Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees

L. Elisa Celis[†], Lingxiao Huang[‡], Vijay Keswani[‡] and Nisheeth K. Vishnoi[†]
† Yale University ‡ EPFL

ABSTRACT

Developing classification algorithms that are fair with respect to sensitive attributes of the data is an important problem due to the increased deployment of classification algorithms in societal contexts. Several recent works have focused on studying classification with respect to specific fairness metrics, modeled the corresponding fair classification problem as constrained optimization problems, and developed tailored algorithms to solve them. Despite this, there still remain important metrics for which there are no fair classifiers with theoretical guarantees; primarily because the resulting optimization problem is non-convex. The main contribution of this paper is a meta-algorithm for classification that can take as input a general class of fairness constraints with respect to multiple non-disjoint and multi-valued sensitive attributes, and which comes with provable guarantees. In particular, our algorithm can handle non-convex "linear fractional" constraints (which includes fairness constraints such as predictive parity) for which no prior algorithm was known. Key to our results is an algorithm for a family of classification problems with convex constraints along with a reduction from classification problems with linear fractional constraints to this family. Empirically, we observe that our algorithm is fast, can achieve near-perfect fairness with respect to various fairness metrics, and the loss in accuracy due to the imposed fairness constraints is often small.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**;

KEYWORDS

Classification, Algorithmic Fairness

ACM Reference Format:
L. Elisa Celis[†], Lingxiao Huang[‡], Vijay Keswani[‡] and Nisheeth K. Vishnoi[†], † Yale University ‡ EPFL. 2019. **Classification with Fairness Constraints:, A Meta-Algorithm with Provable Guarantees. In *FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287586**

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287586

1 INTRODUCTION

Classification algorithms are increasingly being used in many societal contexts such as criminal recidivism [51], predictive policing [35], and job screening [47]. There are growing concerns that these algorithms may introduce significant bias with respect to certain sensitive attributes, e.g., against African-Americans while predicting future criminals [5, 7, 26], granting loans [17] or NYPD stop-and-frisk [30], and against women while recommending jobs [16]. The US Executive Office [52] also voiced concerns about discrimination in automated decision making, including health care delivery and education. Further, introducing bias may be illegal due to anti-discrimination laws [2, 6, 45], and can create social imbalance [1, 56]. Thus, developing classification algorithms that are fair with respect to sensitive attributes has become an important problem.

In classification, one is given a data vector and the goal is to decide whether it satisfies a certain property. The algorithm is allowed to learn from a set of labeled data vectors that may be assumed to come from an unknown distribution. The accuracy of a classifier is measured as the probability that the classifier correctly predicts the label of a data vector drawn from the same distribution. Each data vector, however, may also have a small number of multi-valued sensitive attributes such as race, gender, and political opinion, and each setting of a sensitive attribute gives rise to potentially non-disjoint groups of data points. Since fairness could mean different things in different contexts, a number of different metrics have been used to determine how fair a classifier is with respect to a sensitive group when compared to another, e.g., statistical parity [21], equalized odds [34], and predictive parity [20]. In fact, there are currently at least 21 well-accepted fairness metrics and counting; see [50].

Several recent works use the sensitive attributes and the desired notion of group fairness to place constraints on the classifier – formulating it as a constrained optimization problem that maximizes accuracy – and develop tailored algorithms to find such classifiers, e.g., constrained to statistical parity [29, 46, 60] or equalized odds [34, 46, 59]. However, these algorithms do not always come with provable guarantee, because often the resulting optimization problem turns out to be non-convex; e.g., for statistical parity [42, 60] and equalized odds [59]. Further, it is open whether such approaches would work for other important measures of disparate mistreatment such as predictive parity. Predictive parity, that measures whether the fractions over the class distribution for the predicted labels are close between different group that are important in predicting criminal recidivism [20, 26], stopping-and-frisking pedestrians [30], and predicting heart condition [54]. [59] left as an open problem to find algorithms to solve the fair classification problem with false discovery or false omission parity; two types of predictive parity.

Our contributions. We present a new classification algorithm that takes as input any one of a large class of fairness metrics which can be phrased as "linear-fractional constraints", and produces an (approximately) fair solution. Technically, we achieve this by

- identifying a family of classification problems with linear constraints (see Section 2),
- developing an algorithm to solve this constrained classification problem (see Section 4.2), and
- reducing classification with linear-fractional constraints to solving a small number of linear classification problems above for carefully chosen parameters (see Section 4.3).

