FairRR: Pre-Processing for Group Fairness through Randomized Response

Xianli Zeng*
xlzeng@wharton.upenn.edu
NUS (Chongqing) Research Institute

Joshua Ward*
joshuaward@ucla.edu
University of California
Los Angeles

Guang Cheng
guangcheng@ucla.edu
University of California
Los Angeles

Abstract

The increasing usage of machine learning models in consequential decision-making processes has spurred research into the fairness of these systems. While significant work has been done to study group fairness in the in-processing and post-processing setting, there has been little that theoretically connects these results to the pre-processing domain. This paper proposes that achieving group fairness in downstream models can be formulated as finding the optimal design matrix in which to modify a response variable in a Randomized Response framework. We show that measures of group fairness can be directly controlled for with optimal model utility, proposing a pre-processing algorithm called FairRR [1] that yields excellent downstream model utility and fairness.

1 INTRODUCTION

As the use of machine learning models becomes increasingly prevalent in decision-making processes, concerns about the fairness of algorithms have become more pressing. Case studies from various domains such as criminal justice, healthcare, and employment (Flores et al. [2016], Corbett-Davies et al. [2023], Angwin et al. [2016], Tolan et al. [2019]), have demonstrated that biased algorithms can perpetuate or even amplify discrimination against individuals and groups. In response, a variety of approaches have been developed to ensure fairness focusing on the pre-processing of data, the in-processing of models, or the post-processing of model predictions (Zemel et al. [2013], Louizos et al. [2016], Calmon et al. [2017], Xu et al. [2019a], Celis and Keswani [2019], Cotter et al. [2019], Madras et al. [2018], Creager et al. [2019], Johndrow and Lum [2019], Cho et al. [2020], Zeng et al. [2024]). This litany of methods extends across many metrics of fairness that can roughly be broken into two groups: group fairness (Calders et al. [2009], Dwork et al. [2012], Hardt et al. [2016]) where fairness is defined as ensuring various types of statistical parity across distinct protected groups, and individual fairness (Joseph et al. [2016], Lahoti et al. [2019], Ruoss et al. [2020]) which aims to provide nondiscriminatory predictions for similar individuals.

In this paper, we focus on common group fairness criteria, including demographic parity (Calders et al. [2009], Kamishima et al. [2012], Cho et al. [2020]), equality of opportunity (Hardt et al. [2016], Zhang et al. [2018], Cho et al. [2020]), and predictive equality (Corbett-Davies et al. [2017]) adding to a larger family of diverse pre-processing methods in the supervised classification setting. In general, the goal of pre-processing is to modify the feature space of the original dataset such that when a classifier is trained on this altered data its output is fair. Strategies for this include transforming the data (Feldman et al. [2015], Lum and Johndrow [2016], Calmon et al. [2017], Johndrow and Lum [2019]), fair representation learning (Zemel et al. [2013], Louizos et al. [2016], Madras et al. [2018], Creager et al. [2019]) and fair generative models (Xu et al. [2018], Sattigeri et al. [2019], Xu et al. [2019b], Ramaswamy et al. [2021]). These methods are convenient to apply, as they do not change the training procedure and are generally independent to downstream modelling tasks, allowing for the use of most classifiers. However, they often do not allow for the control of the exact fairness level, they do not always have full

*These authors contributed equally to this work

[1]All code for FairRR with corresponding experiments can be found at: https://github.com/UCLA-Trustworthy-AI-Lab/FairRR

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

coverage of the variety of group fairness metrics in use, and they do not take advantage of recent results from the fair statistical learning literature.

Indeed, with such a variety of potential strategies to deploy in ensuring algorthims are fair, a robust body of literature has developed to answer the question of what exactly is the best theoretical classification strategy in terms of model utility and fairness. In the in-processing domain where fairness is achieved through the modification of a classifier itself, first Corbett-Davies et al. [2017] proved that, under several group fairness metrics, the fair Bayes-optimal classifiers are group-wise thresholding rules with unspecified thresholds. Menon and Williamson [2018] related demographic parity and equality of opportunity to cost-sensitive risks and derived fair Bayes-optimal classifiers under these two fairness measures. Under the setting of perfect demographic parity and equality of opportunity, exact forms of fair Bayes-optimal classifiers were derived in Chzhen et al. [2019] and Schreuder and Chzhen [2021], respectively. Finally, Zeng et al. [2024] and Zeng et al. [2022] showed that in the general case, fair Bayes-optimal classifiers could be derived for any level of disparity in most definitions of group fairness which is an advantage in the applied setting if some level of unfairness is allowed for better model utility. This paper extends this line of research to the pre-processing area where our goal is to develop a method that allows for an explicit level of control of disparity in training data and to extend these theoretical results to create a unified framework for adjusting for disparity at every step of machine learning model development.

Thus, we introduce the classic privacy technique Randomized Response (Warner [1965], Wang et al. [2016]) which privatizes a variable by 'flipping' its labels based on some probability. We propose that measures of group fairness and downstream model utility can be controlled by flipping the response variable in relation to a sensitive attribute. Here, preserving model utility can be thought of as minimizing the probability the label is flipped subject to a fairness constraint that seeks to flip labels to make a training set more fair. To derive this fairness constraint, we use fair group thresholding results from recent work on Fair Bayes-Optimal Classification Zeng et al. [2024, 2022] which allows for fairness to be exactly controlled for. Finally, we find the solution to these optimal flipping probabilities and perturb the response variable with the corresponding randomized response mechanism, finding that downstream models trained on this perturbed variable achieve good utility at various fairness settings.

