $\mathbb{P}(y = 1|\bar{y} = 0, s = 1)$ should not be 0. On the other hand, $\mathbb{P}(y = 0|\bar{y} = 1, s = 1)$ should be 0 because we do not want additional positive labels for $s = 1$. It could also be that the number of positive labels is exactly as it should be, in which case we can just set $y = \bar{y}$ for all data points with $s = 1$.

3.2 Demographic Parity

Demographic Parity is characterized by an independence of predictions $\bar{y}$ and the sensitive attribute s. Given that we have complete control over the *debiasing parameters*, we can ensure this independence by requiring $\mathbb{P}(\bar{y} = 1|s = 0) = \mathbb{P}(\bar{y} = 1|s = 1)$. Our fairness constraint is then that both of these probabilities are equal to the same value, which we will call the target rate PR_t ("PR" as *positive rate*):

$$\mathbb{P}(\bar{y} = 1|s = 0) \stackrel{!}{=} PR_t \quad \text{and}$$
$$\mathbb{P}(\bar{y} = 1|s = 1) \stackrel{!}{=} PR_t \,. \tag{5}$$

This leads us to the following constraints for $s' \in \{0, 1\}$:

$$\begin{aligned} PR_t &= \mathbb{P}(\bar{y} = 1|s = s') \\ &= \sum_y \mathbb{P}(\bar{y} = 1|y, s = s')\, \mathbb{P}(y|s = s'). \end{aligned} \tag{6}$$

We call $\mathbb{P}(y = 1|s = j)$ the base rate PR_b^j which we estimate from the training set:

$$\mathbb{P}(y = 1|s = i) = \frac{\text{number of points with } y = 1 \text{ in group } i}{\text{number of points in group } i} \,.$$

Expanding the sum, we get

$$\begin{aligned} PR_t = \mathbb{P}(\bar{y} = 1|y = 0, s = s') \cdot (1 - PR_b^1) + \\ + \mathbb{P}(\bar{y} = 1|y = 1, s = s') \cdot PR_b^1 \,. \end{aligned} \tag{7}$$

This is a system of linear equations consisting of two equations (one for each value of s') and four free variables: $\mathbb{P}(\bar{y} = 1|y, s)$ with $y, s \in \{0, 1\}$. The two unconstrained degrees of freedom determine how strongly the accuracy will be affected by the fairness constraint. If we set $\mathbb{P}(\bar{y} = 1|y = 1, s)$ to 0.5, then this expresses the fact that a train label y of 1 only implies a target label $\bar{y}$ of 1 in 50% of the cases. In order to minimize the effect on accuracy, we make $\mathbb{P}(\bar{y} = 1|y = 1, s)$ as high as possible and $\mathbb{P}(\bar{y} = 1|y = 0, s)$, conversely, as low as possible. However, the lowest and highest possible values are not always 0 and 1 respectively. To see this, we solve for $\mathbb{P}(\bar{y} = 1|y = 0, s = j)$ in Eq (7):

$$\begin{aligned} &\mathbb{P}(\bar{y} = 1|y = 0, s = j) \\ &= \frac{PR_b^j}{1 - PR_b^j} \left(\frac{PR_t}{PR_b^j} - \mathbb{P}(\bar{y} = 1|y = 1, s = j) \right) \,. \end{aligned} \tag{8}$$

If ${}^{PR_t}/_{PR_b^j}$ were greater than 1, then setting $\mathbb{P}(\bar{y} = 1|y = 0, s = j)$ to 0 would imply a $\mathbb{P}(\bar{y} = 1|y = 1, s = j)$ value greater than 1. A visualization that shows why this happens can be found in the Appendix. Algorithm 2 shows pseudocode of the whole procedure, including the computation of the allowed minimal and maximal value.

We may call the $\mathbb{P}(\bar{y} = 1|y = 1, s)$ ($s \in \{0, 1\}$) the *biased* TPRs. If these TPRs happen to be the same, then this enforces Equality of Opportunity as well; but in general this will not be the case.