Our approach is very flexible – it allows us to provide classifiers that are fair with respect to a host of fairness metrics corresponding to both of linear and non-linear constraints (see Table 1); examples include several prevalent fairness metrics. In particular, we obtain classifiers with predictive parity-type constraints for which there was no previous result with provable guarantees. Additionally, our algorithmic framework can handle multiple fairness metrics simultaneously, and the metrics can be defined with respect to complex sensitive attributes (e.g., multiple attributes, non-disjoint attributes, and/or multi-valued attributes). Further, we conduct an empirical evaluation of our algorithm on the **Adult, German credit** and **COMPAS** datasets and compare it against state-of-the art approaches in fair classification (see Section 5). The results show that our algorithm can often achieve higher fairness than prior work, and that the loss in accuracy due to imposing fairness constraints is often small. Thus, we provide a meta-algorithm for fair classification, which makes it flexible and easy to use in a variety of applications, is approximately optimal for whichever fairness metric is selected, and performs well in practice.

2 OUR MODEL

We consider the Bayesian model for classification. Let $\mathfrak{I}$ denote a joint distribution over the domain $\mathcal{D} = \mathcal{X} \times [p_1] \times \cdots \times [p_n] \times \{0,1\}$ where $\mathcal{X}$ is the feature space. Each sample $(X, Z_1, \ldots, Z_n, Y)$ is drawn from $\mathfrak{I}$ where each $Z_i \in [p_i]$ ($i \in [n]$) represents a sensitive attribute, and $Y \in \{0,1\}$ is the label of $(X, Z_1, \ldots, Z_n)$ that we want to predict. For the sake of readability, we discuss the case where there is only one sensitive attribute $Z \in \{1, 2, \ldots, p\}$ in the main text. This can be generalized to multiple sensitive attributes, by adding fairness constraints for all sensitive attributes, and is discussed in Appendix E.

Fixing different values of Z partitions the domain $\mathcal{D}$ into p groups $G_i := \{(x, i, y) \in \mathcal{D}\}$. Let $\mathcal{F}$ denote the collection of all possible classifiers. Given a loss function $L(\cdot\,;\,\cdot)$ that takes a classifier f and a distribution $\mathfrak{I}$ as arguments, there are two models for fair binary classification which have been studied in the literature and we consider:

(1) If Z is not used for prediction, then the goal is to learn a classifier $f : \mathcal{X} \to \{0,1\}$ that minimizes $L(f; \mathfrak{I})$. In this model, $\mathcal{F} = \{0,1\}^{\mathcal{X}}$.
(2) If Z is used for prediction, then the goal is to learn a classifier $f : \mathcal{X} \times [p] \to \{0,1\}$ that minimizes $L(f; \mathfrak{I})$. In this model, $\mathcal{F} = \{0,1\}^{\mathcal{X} \times [p]}$.

Denote by $\Pr_{\mathfrak{I}}[\cdot]$ the probability with respect to $\mathfrak{I}$. If $\mathfrak{I}$ is clear from context, we simply denote $\Pr_{\mathfrak{I}}[\cdot]$ by $\Pr[\cdot]$. A commonly used loss function is the prediction error, i.e., $L(f; \mathfrak{I}) = \Pr_{\mathfrak{I}}[f \neq Y]$. Here, with some abuse of notation, we use f to represent $f(X)$ for the first model and $f(X, Z)$ for the second model.

Apart from minimizing the loss function as usual, in fair classification problems the goal is also to achieve similar performance across all groups G_i. There are many metrics to measure this *group performance*, including statistical rate, true positive rate, accuracy rate or false discovery rates; see Table 1. For example, the statistical rate of G_i is of the form $\Pr_{\mathfrak{I}}[f = 1 \mid G_i]$, i.e., the probability of an event $(f = 1)$ conditioned on another event (G_i). Group performance can be defined in a general form as follows.