Our contributions are thus summarized as follows:

- We show that a response variable can be made to satisfy many measures of group fairness at any disparity level, proposing a pre-processing method we call Fair Randomized Response (FairRR).
- We extend previous theoretical results from the in-processing to the pre-processing group fairness domain.
- We demonstrate that classifiers trained on modified data from FairRR demonstrate excellent utility and fairness results.

2 PRELIMINARIES

2.1 Fairness

To introduce fair algorithmic design, we consider credit lending as an example, where it is essential to ensure lending decisions are fair in order to comply with legal requirements. This can be formulated as a fair classification problem, where two types of features are observed for potential creditors: standard features $X \in \mathcal{X}$ such as income and education, and protected (or sensitive) features $A \in \mathcal{A}$ such as gender and race. The objective is to predict the label $Y \in \{0, 1\}$, if a creditor were to default on a loan, accurately and fairly with respect to A. Throughout this paper, we set the sensitive feature of $A = 1$ and $A = 0$ respectively be some privileged and the unprivileged groups. In this way, we can split the population into four parts: the positive privileged (PP) group ($A = 1$, $Y = 1$), the positive unprivileged (PN) group ($A = 0$, $Y = 1$), the negative privileged (NP) group ($A = 1$, $Y = 0$), and the negative unprivileged (NN) group ($A = 0$, $Y = 0$).

Researchers have proposed multiple group fairness measures for the fair classification setting. Generally, these measures depend on the constraints imposed on the joint distribution of A, Y, and a classifier's prediction $\widehat{Y}$. Common fairness measures include:

Definition 2.1 (Demographic Parity)**.** *A prediction $\widehat{Y}$ satisfies demographic parity if it achieves the same acceptance rate among protected groups:* $\mathbb{P}(\widehat{Y} = 1|A = 1) = \mathbb{P}(\widehat{Y} = 1|A = 0)$.

Definition 2.2 (Equality of Opportunity)**.** *A prediction $\widehat{Y}$ satisfies demographic parity if it achieves the same true positive rate among protected groups:* $\mathbb{P}(\widehat{Y} = 1|A = 1, Y = 1) = \mathbb{P}(\widehat{Y} = 1|A = 0, Y = 1)$.

Definition 2.3 (Predictive Equality)**.** *A prediction $\widehat{Y}$ satisfies predictive equality if it achieves the same false positive rate among protected groups:* $\mathbb{P}(\widehat{Y} = 1|A = 1, Y = 0) = \mathbb{P}(\widehat{Y} = 1|A = 0, Y = 0)$.

Essentially, these notions of fairness prohibit significant mistreatment of one group over another. When the

equalities holds in the aforementioned definitions, the fairness constraint enforces identical treatment among protected groups, referred to as perfect fairness.

In practice however, a relaxed or approximate versions of these notions could be preferred as perfect fairness may require a large sacrifice of accuracy or may not be possible. This means that instead of demanding identical treatment, we require that there should not be a significant difference in the model decisions between the two groups. Here, the disparity or unfairness of a classifier can be easily quantified by the difference between the groups. Specifically, we use DDP, DEO and DPE to measure the degree of violating demographic parity, equality of opportunity, predictive equality, respectively:

$$\begin{aligned}
\mathrm{DDP}(f) &= \mathbb{P}(\widehat{Y}=1|A=1) - \\
&\qquad \mathbb{P}(\widehat{Y}=1|A=0) \\
\mathrm{DEO}(f) &= \mathbb{P}(\widehat{Y}=1|A=1,Y=1) - \\
&\qquad \mathbb{P}(\widehat{Y}=1|A=0,Y=1) \\
\mathrm{DPE}(f) &= \mathbb{P}(\widehat{Y}=1|A=1,Y=0) - \\
&\qquad \mathbb{P}(\widehat{Y}=1|A=0,Y=0)
\end{aligned} \tag{1}$$

2.2 Fair Bayes Optimal Classifiers under Demographic Parity

In classification problems, the prediction $\widehat{Y}$ is often determined by a classifier f that indicates the probability of predicting $\widehat{Y}=1$ when observing $X=x$ and $A=a$. Specifically, a classifier is a measurable function $f:\mathcal{X}\times\{0,1\}\to[0,1]$ and $Y \mid X \sim \mathrm{Bern}(f(X))$, with $\mathrm{Bern}(p)$ the Bernoulli distribution with success probability p. We denote by $\widehat{Y}_f$ the prediction induced by the classifier f and we call f is fair if its induced prediction $\widehat{f}$ satisfies the fairness constraints. Among all fair classifiers, the Bayes optimal classifier serves as a critical theoretical benchmark, as it establishes the highest achievable accuracy for a given fairness constraint and serves as the theoretical objective that various algorithms aim to estimate. Throughout, we will use $\mathrm{D}(f)$ to denote some level of disparity from 1, depending on the context. We denote by $\mathcal{F}_\delta$ the set of measurable functions satisfying the δ-*parity* constraint

$$\mathcal{F}_\delta = \{f\in\mathcal{F} : |\mathrm{D}(f)|\le\delta\}.$$

A δ-fair Bayes-optimal classifier is defined as

$$f_\delta^\star \in \underset{f\in\mathcal{F}_\delta}{\arg\min}\, R(f) \text{ with } \quad R(f) := \mathbb{P}\left(Y\neq\widehat{Y}_f\right).$$