Once all these probabilities have been found, the debiasing parameters are fully determined by applying Bayes' rule:

$$\mathbb{P}(y = 1|\bar{y}, s) = \frac{\mathbb{P}(\bar{y}|y = 1, s)\mathbb{P}(y = 1|s)}{\mathbb{P}(\bar{y}|s)} \tag{9}$$

Choosing a target rate. Having decided to enforce demographic parity, there is still considerable freedom in choosing the target rate $PR_t := \mathbb{P}(\bar{y} = 1)$. However, this choice affects the accuracy as well, as Theorem 1 and Corollary 1.1 in the following state.

Before we get to the theorem, we introduce some notation. We are given a dataset $\mathcal{D} = \{(x_i, y_i)\}_i$, where the x_i are vectors of features and the y_i the corresponding labels. We refer to the tuples (x, y) as the *samples* of the dataset. The number of samples is $N = |\mathcal{D}|$.

We assume binary labels ($y \in \{0, 1\}$) and thus can form the (disjoint) subsets $\mathcal{Y}^0$ and $\mathcal{Y}^1$ with

$$\mathcal{Y}^j = \{(x, y) \in \mathcal{D} | y = j\} \quad \text{with } j \in \{0, 1\} \ . \tag{10}$$

Furthermore, we associate each sample with a classification $\hat{y} \in \{0, 1\}$. The task of making the classification $\hat{y} = 0$ or $\hat{y} = 1$ can be understood as sorting each sample from $\mathcal{D}$ into one of two sets: $\mathcal{C}^0$ and $\mathcal{C}^1$, such that $\mathcal{C}^0 \cup \mathcal{C}^1 = \mathcal{D}$ and $\mathcal{C}^0 \cap \mathcal{C}^1 = \emptyset$.

We refer to the set $\mathcal{A} = (\mathcal{C}^0 \cap \mathcal{Y}^0) \cup (\mathcal{C}^1 \cap \mathcal{Y}^1)$ as the set of correct (or accurate) predictions. The *accuracy* is given by $acc = N^{-1} \cdot |\mathcal{A}|$.

Definition 1.

$$r_a := \frac{|\mathcal{Y}^1|}{|\mathcal{D}|} = \frac{|\mathcal{Y}^1|}{N} \tag{11}$$

is called the *base acceptance rate* of the dataset $\mathcal{D}$.

Definition 2.

$$\hat{r}_a = \frac{|\mathcal{C}^1|}{|\mathcal{D}|} = \frac{|\mathcal{C}^1|}{N} \tag{12}$$

is called the *predictive acceptance rate* of the predictions.

Theorem 1. *For a dataset with the base rate r_a and corresponding predictions with a predictive acceptance rate of $\hat{r}_a$, the accuracy is limited by*

$$acc \leq 1 - |\hat{r}_a - r_a| \ . \tag{13}$$

Corollary 1.1. *Given a dataset that consists of two subsets $\mathcal{S}_0$ and $\mathcal{S}_1$ ($\mathcal{D} = \mathcal{S}_0 \cup \mathcal{S}_1$) where p is the ratio of $|\mathcal{S}_0|$ to $|\mathcal{D}|$ and given corresponding acceptance rates r_a^0 and r_a^1 and predictions with target rates $\hat{r}_a^0$ and $\hat{r}_a^1$, the accuracy is limited by*

$$acc \leq 1 - p \cdot \left|\hat{r}_a^0 - r_a^0\right| - (1 - p) \cdot \left|\hat{r}_a^1 - r_a^1\right| \ . \tag{14}$$

The proofs are fairly straightforward and can be found in Appendix X.

Corollary 1.1 implies that in the common case where group $s = 0$ is disadvantaged ($r_a^0 < r_a^1$) and also underrepresented ($p < \frac{1}{2}$), the highest accuracy under demographic parity can be achieved at $PR_t = r_a^1$ with

$$acc \leq 1 - p \cdot \left(r_a^1 - r_a^0\right) \ . \tag{15}$$

However, this means willingly accepting a bad accuracy in the (smaller) subset $\mathcal{S}_0$ that is compensated by a very good accuracy in the (larger) subset $\mathcal{S}_1$. An decidedly "fairer" approach is to aim for the same accuracy in both subsets which is achieved by using the average of the base acceptance rates for the target rate. We compare the two choices (PR_t^{max} and PR_t^{avg}) in Section 5.