Definition 2.1 (**Group performance & group performance function**). *Given a classifier $f \in \mathcal{F}$ and $i \in [p]$, we call $q_i^{\mathfrak{I}}(f)$ the group performance of G_i if*

$$q_i^{\mathfrak{I}}(f) = \Pr_{\mathfrak{I}}\left[\mathcal{E} \mid G_i, \mathcal{E}'\right]$$

for some events $\mathcal{E}, \mathcal{E}'$ that might depend on the choice of f. Define a group performance function $q^{\mathfrak{I}} : \mathcal{F} \to [0,1]^p$ for any classifier $f \in \mathcal{F}$ as

$$q^{\mathfrak{I}}(f) = \left(q_1^{\mathfrak{I}}(f), \ldots, q_p^{\mathfrak{I}}(f)\right).$$

When $\mathfrak{I}$ is clear from context, we denote $q^{\mathfrak{I}}$ by q. At a high level, a classifier f is considered to be fair w.r.t. to q if $q_i(f) \approx q_j(f)$ for all i, j.

Consider the following examples of q.

(1) **Accuracy Rate:** Here $\mathcal{E} := (f = Y)$ and $\mathcal{E}' := \emptyset$, i.e., $q_i(f)$ is the accuracy of the classifier on group G_i, we can rewrite $q_i(f)$ as follows (see Lemma A.1 in Appendix A):

$$q_i(f) = \Pr[Y = 0 \mid G_i] + \Pr[Y = 1 \mid G_i] \cdot \Pr[f = 1 \mid Y = 1, G_i] - \Pr[Y = 0 \mid G_i] \cdot \Pr[f = 1 \mid Y = 0, G_i],$$

i.e., a linear combination of conditional probabilities $\Pr[f = 1 \mid G_i, Y = 0]$ and $\Pr[f = 1 \mid G_i, Y = 1]$.

(2) **False Discovery Rate:** Here $\mathcal{E} := (Y = 0)$ and $\mathcal{E}' := (f = 1)$, i.e., $q_i(f)$ is the prediction error on the sub-group of G_i with positive predicted labels, we can rewrite $q_i(f)$ as follows (see Lemma A.2 in Appendix A):

$$q_i(f) = \frac{\Pr[Y = 0, G_i] \cdot \Pr[f = 1 \mid G_i, Y = 0]}{\Pr[G_i] \cdot \Pr[f = 1 \mid G_i]},$$

i.e., the fraction of two conditional probabilities $\Pr[f = 1 \mid G_i, Y = 0]$ and $\Pr[f = 1 \mid G_i]$.

In both these examples, $q_i(f)$ can be written in terms of probabilities $\Pr[f = 1 \mid G_i, \cdot]$ as either a linear combination, or as a quotient of linear combinations. Below we define two general classes of group performance functions that generalize these two examples respectively.

Definition 2.2 (**Linear-fractional/linear group performance functions**). *A group performance function q is called **linear-fractional** if for any $f \in \mathcal{F}$ and $i \in [p]$, $q_i(f)$ can be written as*

$$q_i(f) = \frac{\alpha_0^{(i)} + \sum_{j=1}^{k} \alpha_j^{(i)} \cdot \Pr_{\mathfrak{I}}\left[f = 1 \mid G_i, \mathcal{A}_j^{(i)}\right]}{\beta_0^{(i)} + \sum_{j=1}^{l} \beta_j^{(i)} \cdot \Pr_{\mathfrak{I}}\left[f = 1 \mid G_i, \mathcal{B}_j^{(i)}\right]} \tag{1}$$

Table 1: Summary of prior work. The symbol ✓ (or ⋆) represents that the corresponding framework works for (or can be extended to handle) the corresponding fairness metric. The events $\mathcal{E}$ and $\mathcal{E}'$ determine the group performance function $q_i(f)$ of the fairness metric (see Defn. 2.1), while L/LF represents whether this group performance function is linear or linear-fractional.