Zeng et al. [2024] and Zeng et al. [2022] studied the explicit form of fair Bayes-optimal classifiers. They found that, for many fairness metrics, the fair Bayes-optimal classifiers are group-wise thresholding rules with adjusted thresholds. Specifically, the standard Bayes-optimal classifiers $f^*:\mathcal{X}\times\{0,1\}\to[0,1]$ of the form: $f^*(x,a)=I\left(\eta_a(x)>1/2\right)$ can be modified to satisfy group fairness measures:

$$f_\delta^\star(x,a) = I\left(\eta_a(x) > \frac{1+(2a-1)T_a(t_\delta^\star)}{2}\right) \tag{2}$$

Here, $(x,a)\in\mathcal{X}\times\{0,1\}$, $\eta_a(x)=\mathbb{P}(Y=1|,A=a,X=x)$. $T_1(\cdot):\mathbb{R}\to[-1,1]$ and $T_0(\cdot):\mathbb{R}\to[-1,1]$ are two monotone non-decreasing functions with $T_1(0)=T_0(0)=0$ that are decided by the fairness metric and group-wise probabilities. In particular, with $p_{ay}=\mathbb{P}(A=a,Y=y),(a,y)\in\{0,1\}^2$, we have $T_a(t)=t/(p_{a1}+p_{a0})$ for demographic parity, $T_a(t)=t/[2p_{a1}-(2a-1)t]$ for equality of opportunity, and $T_a(t)=t/[2p_{a0}+(2a-1)t]$ for predictive equality.

The parameter $t_\delta^\star$ is decided by the disparity level δ where for a given t in a proper range, the classifier $f_t(x,a)=I\left(\eta_a(x)>\frac{1+(2a-1)T_a(t)}{2}\right)$ is a fair Bayes-optimal classifier for a certain disparity level δ_t. In particular, the disparity level $D(t)=D(f_t)$ is a monotone non-increasing function of t. In other words, $t_\delta^\star$ can be thought of as a term that balances the fairness-accuracy tradeoff of the fair Bayes-optimal classifier. Details on estimating $t_\delta^\star$ can be found in the next section, but in practice it can also be treated as a hyperparameter to control for disparity.

2.3 Design Matrices in Randomized Response

Randomized Response was first proposed by Warner [1965] to preserve the privacy of survey respondents' answers when asked sensitive questions and is a classic privacy technique. To start, suppose n individuals each have a response for some sensitive binary attribute Y, $y_i\in 0,1$. Each individual wishes to preserve the privacy of their response and so they send to an untrusted server a modified version of y_i in which the label is flipped to $\widetilde{y}_i$ by some probability. The probabilities in which y_i is flipped are determined by a design matrix which in the binary case can be written as:

$$\mathbf{P} = \begin{bmatrix} P(\widetilde{Y}=1|Y=1) & P(\widetilde{Y}=1|Y=0) \\ P(\widetilde{Y}=0|Y=1) & P(\widetilde{Y}=0|Y=0) \end{bmatrix} \tag{3}$$

To anonymize Y across a second binary variable $A\in\{0,1\}$, we can rewrite 3 to consist of separate design matrices:

$$\mathbf{P}_1 = \begin{bmatrix} P(\widetilde{Y}=1|A=1,Y=1) & P(\widetilde{Y}=1|A=1,Y=0) \\ P(\widetilde{Y}=0|A=1,Y=1) & P(\widetilde{Y}=0|A=1,Y=0) \end{bmatrix}$$

and

$$\mathbf{P}_0 = \begin{bmatrix} P(\widetilde{Y}=1|A=0,Y=1) & P(\widetilde{Y}=1|A=0,Y=0) \\ P(\widetilde{Y}=0|A=0,Y=1) & P(\widetilde{Y}=0|A=0,Y=0) \end{bmatrix}$$

Since the columns for each matrix must sum to 1, $\mathbf{P}_1$ and $\mathbf{P}_0$ can be expressed as the randomization mechanism $\mathcal{R}_{(\theta_{11},\theta_{10},\theta_{01},\theta_{00})}$ where $\theta_{ay} = \mathbb{P}(\widetilde{Y}=y|A=a,Y=y)$:

$$\mathbf{P}_1 = \begin{bmatrix} \theta_{11} & 1-\theta_{10} \\ 1-\theta_{11} & \theta_{10} \end{bmatrix}$$

and

$$\mathbf{P}_0 = \begin{bmatrix} \theta_{01} & 1-\theta_{00} \\ 1-\theta_{01} & \theta_{00} \end{bmatrix}$$

3 METHOD

3.1 Overview

We therefore have the preliminaries to begin developing a pre-processing method to perturb Y to be fair. Here, the goal is to find the randomization mechanism $\mathcal{R}_{(\theta_{11},\theta_{10},\theta_{01},\theta_{00})}$ that maximizes downstream model utility subject to fairness constraints. The design matrix for the best randomization mechanism can be easily found before then being applied to the training dataset. After this application, a final classifier can then be fit for the original X and now perturbed label variable $\widetilde{Y}$.