Algorithm 2 Targeting PR (e.g. demographic parity)

Input: target rate PR_t, biased acceptance rate PR_b^i
Output: debiasing parameter $d_{\bar{y}=j}^{s=i}$

if $PR_t > PR_b^i$ **then**
  $\mathbb{P}(\bar{y} = 1 | y = 1, s = i) \leftarrow 1$
else
  $\mathbb{P}(\bar{y} = 1 | y = 1, s = i) \leftarrow \frac{PR_t}{PR_b^i}$
end if
if j=0 **then**
  $\mathbb{P}(\bar{y} = 0 | y = 1, s = i) \leftarrow 1 - \mathbb{P}(\bar{y} = 1 | y = 1, s = i)$
  $d_{\bar{y}=0}^{s=i} \leftarrow \frac{\mathbb{P}(\bar{y}=0|y=1,s=i) \cdot PR_b^i}{1 - PR_t}$
else if j=1 **then**
  $d_{\bar{y}=1}^{s=i} \leftarrow \frac{\mathbb{P}(\bar{y}=1|y=1,s=i) \cdot PR_b^i}{PR_t}$
end if

3.3 Equality of Opportunity

Equality of opportunity — another fairness definition — enforces independence of the prediction and the sensitive attribute s conditional on $y = 1$. If we see $\bar{y}$ as the prediction (which is not entirely correct as discussed below), then this fairness constraint can be expressed as $\bar{y} \perp s \mid (y = 1)$. Similarly to before, we enforce this by setting $\mathbb{P}(\bar{y} = 1 | y = 1, s)$ for both $s = 0$ and $s = 1$ to the same target TPR: TPR_t.

We can do the same for the TNR in order to enforce equalized odds: $TNR_t = \mathbb{P}(\bar{y} = 0 | y = 0, s)$. Together with the base rate $\mathbb{P}(y|s)$, TPR and TNR already determine the debiasing parameters uniquely; $\mathbb{P}(\bar{y}|s)$ is obtained from the sum rule. Thus, the target rate $\mathbb{P}(\bar{y}|s)$ is not free to set, and the mutual exclusivity of demographic parity and equality of opportunity (Kleinberg, Mullainathan, and Raghavan 2016; Chouldechova 2017) naturally falls out of our framework.

At first glance, it seems desirable to set both the target TNR (true negative rate), TNR_t, and the target TPR, TPR_t, to 1, because any value lower than 1 necessarily reduces the accuracy (as the accuracy is a weighted average of TPR and TNR). However, when target TNRs and target TPRs are all set to 1, then the debiasing parameters are all either 1 or 0 in a way that is equivalent to $\bar{y} = y$ for all data-points, making the $\bar{y}$ pointless. This problem is connected to the assumption above that we can substitute $\hat{y}$ with $\bar{y}$. This is only true for a perfect predictor which predicts all labels correctly. Such a predictor would fulfill equality of opportunity even if trained on y. Generally, our predictors are not perfect, however, so they make some classification errors. What equality of opportunity demands is that this classification error is the same for all specified groups. We thus set the target TPR (TPR_t) to a lower value that is the same for all groups, and thereby purposefully sacrifice some accuracy to make the errors the same. The value of TPR_t should be such that the classifier can achieve this TPR for all groups, but not lower than that.

Choosing values for the TPR target rate (and the TNR target rate) is not as straightforward as in the case of targeting an acceptance rate because TPR and TNR are inextricably linked to the classifier that is used. We additionally found that the achieved TPR does not just depend on the target TPR, but also on the target TNR. More specifically, targeting

Table 1: Accuracy and fairness (with respect to *demographic parity*) for various methods on the Adult dataset. Fairness is defined as $PR_{s=0}/PR_{s=1}$ (a completely fair model would achieve a value of 1.0). Left: using **race** as the sensitive attribute. Right: using **gender** as the sensitive attribute. The mean and std of 10 repeated experiments.