		$q_i(f)$		$\mathcal{Q}_{\text{lin}}/\mathcal{Q}_{\text{linf}}$	This paper	[34]	[57]	[60]	[59]	[46]	[29]	[42]
		$\mathcal{E}$	$\mathcal{E}'$									
fairness metrics	statistical	$f=1$	$\emptyset$	$\mathcal{Q}_{\text{lin}}$	✓			✓		✓	✓	✓
	conditional statistical	$f=1$	$X \in S$	$\mathcal{Q}_{\text{lin}}$	✓			✓		⋆	⋆	
	false positive	$f=1$	$Y=0$	$\mathcal{Q}_{\text{lin}}$	✓	✓	✓		✓	⋆		✓
	false negative	$f=0$	$Y=1$	$\mathcal{Q}_{\text{lin}}$	✓	⋆	⋆		✓	⋆		✓
	true positive	$f=1$	$Y=1$	$\mathcal{Q}_{\text{lin}}$	✓	✓	✓		⋆	✓		
	true negative	$f=0$	$Y=0$	$\mathcal{Q}_{\text{lin}}$	✓	⋆	⋆		⋆	⋆		
	accuracy	$f=Y$	$\emptyset$	$\mathcal{Q}_{\text{lin}}$	✓				✓	⋆		
	false discovery	$Y=0$	$f=1$	$\mathcal{Q}_{\text{linf}}$	✓							
	false omission	$Y=1$	$f=0$	$\mathcal{Q}_{\text{linf}}$	✓							
	positive predictive	$Y=1$	$f=1$	$\mathcal{Q}_{\text{linf}}$	✓							
	negative predictive	$Y=0$	$f=0$	$\mathcal{Q}_{\text{linf}}$	✓							

*for two integers $k, l \geq 0$, events $\mathcal{A}_1^{(i)}, \ldots, \mathcal{A}_k^{(i)}, \mathcal{B}_1^{(i)}, \ldots, \mathcal{B}_l^{(i)}$ that are independent of the choice of f, and parameters $\alpha_0^{(i)}, \ldots, \alpha_k^{(i)}$, $\beta_0^{(i)}, \ldots, \beta_l^{(i)} \in \mathbb{R}$ that may depend on $\mathfrak{I}$ but are independent of the choice of f. Denote $\mathcal{Q}_{\text{linf}}$ to be the collection of all linear-fractional group performance functions. Specifically, if $l=0$ and $\beta_0^{(i)}=1$ for all $i \in [p]$, q is said to be **linear**. Denote $\mathcal{Q}_{\text{lin}} \subseteq \mathcal{Q}_{\text{linf}}$ to be the collection of all linear group performance functions.*

In Appendix A, we show that all q in Table 1 are linear-fractional, and, in fact, many are linear (see the $\mathcal{Q}_{\text{lin}}/\mathcal{Q}_{\text{linf}}$ column).

A classifier f is said to satisfy τ-rule w.r.t. to a given group performance function q if $\rho_q(f) := \min_{i\in[p]} q_i(f)/\max_{i\in[p]} q_i(f) \geq \tau$; see [24, 46, 59, 60]. The closer τ is close to 1, the fairer f is with respect to q. Assume there are m fractional group performance functions $q^{(1)}, \ldots, q^{(m)} \in \mathcal{Q}_{\text{linf}}$ and $L(f;\mathfrak{I}) = \Pr_{\mathfrak{I}}[f \neq Y]$. Given $\tau_1, \ldots, \tau_m \in [0,1]$, our main objective is to solve the following fair classification program induced by ρ_q that we refer to as ρ-Fair.

$$\min_{f\in\mathcal{F}} \Pr_{\mathfrak{I}}[f \neq Y] \quad s.t. \quad \rho_{q^{(i)}}(f) = \min_{j\in[p]} q_j^{(i)}(f)/\max_{j\in[p]} q_j^{(i)}(f) \geq \tau_i,\ \forall i \in [m]. \tag{ρ-Fair}$$

It follows that by setting q appropriately, ρ-Fair captures several existing constrained classification problems as special cases, e.g., statistical rate [3, 46], true positive rate [3, 46], or predictive rate [55].

Remark 2.3. *If $\tau_i = 1$, the program above computes a classifier f with perfect fairness w.r.t. to $q^{(i)}$. This setting is well studied in the literature [9, 21, 24, 34, 59, 60, 62]. However, perfect fairness is known to have deficiencies [27, 34, 41] and, hence, prior work considers the relaxed fairness metric τ-rule where $\tau_i < 1$. Another relaxed fairness metric is defined by $\delta_q(f) := \min_{i\in[p]} q_i(f) - \max_{i\in[p]} q_i(f)$. Computing a classifier f such that $\delta_q(f) \geq \tau$ ($\tau \in [-1, 0]$) has also been investigated in the literature [9, 46]. We refer the reader to a survey [63] for other relaxed fairness metrics, e.g., AUC and correlation.*

Computationally, the constraints of ρ-Fair are non-convex making the problem of solving it (even approximately) intractable in general. To bypass this, we introduce a fair classification problem with linear constraints, which we call Group-Fair, that has additional parameters corresponding to lower and upper bounds and corresponding fairness constraints.