To start, we propose that the best $\mathcal{R}_{(\theta_{11},\theta_{10},\theta_{01},\theta_{00})}$ from solely a utility perspective would be the one that does not flip Y at all as it would not inject any noise into the training dataset. Thus, we wish to maximize $\mathbb{P}(\widetilde{Y}=Y)$ or:

$$\max p_{11}\theta_{11} + p_{10}\theta_{10} + p_{01}\theta_{01} + p_{00}\theta_{00}$$

where $\frac{1}{2} \leq \theta_{11},\theta_{10},\theta_{01},\theta_{00} \leq 1$. We will show that common group definitions of fairness can be written as linear equality constraints for this function.

3.2 Fairness through Randomized Response

3.2.1 Randomized Response and the Fair Bayes-Optimal Classifier

To illustrate how fairness can be achieved by randomized response, we first consider the Bayes-Optimal classifier with no protected attribute. Specifically, let $\mathcal{R}_{\theta_1,\theta_2}$ denote the randomization mechanism as follows: $\widetilde{Y} = \mathcal{R}_{\theta_1,\theta_0}(Y)$ has the property that $\mathbb{P}(\widetilde{Y}=Y) = \theta_1$ for $Y=1$ and θ_0 for $Y=0$ for any $(\theta_1,\theta_0) \in [1/2,1]\times[1/2,1]$. Note that this mechanism is imbalanced if $\theta_1 \neq \theta_0$. Denote by $\eta(x)$ and $\widetilde{\eta}(x)$ the conditional distribution of Y and $\widetilde{Y}$ given $X=x$, respectively. It can be verified that:

$$\widetilde{\eta}(x) = \theta_1\eta(x) + (1-\theta_2)(1-\eta(x)).$$

Clearly, $\widetilde{\eta}(x) > 1/2$ is equivalent to $\eta(x) > (\theta_0 - 1/2)/(\theta_1+\theta_0-1)$. Hence, $\mathcal{R}_{\theta_1,\theta_0}(Y)$ essentially shifts the thresholds of the decision rule when $\theta_1 \neq \theta_0$. As we discussed in section 2, the optimal fair classifiers are known to be group-wise thresholding rules for many fairness-metrics. The aforementioned technical connection enables us to generate a fair dataset through an im-balanced randomization of response.

Theorem 3.1. *Let (X,A,Y) follow a distribution $\mathbb{P}$ on $\mathcal{X}\times\{0,1\}\times\{0,1\}$. Consider a group-wise im-balanced randomized response mechanism $\widetilde{Y} = \mathcal{R}_{\theta_{11},\theta_{10},\theta_{01},\theta_{00}}(A,Y)$ with, for $a\in\{0,1\}$,*

$$\mathbb{P}(\widetilde{Y}=Y|A=a) = \begin{cases} \theta_{a1}, & \text{for } Y=1; \\ 1-\theta_{a0}, & \text{for } Y=0. \end{cases} \quad (4)$$

When the flipping probabilities satisfy:

$$(T_1(t^\star_\delta)+1)\theta_{11} + (T_1(t^\star_\delta)-1)\theta_{10} = T_1(t^\star_\delta);$$
$$(T_0(t^\star_\delta)-1)\theta_{01} + (T_0(t^\star_\delta)+1)\theta_{00} = T_0(t^\star_\delta);$$

where $T_1(\cdot)$, $T_0(\cdot)$ and $t^\star$ are the same as in (2). Denote $\widetilde{\mathbb{P}}$ as the joint distribution of $(X,A,\widetilde{Y})$. Then, the Bayes optimal classifier learned on $\widetilde{\mathbb{P}}$ is a δ-fair Bayes-optimal classifier (2) learned on $\mathbb{P}$.

Remark 3.2. *We need to maximize $\theta_{ay} \in [1/2,1]$ to maximize the objective function (3.1). As a result, we can take, when $t^\star_\delta > 0$,*

$$(\theta_{11},\theta_{10},\theta_{01},\theta_{00}) = \left(\frac{1}{1+T_1(t^\star_\delta)}, 1, 1, \frac{1}{1+T_0(t^\star_\delta)}\right), \quad (5)$$

and, when $t^\star_\delta < 0$,

$$(\theta_{11},\theta_{10},\theta_{01},\theta_{00}) = \left(1, \frac{1}{1-T_1(t^\star_\delta)}, \frac{1}{1-T_0(t^\star_\delta)}, 1\right), \quad (6)$$

By Theorem 3.1 we can express group fairness definitions such as Demographic Parity, Equalized Opportunity, and Predictive Equality in terms of the randomization mechanism which double as equality constraints.