Algorithm	Fair $\rightarrow 1.0 \leftarrow$	Accuracy $\uparrow$	Fair $\rightarrow 1.0 \leftarrow$	Accuracy $\uparrow$
GP	0.56 ± 0.02	0.853 ± 0.002	0.32 ± 0.03	0.854 ± 0.003
LR	0.57 ± 0.03	0.846 ± 0.003	0.33 ± 0.02	0.847 ± 0.002
SVM	0.61 ± 0.02	0.859 ± 0.002	0.26 ± 0.02	0.857 ± 0.003
FairGP (ours)	0.62 ± 0.03	0.853 ± 0.003	0.65 ± 0.04	0.846 ± 0.004
FairLR (ours)	0.73 ± 0.04	0.844 ± 0.003	0.67 ± 0.03	0.839 ± 0.003
ZafarAccuracy (Zafar et al. 2017b)	1.31 ± 0.36	0.800 ± 0.012	1.22 ± 0.40	0.793 ± 0.009
ZafarFairness (Zafar et al. 2017b)	0.58 ± 0.14	0.846 ± 0.003	0.35 ± 0.10	0.846 ± 0.003
Kamiran&Calders (2012)	0.60 ± 0.03	0.831 ± 0.003	0.64 ± 0.03	0.847 ± 0.003
Agarwal et al. (2018)	0.61 ± 0.06	0.847 ± 0.003	0.35 ± 0.03	0.847 ± 0.003

a lower TNR makes it easier to achieve a higher TPR. This is perhaps not surprising, as lowering the TNR will result in more positive predictions ($\hat{y} = 1$), which means that the general threshold for a positive predictions is lowered. This lowered threshold makes it more likely that a given false negative prediction will be flipped; i.e., becomes a true positive prediction. A decrease of false negatives coupled with an increase in true positives will increase the TPR.

TPR_t and TNR_t have to be tuned for each dataset and each machine learning algorithm separately, because these settings interact deeply with the internals of the algorithm. However, we can give some guidelines to finding good values for TPR_t and TNR_t. As mentioned, tweaking either value can achieve the desired fairness. However, if the prediction task is such that false negatives are especially undesirable then TPR_t should be held close to 1.0 and TNR_t should be lowered instead. The opposite holds if false positives are undesirable.

4 Related work

There are several ways to enforce fairness in machine learning models: as a pre-processing step (Kamiran and Calders 2012; Louizos et al. 2016; Lum and Johndrow 2016; Zemel et al. 2013; Chiappa 2019), as a post-processing step (Feldman et al. 2015; Hardt et al. 2016), or as a constraint during the learning phase (Calders, Kamiran, and Pechenizkiy 2009; Zafar et al. 2017a; 2017b; Donini et al. 2018; Dimitrakakis et al. 2019). Our method enforces fairness during the learning phase (an in-processing approach) but, unlike other approaches, we do not cast fair-learning as a *constrained* optimization problem. Constrained optimization requiress a customized procedure. In Goh et al. (2016), Zafar et al. (2017a), and (2017b), suitable majorization-minimization/convex-concave procedures (Lanckriet and Sriperumbudur 2009) were derived. Furthermore, such constrained optimization approaches may lead to more unstable training, and often yield classifiers with both worse accuracy and more unfair (Cotter et al. 2018).

The approaches most closely related to ours were given by Kamiran and Calders (2012) who present four pre-processing methods: *Supression*, *Massaging the dataset*, *Reweighing*, and *Sampling*. In our comparison we focus on methods 2, 3 and 4, because the first one simply removes sensitive attributes and those features that are highly correlated with them. All the methods given by Kamiran and Calders (2012) aim only at enforcing demographic parity.

Massaging the dataset approach uses a classifier to first rank all samples according to their probability of having a positive label ($y = 1$) and then flips the labels that are closest to the decision boundary such that the data then satisfies demographic parity. This *pre-processing* approach is similar in spirit to our *in-processing* method but differs in the execution. In our method (Section 3.2), "ranking" and classification happen in one step and labels are not flipped but assigned probabilities of them being wrong.