Definition 2.4 (**Group-Fair**). *Given $\ell_j^{(i)}, u_j^{(i)} \geq 0$ for all $i \in [m]$ and $j \in [p]$, we consider the following classification problem with fairness constraints:*

$$\min_{f\in\mathcal{F}} \Pr_{\mathfrak{I}}[f \neq Y] \quad s.t.,\ \ell_j^{(i)} \leq q_j^{(i)}(f) \leq u_j^{(i)},\ \forall i \in [m], j \in [p]. \tag{Group-Fair}$$

On the positive side, these constraints are linear, resulting in a convex programming problem and give us a finer control over the group performance function. In particular, by selecting non-uniform parameters $\ell_j^{(i)}, u_j^{(i)}$, Group-Fair can treat different groups differently while all groups are symmetrically regarded in ρ-Fair. Moreover, it is easy to see that for any feasible classifier f of Group-Fair and any $i \in [m]$, f satisfies $\frac{\min_{j\in[p]} \ell_j^{(i)}}{\max_{j\in[p]} u_j^{(i)}}$-rule w.r.t. to $q^{(i)}$.

While being tractable, ρ-Fair raises the problem of finding the appropriate values of the lower and upper bounds that are not part of the input to Group-Fair. In Section 4.1, we show how to solve Group-Fair by a small number of calls to ρ-Fair where we set the lower and upper bound parameters in each call to ρ-Fair carefully.

Remark 2.5. *We remark that our algorithms assume the existence of an oracle that evaluates $q_i^{\mathfrak{I}}(f)$ sufficiently well for any given classifier f. To overcome this issue, we note that we can estimate $q_i^{\mathfrak{I}}(f) = \Pr_{\mathfrak{I}}[\mathcal{E} \mid G_i, \mathcal{E}']$ by the empirical probability of samples drawn from $\mathfrak{I}$, i.e., the ratio between the number of samples satisfying $\mathcal{E} \cap G_i \cap \mathcal{E}'$ and the number of samples satisfying $G_i \cap \mathcal{E}'$.*

3 RELATED WORK

From a technical perspective, the most relevant prior work includes [15, 39, 46], which considered the Bayesian classification model for statistical parity or equalized odds. [46] reduce their constrained classification problems to unconstrained optimization problem by the Lagrangian principle, while [15, 39] aim to find

optimal threshold rules or use regularizers to find a fair classifier. Our framework also uses the Lagrangian principle, but works for a much wider class than these works, i.e., any linear-fractional group performance function.

Some very recent work has also taken steps towards providing a unifying approach to fair classification. [55] encode fairness constraints, including statistical parity, equalized odds, and predictive parity, as a distance between the distributions for different values of a single binary sensitive attribute, and then use the privileged learning framework to optimize loss with respect to fairness constraints. While this results in an interesting heuristic, they do not provide theoretical guarantees for their approach. [3] give a method to compute a nearly optimal fair classifier with respect to statistical parity or equalized odds by the Lagrangian principle. In particular, their framework supports fairness constraints that are linearly dependent on the conditional moments of the form $\mathbb{E}[g(\cdot, f) \mid \mathcal{E}]$, where g is a function that depends on the classifier f along with features of the element while $\mathcal{E}$ is an event that does not depend on f. However linear-fractional constraints cannot be directly represented in this form, since here the event we condition on, $\mathcal{E}$, depends on the classifier f, which is why their framework does not support constraints like predictive parity. [54] use post-processing techniques to achieve calibration[1] along with single error constraints, such as false-positive or false-negative parity, but do not provide any provably guarantee with respect to predictive parity. We also note that our linear fairness constraints are inspired by works on other fundamental algorithmic problems such as data summarization [11], ranking [12, 58], elections [10], and personalization [13].