Definition 3.3 (Demographic Parity)**.** *A randomization mechanism achieves Demographic Parity if it satisfies:*

$$(p_{11}+p_{10}+t^\star_\delta)\theta_{11} + (t^\star_\delta - p_{11} - p_{10})\theta_{10} = t^\star_\delta;$$
$$(t^\star_\delta - p_{01} - p_{00})\theta_{01} + (t^\star_\delta + p_{01} + p_{00})\theta_{00} = t^\star_\delta.$$

Definition 3.4 (Equality of Opportunity)**.** *A randomization mechanism achieves Equalized Opportunity if it satisfies:*

$$2p_{11}\theta_{11} + 2(t_\delta^\star - p_{11})\theta_{10} = t_\delta^\star;$$
$$-2p_{01}\theta_{01} + 2(t_\delta^\star + p_{01})\theta_{00} = t_\delta^\star.$$

Definition 3.5 (Predictive Equality)**.** *A randomization mechanism achieves Predictive Equality if it satisfies:*

$$2(t_\delta^\star + p_{10})\theta_{11} - 2p_{10}\theta_{10} = t_\delta^\star;$$
$$2(t_\delta^\star - p_{00})\theta_{01} + 2p_{00}\theta_{00} = t_\delta^\star.$$

3.2.2 FairRR: a Randomized Response Mechanism for Fair Classification

In this section, we propose the Randomized Response Mechanism that removes the discrimination from the training dataset. Based on the aforementioned theory, we are able to derive the optimal fair flipping probabilities as long as we estimate $p_{ay}, (a, y) \in \{0, 1\}^2$ and $t_\delta^\star$ from the training data. p_{ay} can be estimated directly by using its empirical estimator and $t_\delta^\star$ can be conveniently estimated using bisection methods due to its monotonic relationship with the decision disparity.

Here, we set $t_{min} = \inf_t : \{|T_a(t)| \leq 1 \text{ for } a \in \{0, 1\}, \}$ and $t_{max} = \sup_t : \{|T_a(t)| \leq 1 \text{ for } a \in \{0, 1\}, \}$. In each iteration, we update $t = (t_{max} + t_{min})/2$ and calculate the flipping probabilities as referenced in (5) and (6). Then, classifier $\hat{f}_t$ is learned from $(X, A, \widetilde{Y})$ with $\widetilde{Y} = \mathcal{R}_{\theta_{11},\theta_{10},\theta_{01},\theta_{00}}(Y)$. If the disparity level of $\hat{f}_t$ is greater than the pre-specified disparity level, we set $t_{min} = t_{mid}$ iterate until $t_\delta^\star$ is found.

Thus, with p_{ay} and $t_\delta^\star$ estimated, the optimal $(\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ can be solved for using (5), (6) and a corresponding group fairness definition, which maximizes 3.1 subject to the constraints of either 3.3, 3.4, or 3.5. With this randomization mechanism, the values in the privileged group $A = 1$ are randomly flipped from $Y = 1$ to $Y = 0$ and values in the unprivileged group $A = 0$ are randomly flipped from $Y = 0$ to $Y = 1$ such that a new perturbed response variable $\widetilde{Y}$ is created. Any classifier can then be fit to the original data X with perturbed $\widetilde{Y}$.

The time complexity of this method is dependent on which classifier is chosen for estimating $t_\delta^\star$ in the aforementioned bisection method. Here, a classifier has to be iteratively trained and evaluated to find the desired $t_\delta^\star$. In practice, we find that this takes relatively few iterations. In the case of using logistic regression with the LBFGS solver for example, the training complexity is $\mathrm{O}(p \times m)$ where p is the number of parameters and m is the number of memory corrections (Saputro and Widyaningsih [2017]). The evaluation of each iteration simplifies to $\mathrm{O}(n \times p)$ where n is the size of the evaluation set. Once $t_\delta^\star$ is estimated, the perturbation of Y is an $\mathrm{O}(N)$ process where N is the sample size of all data to be perturbed. The overall time complexity is thus dependent on n, m, p, and N as to what the final complexity reduces to.

4 EXPERIMENTS

4.1 Empirical Data Analysis

Datasets: We test FairRR on three benchmark datasets for fair classification: Adult Dua and Graff [2017], COMPAS Angwin et al. [2016] and Law School Wightman [1998].

- *Adult:* The target variable Y is whether the income of an individual is more than \$50,000. Age, marriage status, education level and other related variables are included in X, and the protected attribute A refers to gender.
- *COMPAS:* In the COMPAS dataset, the target is to predict recidivism. Here Y indicates whether or not a criminal will reoffend, while X includes prior criminal records, age and an indicator of misdemeanor. The protected attribute A is the race of an individual, "white-vs-non-white".
- *Law School:* The task of interest in Law School data set is to predict whether an applicant gets an admission from a law school based on common features include LSAT score and undergraduate GPA. The protected attribute A is the race of the individual: "white-vs-non-white"

Compared algorithms: In addition to FairRR, we also consider several benchmark methods in our experiments. As FairRR is a pre-processing method, we only include other pre-processing methods for comparison. Specifically, we consider the following:

- (1) Fair Sampling

  Fair Sampling Kamiran and Calders [2012] is a method based on adjusting the size of PP, PN, NP and NN groups. Its idea is to apply over/down sampling such that the label on the training data is independent of the sensitive attribute. Specifically, size of group PP, PN, NP and NN after sampling are:

$$n_{ay} = \frac{(n_{a1} + n_{a0})(n_{1y} + n_{0y})}{n_{11} + n_{10} + n_{01} + n_{00}}$$