The reweighing method re-weights samples based on whether they belong to an overrepresented or underrepresented demographic group. The sampling approach is based on the same idea but works by re-sampling instead of re-weighting. The weights are determined by a completely different procedure than ours. The prior work defines the weights based on observed proxy labels, while our method defines the weights (called debiasing parameters) based on latent target labels.

One approach in Calders and Verwer (2010) is also worth mentioning. It is based on a *generative* Naïve Bayes model in which a latent variable L is introduced which is reminiscent to our target label $\bar{y}$. We provide a *discriminative* version of this approach. In discriminative models, parameters capture the conditional relationship of an output given an input, while in generative models, the joint distribution of input-output is parameterized. With this conditional relationship formulation ($\mathbb{P}(y|\bar{y}, s) = {}^{\mathbb{P}(\bar{y}|y,s)\mathbb{P}(y|s)}/_{\mathbb{P}(\bar{y}|s)}$), we can show the theoretical mutual exclusivity of *demographic parity* and *equalized odds*, and we can have detailed control in setting the target rate. Calders and Verwer (2010) focuses only on the demographic parity fairness metric.

5 Experiments

We compare the performance of our target-label model with other existing models based on two real-world datasets. These datasets have been previously considered in the fairness-aware machine learning literature.

Implementation. The proposed method is compatible with any likelihood-based algorithm. We consider both a nonparametric and a parametric model. The nonparametric model is a Gaussian process model, and Logistic regression is the parametric counterpart. Since our fairness approach is not being framed as a constrained optimization problem, we can reuse off-the-shelf toolboxes including the GPyTorch library by Gardner et al. (2018) for Gaussian process models. This library incorporates recent advances in scalable variational inference including variational *inducing inputs* and likelihood ratio/REINFORCE estimators. The variational posterior can be derived from the likelihood and the prior. We need just need to modify the likelihood to take into account the target labels (Algorithm 1).

Data. The first dataset is the **Adult Income** dataset (Dheeru and Karra Taniskidou 2017). It contains 33,561 data points with census information from US citizens. The labels indicate whether the individual earns more ($y = 1$) or less ($y = 0$) than \$50,000 per year. We use the dataset with *race* and *gender* as the sensitive attribute. The input dimension, excluding the sensitive attributes, is 12 in the raw data; the categorical features are then one-hot encoded. For the experiments, we removed 2,399 instances with missing data and used only the training data, which we split randomly for each trial run. The second dataset is the **ProPublica recidivism** dataset. It contains data from 6,167 individuals that were arrested. The data was collected for the COMPAS risk assessment tool (Angwin et al. 2016). The task is to predict whether the person was rearrested within two years ($y = 1$ if they were rearrested, $y = 0$ otherwise). We again use the dataset with *race* and *gender* as the sensitive attributes.

Method. We evaluate two versions of our target label model[1]: *FairGP*, which is based on Gaussian Process models, and *FairLR*, which is based on logistic regression. We also train baseline models that do not take fairness into account.

The fair GP models and the baseline GP model are all based on variational inference and use the same settings. During training, each batch is equivalent to the whole dataset. The number of inducing inputs is 500 on the ProPublica dataset and 2500 on the Adult dataset which corresponds to approximately $1/8$ of the number of training points for each dataset. We use a squared-exponential (SE) kernel with automatic relevance determination (ARD) and the probit function as the likelihood function. We optimize the hyperparameters and the variational parameters with the Adam method (Kingma and Ba 2015) with the default parameters. We use the full covariance matrix for the Gaussian variational distribution. The logistic regression is trained with RAdam (Liu et al. 2019) and uses L2 regularization. For the regularization coefficient, we conducted a hyperparameter search over 10 folds of the data. For each fold, we picked the hyperparameter which achieved the best fairness among those 5 with the best accuracy scores. We then averaged over the 10 hyperparameter values chosen in this way and then used this average for all runs to obtain our final results.

[1]The code is available in the supplementary folder.