There are increasingly many works on fairness in machine learning with provable guarantees, including [25, 34, 54, 57], that provide different classification algorithms with constraints on statistical parity or equalized odds, and [36, 40] for fairness in multi-armed bandit settings or ranking problems respectively. To the best of our knowledge, our algorithm is the first unifying framework for all current [50] fairness metrics, with provable guarantees.

Many alternate approaches to improve the fairness of classification have also been studied. One approach is to make predictions without the information of sensitive attributes, which avoids disparate treatment [2]. However, since the learning data may contain historical biases, classifiers trained on such data may still have indirect discrimination for certain sensitive groups [53]. Another approach is to modify the classification problem to incorporate constraints of some kind. For example, one approach proposes other fairness metrics as a proxy of statistical parity or equalized odds, e.g., [29, 59, 60]. [59, 60] propose a covariance-type constraint for statistical parity and equalized odds. Their model does not require the sensitive attribute Z to be explicitly provided during prediction, thereby preventing disparate treatment. Yet another approach post-processes a baseline classifier by shifting the decision boundary (can be different for different groups), e.g., [22, 25, 31, 34, 54, 57]. [34] use the sensitive attribute Z during prediction. Their goal can be regarded as learning a different classifier for each G_i. Alternatively, another line of research is to pre-process on the training data and achieve an unbiased dataset for learning, e.g., [24, 37, 38, 42, 44, 62]. This approach is quite different from ours since we focus on learning classifiers and investigating the accuracy-fairness tradeoff from the feeding dataset.

Beyond group fairness, recent works also proposed other notions of fairness in classification. [21] and [62] discussed a notion of *individual fairness* that similar individuals should be treated similarly. [61] defined *preference fairness* based on the concepts of fair division and envy-freeness in economics. Moreover, [32, 33] discussed *procedural fairness* that investigates which input features are fair to use in the decision process and how including or excluding the features would affect outcomes. Finally, [14] and [41] investigated the inherent tradeoff between equalized odds and predictive parity (called well-calibrated in their papers).

[1] Calibration is a stronger fairness constraint that satisfies predictive parity.

4 THEORETICAL RESULTS

In this section, we present an efficient algorithm to approximately solve ρ-Fair (Theorem 4.4, Section 4.3). Towards this goal, we first show that ρ-Fair can be efficiently reduced to a family of programs with linear fairness constraints (Group-Fair - Section 4.1). Subsequently, we show that there exists a polynomial time algorithm that computes an approximately optimal classifier for Group-Fair (Section 4.2). For convenience, we only consider $m = 1$ in this section, i.e., there is only one group performance function q and we require $\rho_q(f) \geq \tau$ for some $\tau \in [0, 1]$. This can be generalized to multiple group performance functions as discussed in Appendix E.

4.1 Reduction from ρ-Fair to Group-Fair

We first show the generality of Group-Fair, i.e., approximately solving ρ-Fair can be reduced to solving a family of Group-Fair. A β-approximate algorithm for Group-Fair ($\beta \geq 1$) is an efficient algorithm that computes a feasible classifier with prediction error at most β times the optimal prediction error of Group-Fair.

Theorem 4.1 (**Reduction from ρ-Fair to Group-Fair**). *Given $\tau \in [0, 1]$, let $f_\tau^\star$ denote an optimal fair classifier for ρ-Fair. Given a β-approximate algorithm A for Group-Fair ($\beta \geq 1$) and any $\varepsilon > 0$, there exists an algorithm that calls A at most $\lceil \tau/\varepsilon \rceil$ times and computes a classifier $f \in \mathcal{F}$ such that*

(1) $\Pr[f \neq Y] \leq \beta \cdot \Pr\left[f_\tau^\star \neq Y\right]$;
(2) $\min_{i \in [p]} q_i(f) \geq \tau \cdot \max_{i \in [p]} q_i(f) - \varepsilon$.

Theorem 4.1 asserts that the ρ-Fair program can be solved efficiently by solving at most $\lceil \tau/\varepsilon \rceil$ different Group-Fair programs. The resulting classifier f slightly violates the τ-rule since there is an additive error term ε in the right side of (2). As the error ε goes to 0, the violation becomes small and hence the resulting classifier is guaranteed to be more fair with respect to q. However, the running time becomes longer since it depends on the term $\lceil \tau/\varepsilon \rceil$.