- (2) FAWOS

Table 1: Benchmarking Results: Original vs FairRR Pre-processed Datasets

Panel A: Original Datasets

	Metrics				
Datasets	Acc	f_1	DDP	DEO	DPE
Adult	0.841 (0.003)	0.620 (0.007)	0.188 (0.006)	0.184 (0.026)	0.086 (0.005)
COMPAS	0.676 (0.015)	0.632 (0.016)	0.283 (0.031)	0.313 (0.052)	0.186 (0.035)
Law School	0.787 (0.003)	0.499 (0.005)	0.060 (0.005)	0.084 (0.015)	0.024 (0.004)

Panel B: Fair Randomized Response

	Fairness Criteria								
	Demographic Parity			Equality of Opportunity			Predictive Equality		
	Metrics			Metrics			Metrics		
Datasets	Acc	f_1	DDP	Acc	f_1	DEO	Acc	f_1	DPE
Adult	0.820 (0.004)	0.534 (0.009)	0.007 (0.005)	0.839 (0.003)	0.608 (0.007)	0.024 (0.02)	0.829 (0.004)	0.563 (0.009)	0.005 (0.004)
COMPAS	0.660 (0.015)	0.608 (0.017)	0.027 (0.019)	0.661 (0.014)	0.610 (0.016)	0.046 (0.037)	0.667 (0.014)	0.614 (0.016)	0.031 (0.024)
Law School	0.785 (0.003)	0.486 (0.005)	0.006 (0.004)	0.785 (0.003)	0.489 (0.005)	0.015 (0.011)	0.786 (0.003)	0.493 (0.005)	0.004 (0.004)

FAWOS Salazar et al. [2021] is another sampling method for fairness proposed recently. Unlike fair sampling that adjust the sizes of all four groups, FAWOS only applies SMOTE (Chawla et al. [2002], a popular oversampling method for unbalanced classification problem) to over-sample the points in the NN group where the number of points generated is:

$$N = \alpha \times \left(\frac{n_{11} n_{00}}{n_{10}} - n_{01} \right)$$

- (3) TabFairGAN

  TabFairGAN Rajabi and Garibay [2021] is a fair synthetic generation method based on the framework of generative adversarial network which adds a fairness penalty term to the generator loss of a standard WGAN model. Specifically, the fairness penalty is equal to the demographic parity of the generated data squared.

Experimental Setting: The goal of fair classification is to learn a classifier with the highest model utility, subject to some fairness constraint. Thus to test and benchmark FairRR we first apply each aforementioned fair pre-processing algorithm to each training dataset. A logistic regression classifier is then learned on the returned de-biased training dataset where it is then evaluated based on the average accuracy, f_1 score and disparity levels over 100 random 80:20 train/test splits. All model hyperparameters are left as the scikit-learn defaults for reproduciblity. The standard deviations of these metrics are also reported. All training and evaluations were processed using an Apple M1 CPU.

4.2 Results

We first evaluate the performance of FairRR controlling for either Demographic Parity, Equalized Opportunity, or Predictive Equality. We present the simulation results in Table 1. We observe that FairRR significantly controls for disparity across each fairness metric while seeing minimal decreases of model utility measured by accuracy and f_1 score. We then benchmark FairRR with other existing pre-processing methods. Here, only demographic parity is considered as it is the only common fairness metric supported across all pre-processing methods. We present these benchmarking results in Table 2.

Finally, we showcase the ability of FairRR to control for

Table 2: Benchmarking Results: Pre-processing Methods

		Methods				
Datasets	Metrics	Original	FairRR	TabFairGan	FS	FAWOS
Adult	Acc	0.841	0.820	0.804	0.836	0.786
		(0.003)	(0.004)	(0.008)	(0.003)	(0.004)
	DDP	0.188	0.007	0.023	0.091	0.008
		(0.006)	(0.005)	(0.024)	(0.008)	(0.006)
COMPAS	Acc	0.676	0.660	0.631	0.659	0.632
		(0.015)	(0.015)	(0.034)	(0.014)	(0.015)
	DDP	0.283	0.027	0.150	0.033	0.022
		(0.031)	(0.019)	(0.110)	(0.026)	(0.017)
Law School	Acc	0.787	0.785	0.774	0.784	0.782
		(0.003)	(0.003)	(0.030)	(0.003)	(0.003)
	DDP	0.060	0.006	0.060	0.006	0.006
		(0.005)	(0.004)	(0.153)	(0.004)	(0.004)

specific levels of disparity. In Table 3, FairRR was set to control for disparity at the quintiles between perfect demographic parity and the DDP level of the original dataset. The corresponding average DDP values in the final logistic regression and corresponding accuracies with standard deviations over 100 random seeds are reported. Figure 1 plots this experiment to highlight the accuracy/ disparity trade-off, comparing FairRR to FAWOS at these quintiles of controlled-for disparity. Figure 2 showcases the Pareto Curves of FairRR, Fair Sampling, FAWOS, and FairTabGAN when an SVM is trained on pre-processed data from the Adult Dataset.