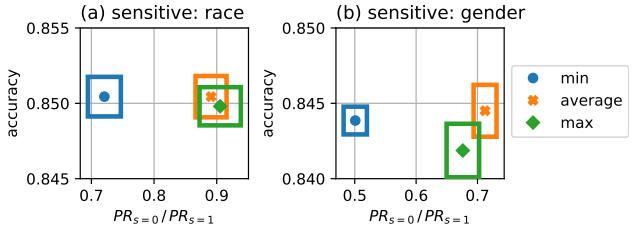

Figure 1: Accuracy and fairness (demographic parity) for various target choices. (a): Adult dataset using race as the sensitive attribute; (b): Adult dataset using gender. Center of the box is the mean; height and width of the box encode half of standard derivation of accuracy and disparate impact.

In addition to the GP and LR baselines, we compare our proposed model with the following methods: Support Vector Machine (*SVM*), *Kamiran & Calders* (Kamiran and Calders 2012) ("reweighing" method), *Agarwal et al.* (Agarwal et al. 2018) (using logistic regression as the classifier) and several methods given by Zafar et al. (Zafar et al. 2017b; 2017a), which include maximizing accuracy under demographic parity fairness constraints (*ZafarFairness*), maximizing demographic parity fairness under accuracy constraints (*ZafarAccuracy*), and removing disparate mistreatment by constraining the false negative rate (*ZafarEqOpp*). Every method is evaluated over 10 repeats that each have different splits of the training and test set.

Results for Demographic Parity on Adult dataset. Following Zafar et al. (Zafar et al. 2017b) we evaluate demographic parity on the Adult dataset. Table 1 shows the accuracy and fairness for several algorithms. In the table, and in the following, we use $PR_{s=i}$ to denote the observed rate of positive predictions per demographic group $\mathbb{P}(\hat{y} = 1|s = i)$. Thus, $PR_{s=0}/PR_{s=1}$ is a measure for demographic parity where a completely fair model would attain a value of 1.0. This measure for demographic parity is also called "disparate impact" (see e.g. Feldman et al.; Zafar et al. (2015; 2017a)). As the results in Table 1 show, FairGP is clearly fairer than the baseline GP. We use the mean (PR_t^{avg}) for the target acceptance rate. In Fig. 1, we investigate which choice of target (PR_t^{avg}, PR_t^{min} or PR_t^{max}) gives the best result. We use PR_t^{avg} for all following experiments as this is the fairest choice (cf. Section 3.2). The Fig.1(a) shows results from Adult dataset with *race* as sensitive attribute where we have $PR_t^{min} = 0.156$, $PR_t^{max} = 0.267$ and $PR_t^{avg} = 0.211$. PR_t^{avg} performs equal (for *race* as sensitive attribute) or better (for *gender*) compared with the two other possibilities. *ZafarAccuracy* can achieve good fairness results at the cost of accuracy. The results of FairGP are characterized by high fairness and high accuracy. FairLR achieves similar results to FairGP, but with generally slightly lower accuracy but better fairness. We used the two step procedure of Donini et al. (Donini et al. 2018) to verify that we cannot achieve the same fairness result with just parameter search on LR.

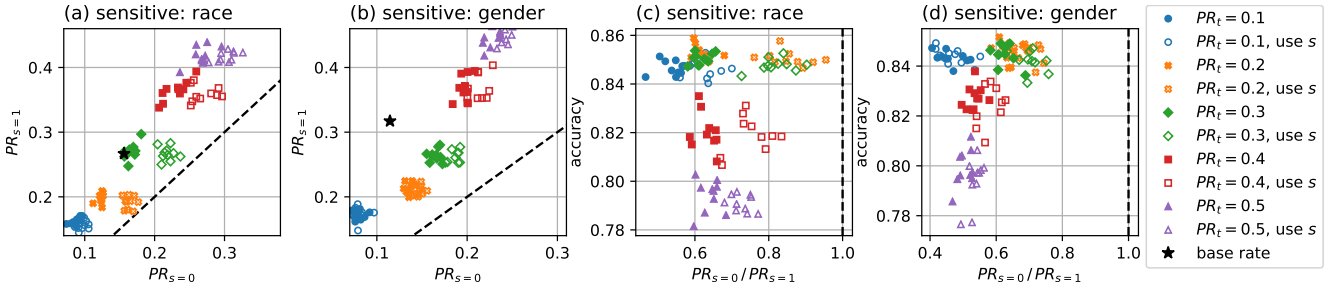

Figure 2: Predictions with different target acceptance rates (demographic parity) on Adult dataset. **(a)** and **(b)**: $PR_{s=0}$ vs $PR_{s=1}$. **(c)** and **(d)**: $PR_{s=0}/PR_{s=1}$ vs accuracy. Left column: using race as the sensitive attribute; Right column: using gender. The *base rate* indicates the positive rates of the training data. "use s" indicates that s was appended to the input features.

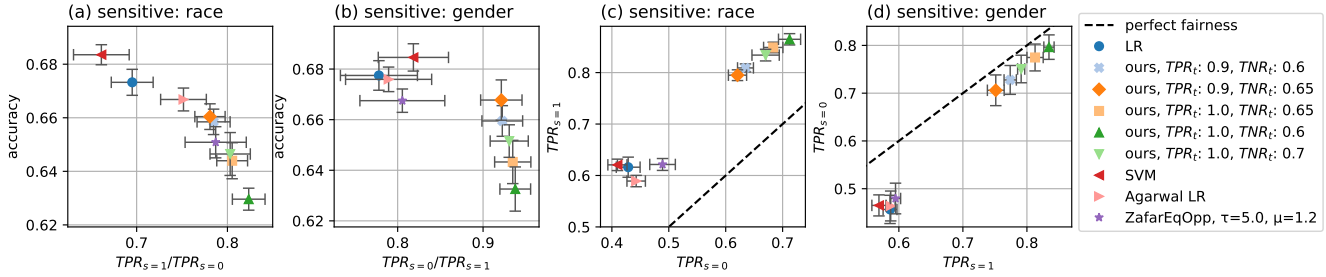

Figure 3: Predictions with different target true positive rates (TPR_t; equality of opportunity) on ProPublica dataset. Our results were obtained with the Logistic Regression model. **(a)** and **(b)**: $TPR_{s=0}/TPR_{s=1}$ vs accuracy. **(c)** and **(d)**: $TPR_{s=0}$ vs $TPR_{s=1}$. Left column: using race as the sensitive attribute; Right column: using gender. s is *not* used as an input feature.

Fig. 2(a) and (b) show runs of FairGP where we explicitly set a target acceptance rate, $PR_t := \mathbb{P}(\bar{y} = 1)$, instead of taking the mean PR_t^{avg}. A perfect targeting mechanism would produce a diagonal. The data points are not exactly on the diagonal but they show that setting the target rate has the expected effect on the observed acceptance rate. This tuning of the target rate is the unique aspect of the approach. This would be very difficult to achieve with existing fairness methods; a new constraint would have to be added. Fig. 2(c) and (d) show the same data as Fig. 2(a) and (b) but with different axes. It can be seen from from this Fig. 2(a) and (b) that the target acceptance rate can be used to *control* the trade-off between accuracy and fairness. In this specific case, changing the target rate barely affects fairness and it only affects the accuracy because target acceptance rates that are different from the base acceptance rate necessarily lead to "missclassifications".

Results for Equality of Opportunity on ProPublica dataset. For equality of opportunity, we again follow Zafar et al. (Zafar et al. 2017a) and evaluate the algorithm on the ProPublica dataset. As we did for demographic parity, we define a measure of equality of opportunity via the ratio of the true positive rates (TPRs) within the demographic groups. We use $TPR_{s=i}$ to denote the observed TPR in group i: $\mathbb{P}(\hat{y} = 1|y = 1, s = i)$, and $TNR_{s=i}$ for the observed true negative rate (TNR) in the same manner. The measure is then given by $TPR_{s=0}/TPR_{s=1}$. A perfectly fair algorithm would achieve 1.0 on the measure.