5 DISCUSSION

Overall, FairRR achieves favorable or comparable-to-the-leader accuracy and disparity scores across the three benchmarking datasets. FairRR effectively maintains model utility while enforcing small amounts of disparity, regardless of the chosen group fairness definitions. One surprising result was the stability of the algorithm. One potential downside to FairRR could be with it randomly flipping labels the effectiveness could vary widely depending on the random seed. With low standard deviations across evaluation metrics though, FairRR proves to be also be robust. One interesting finding is that FairRR, Fair Sampling (FS), and FAWOS generally performed better than TabFairGan. We suspect this is because TabFairGan learns both X and y, which has advantages for applications such as privacy, but likely makes it weaker for pure fair classification tasks where FairRR and the over/under sampling strategies in FS and FAWOS perturb the feature space less.

FairRR also favorably controls for exact levels of disparity, an added benefit in applications where perfect group fairness is impractical or not needed. Table 3 shows empirically that disparity can be set to a level a-priori to model training and the downstream model will have that final level of disparity. Similarly, the trade-off between accuracy and disparity is better than competing methods that have this feature. With FAWOS conveniently allowing for the control of disparity we compare it with FairRR in Figure 1, showing that FairRR has a preferable utility curve to FAWOS in that at nearly all levels of disparity, the model trained with FairRR-processed data has better accuracy. This is further shown in Figure 2 which evaluates all competing methods on the Adult dataset with Support Vector Machines. Here, FairRR dominated the Pareto Frontier, noting that Fair Sampling does not allow for disparity control.

Another component investigated was the corresponding privacy offered by FairRR. As Randomized Response was first introduced as a privacy method, a natural extension of FairRR is to investigate the relationship between its utility/ fairness trade-off and the additional privacy it provides. This proves to be technically challenging. While Randomized Response is shown to satisfy (ϵ, δ)- Label Differential Privacy (Wang et al. [2016], Shirong et al. [2023]), the added fairness component of FairRR complicates a typical privacy analysis as it makes the privacy mechanism no longer independent of the data it is privatising. This is highlighted in the estimation of $t^{\star}_{\delta}$ where the design matrix is explicitly calculated based off of the disparity level in the original dataset. While $\widetilde{Y}$ is more private than Y, it remains unsolved how to quantify exactly how much more private it is in the context of some privacy budget. However, in application an advantage of pre-processing is that other

Table 3: Direct Control on Pre-specified Disparity Levels (δ)

Datasets		Metrics				
Adult	δ	0.000	0.040	0.080	0.120	0.160
	DDP	0.007 (0.005)	0.040 (0.008)	0.081 (0.008)	0.121 (0.008)	0.161 (0.007)
	Acc	0.820 (0.004)	0.826 (0.004)	0.833 (0.003)	0.838 (0.003)	0.841 (0.003)
COMPAS	δ	0.000	0.060	0.120	0.180	0.240
	DDP	0.027 (0.019)	0.062 (0.030)	0.123 (0.032)	0.182 (0.030)	0.239 (0.032)
	Acc	0.660 (0.015)	0.665 (0.014)	0.669 (0.014)	0.674 (0.015)	0.676 (0.015)
Law School	δ	0.000	0.012	0.024	0.036	0.048
	DDP	0.006 (0.004)	0.013 (0.006)	0.025 (0.007)	0.036 (0.006)	0.049 (0.006)
	Acc	0.785 (0.003)	0.785 (0.003)	0.786 (0.003)	0.786 (0.003)	0.786 (0.003)

pre-processing techniques can also be applied to the training data and in the context of privacy, FairRR is well-suited to be used in conjunction with other privacy mechanisms such as Laplacian and Exponential Noise (Jain et al. [2018]).

6 CONCLUSION

FairRR can be an excellent choice achieving group fairness in that it is a downstream model agnostic, efficient, and theory motivated algorithm that supports most group fairness definitions. In benchmarking, it performs comparably or better than other choices for pre-processing algorithms and additionally connects previous fair statistical learning theory to the pre-processing domain.

There are a variety of future research opportunities with FairRR. For starters, this paper only addresses the single binary sensitive attribute, single binary outcome problem formulation of fair classification. We believe that FairRR could be generalized to work in settings where multiple sensitive attributes are needed. Another interesting line of work is studying FairRR in the context of privacy, what Randomized Response was initially designed for. While this is technically challenging, we believe that extensions on FairRR could help shed light into the theoretical trade-offs between fairness and privacy. Lastly, we suspect there are a variety of additional mechanisms outside of randomized response to further apply the idea of pre-processing or post-processing data based on the Fair Optimal Bayes thresholding to achieve group fairness.

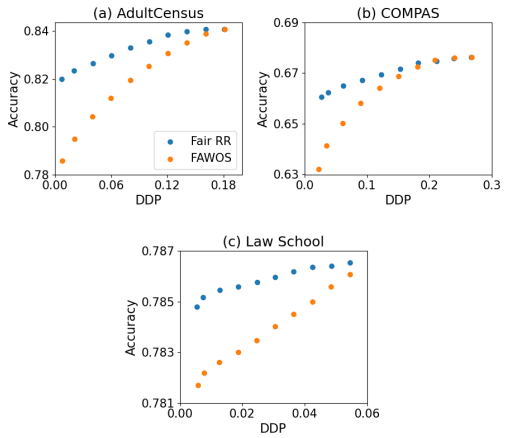

Figure 1: Logistic Regression Accuracy/ Disparity Trade-offs: FairRR and FAWOS comparison across datasets.