In order to demonstrate the *tuning* aspect of our proposed framework, we set different target TPRs and TNRs. By setting a low target TNR we can achieve very high TPRs. The results of 10 runs are shown in Fig. 3. Fig. 3(a) and (b) show the accuracy-fairness trade-off; (c) and (d) show the achieved TPRs. The latter two plots make clear that the TPR ratio does not tell the whole story: the realization of the fairness constraint can differ substantially. Tuning these hidden aspects of fairness is the strength of our method.

6 Discussion and conclusion

We have developed a machine learning framework which allows us to set a *target rate* for a variety of fairness notions, including demographic parity and equality of opportunity. For example, we can set the target true positive rate for equality of opportunity to be 0.6 for different groups. This capability is unique to our approach and can be used as an intuitive mechanism to control the trade-off between fairness and accuracy. In contrast to other methods which rely on unintuitive parameters, such as covariance thresholds, to enforce fairness, our method enables more control over tuning fairness, by introducing control parameters that are understandable to the general public: true positive rates, false positive rates, and positive rates. Our framework is general and will be applicable for sensitive variables with binary and multi-level values. The current work focuses on a single binary sensitive variable. Future work could extend our tuning approach to other fairness concepts like the closely related predictive parity group fairness (Chouldechova 2017) or individual fairness (Dwork et al. 2012).

References

Agarwal, A.; Beygelzimer, A.; Dudik, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *ICML*, volume 80, 60–69.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. *ProPublica* 23.

Barocas, S., and Selbst, A. D. 2016. Big data's disparate impact. *California Law Review* 104:671–732.

Bonilla, E. V.; Krauth, K.; and Dezfouli, A. 2016. Generic inference in latent gaussian process models. *arXiv preprint arXiv:1609.00577*.

Calders, T., and Verwer, S. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21(2):277–292.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *ICDM Workshops*, 13–18. IEEE.

Chiappa, S. 2019. Path-specific counterfactual fairness. In *AAAI*.

Chouldechova, A., and Roth, A. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

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

Cotter, A.; Jiang, H.; Wang, S.; Narayan, T.; Gupta, M. R.; You, S.; and Sridharan, K. 2018. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *arXiv preprint arXiv:1809.04198*.

Dheeru, D., and Karra Taniskidou, E. 2017. UCI machine learning repository.

Dimitrakakis, C.; Liu, Y.; Parkes, D. C.; and Radanovic, G. 2019. Bayesian fairness. In *AAAI*.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J. S.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. In *NeurIPS*, 2796–2806.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *ITCS*, 214–226.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *KDD*, 259–268. ACM.

Gardner, J. R.; Pleiss, G.; Bindel, D.; Weinberger, K. Q.; and Wilson, A. G. 2018. GPyTorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. In *NeurIPS*, 7587–7597.

Goh, G.; Cotter, A.; Gupta, M.; and Friedlander, M. P. 2016. Satisfying real-world goals with dataset constraints. In *NIPS*, 2415–2423.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *NIPS*, 3315–3323.

Joseph, M.; Kearns, M.; Morgenstern, J. H.; and Roth, A. 2016. Fairness in learning: Classic and contextual bandits. In *NIPS*, 325–333.

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

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *ECML PKDD*, 35–50. Springer.

Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

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

Krauth, K.; Bonilla, E. V.; Cutajar, K.; and Filippone, M. 2016. AutoGP: Exploring the capabilities and limitations of Gaussian Process models. *arXiv preprint arXiv:1610.05392*.

Lanckriet, G. R., and Sriperumbudur, B. K. 2009. On the convergence of the concave-convex procedure. In *NIPS*, 1759–1767.

Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; and Han, J. 2019. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2016. The variational fair autoencoder. In *ICLR*.

Lum, K., and Johndrow, J. 2016. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. S. 2018. Learning adversarially fair and transferable representations. In *ICML*, 3381–3390.

Quadrianto, N., and Sharmanska, V. 2017. Recycling privileged learning and distribution matching for fairness. In *NIPS*, 677–688.

Tolan, S. 2019. Fair and unbiased algorithmic decision making: Current state and future challenges. *arXiv preprint arXiv:1901.04730*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, 1171–1180.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017b. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 962–970.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *ICML*, 325–333.