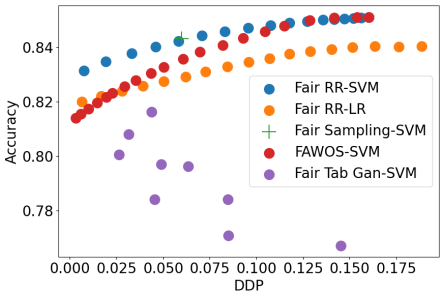

Figure 2: Accuracy/ Disparity Pareto Curves of various pre-processing algorithms on the Adult dataset evaluated with Support Vector Machines.

Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions. This work was partially supported by the JP Morgan Chase Faculty Research Award, NSF – CNS (2247795), and the Office of Naval Research (ONR N00014-22-1-2680). This work is also partially supported by the National Natural Science Foundation of China, No. 72033002.

References

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, May 2016.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009.

F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized preprocessing for discrimination prevention. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

L. E. Celis and V. Keswani. Improved adversarial learning for fair classification. *arXiv preprint arXiv:1901.10443*, 2019.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002.

J. Cho, G. Hwang, and C. Suh. A fair classifier using kernel density estimation. In *Advances in Neural Information Processing Systems*, pages 15088–15099. Curran Associates, Inc., 2020.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 797–806. Association for Computing Machinery, 2017.

S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312):1–117, 2023.

A. Cotter, H. Jiang, M. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.

E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel. Flexibly fair representation learning by disentanglement. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1436–1445. PMLR, 2019.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. Association for Computing Machinery, 2012.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. Association for Computing Machinery, 2015.

A. W. Flores, K. Bechtel, and C. T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to Machine Bias: "There's software used across the country to predict future criminals. And it's biased against blacks". *Federal Probation*, 80(2):38–46, 2016.

M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016.

P. Jain, M. Gyanchandani, and N. Khare. Differential privacy: its technological prescriptive using big data. *Journal of Big Data*, 5(15), 2018.

J. E. Johndrow and K. Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019.

M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016.

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

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer Berlin, Heidelberg, 2012.

P. Lahoti, K. P. Gummadi, and G. Weikum. iFair: Learning individually fair data representations for

algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE, 2019.

C. Louizos, K. Swersky, Y. Li, M. Welling, and R. S. Zemel. The variational fair autoencoder. In *4th International Conference on Learning Representations (ICLR)*, 2016.

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

D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, 2018.

A. Rajabi and O. O. Garibay. TabFairGAN: Fair tabular data generation with generative adversarial networks. *arXiv preprint arXiv:2109.00666*, 2021.

V. V. Ramaswamy, S. S. Y. Kim, and O. Russakovsky. Fair attribute classification through latent space debiasing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

A. Ruoss, M. Balunovic, M. Fischer, and M. Vechev. Learning certified individually fair representations. In *Advances in Neural Information Processing Systems*, pages 7584–7596. Curran Associates, Inc., 2020.

T. Salazar, M. S. Santos, H. Araújo, and P. H. Abreu. FAWOS: Fairness-aware oversampling algorithm based on distributions of sensitive attributes. *IEEE Access*, 9:81370–81379, 2021.

D. R. S. Saputro and P. Widyaningsih. Limited memory broyden-fletcher-goldfarb-shanno (l-bfgs) method for the parameter estimation on geographically weighted ordinal logistic regression model (gwolr). *AIP Conference Proceedings*, 1868(1):040009, 2017.

P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3:1–3:9, 2019.

N. Schreuder and E. Chzhen. Classification with abstention but without disparities. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 1227–1236. PMLR, 2021.

X. Shirong, C. Wang, W. W. Sun, and G. Cheng. Binary classification under local label differential privacy using randomized response mechanisms. *Transactions on Machine Learning Research*, 2023.

S. Tolan, M. Miron, E. Gómez, and C. Castillo. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 83–92. Association for Computing Machinery, 2019.

Y. Wang, X. Wu, and D. Hu. Using randomized response for differential privacy preserving data collection. In *EDBT/ICDT Workshops*, 2016.

S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

L. F. Wightman. Lsac national longitudinal bar passage study. lsac research report series, 1998. URL https://archive.lawschooltransparency.com/reform/projects/investigations/2015/documents/NLBPS.pdf.

D. Xu, S. Yuan, L. Zhang, and X. Wu. FairGAN: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018.

D. Xu, Y. Wu, S. Yuan, L. Zhang, and X. Wu. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1452–1458. International Joint Conferences on Artificial Intelligence Organization, 2019a.

D. Xu, S. Yuan, L. Zhang, and X. Wu. FairGAN+: Achieving fair data generation and classification through generative adversarial nets. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1401–1406, 2019b.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pages 325–333. PMLR, 2013.

X. Zeng, E. Dobriban, and G. Cheng. Fair Bayes-optimal classifiers under predictive parity. In *Advances in Neural Information Processing Systems*, pages 27692–27705. Curran Associates, Inc., 2022.

X. Zeng, G. Cheng, and E. Dobriban. Bayes-optimal fair classification with linear disparity constraints via pre-, in-, and post-processing. *arXiv preprint arXiv:2402.02817*, 2024.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. Association for Computing Machinery, 2018.