Proof of Theorem 4.1. Let $T := \lceil \tau/\varepsilon \rceil$. For each $t \in [T]$, denote $a_t := (t-1) \cdot \varepsilon$ and $b_t := t \cdot \varepsilon/\tau$. For each $t \in [T]$, we construct a Group-Fair program P_t with $\ell_j = a_t$ and $u_j = b_t$ for all $j \in [p]$. Then we apply A to compute $f_t \in \mathcal{F}$ as a solution of P_t. Among all f_t, we output f such that $\Pr[f \neq Y]$ is minimized. Next, we verify that f satisfies the conditions in the theorem.

Note that $a_t \geq \tau \cdot b_t - \varepsilon$ for each $t \in [T]$. We have

$$\min_{i \in [p]} q_i(f) \geq \tau \cdot \max_{i \in [p]} q_i(f) - \varepsilon.$$

On the other hand, assume that $(t'-1)\cdot\varepsilon \le \min_{i\in[p]} q_i(f_\tau^\star) < t'\cdot\varepsilon$ for some $t' \in [T]$. Since $f_\tau^\star$ is a feasible solution of ρ-Fair,

$$\max_{i\in[p]} q_i(f_\tau^\star) \le \frac{\min_{i\in[p]} q_i(f_\tau^\star)}{\tau} < t'\cdot\varepsilon/\tau.$$

Hence, $f_\tau^\star$ is a feasible solution of Program P_j. By the definitions of A and f, we have

$$\Pr[f \neq Y] \le \Pr[f_{t'} \neq Y] \le \beta\cdot\Pr\left[f_\tau^\star \neq Y\right].$$

□

The above theorem can be generalized to any loss function instead of the prediction error. The reduction also holds for the $m > 1$ case. The only difference is that we need to call algorithm A roughly ε^{-m} times. This enables us to simultaneously handle multiple fairness requirements; see Appendix E for details. The reduction is efficient with respect to running time as well and so to efficiently solve a ρ-Fair program we just need to construct an algorithm for the Group-Fair program.

4.2 Algorithm for Group-Fair

In this section, we propose an algorithm for Group-Fair. Due to space limitations, we omit many details (see Appendix B). For concreteness, we first consider the setting where $\mathcal{F} = \{0,1\}^{\mathcal{X}}$ and $q \in \mathcal{Q}_{\text{lin}}$ and subsequently discuss the $q \in \mathcal{Q}_{\text{linf}}$ case.

By Definition 2.2, assume that

$$q_i(f) = \alpha_0^{(i)} + \sum_{j=1}^{k} \alpha_j^{(i)}\cdot\Pr\left[f=1 \mid G_i, \mathcal{A}_j^{(i)}\right]$$

for $f \in \mathcal{F}$ and $i \in [p]$. Without fairness constraints, it can be shown that

$$f^\star := \mathrm{I}[\Pr[Y=1 \mid X=x] - 0.5 > 0]$$

is an optimal classifier minimizing the prediction error $\Pr[f \neq Y]$, where $I[\cdot]$ is the indicator function. But such a classifier $f^\star$ might not satisfy all the fairness constraints. Hence, we introduce a regularization parameter $\lambda \in \mathbb{R}^p$ and study the following problem

$$f_\lambda^\star := \arg\min_{f\in\mathcal{F}} \Pr[f \neq Y] - \sum_{i\in[p]} \lambda_i\cdot q_i(f). \tag{2}$$

Now we can "control" $q_i(f_\lambda^\star)$ by adjusting λ. Intuitively, increasing λ_i leads to an increase in $q_i(f_\lambda^\star)$. By selecting suitable λ, we can expect that $f_\lambda^\star$ satisfies all fairness constraints. We will show that there exists some $\lambda \in \mathbb{R}^p$ such that Group-Fair is equivalent to (2), by the Lagrangian principle. Moreover, $f_\lambda^\star$ can be shown to be an instance-dependent threshold function with the threshold

$$s_\lambda(x) := \Pr[Y=1 \mid X=x] - 0.5 + \sum_{i\in[p]} \lambda_i\cdot\psi_i(x), \tag{3}$$

where $\psi_i(x) = \sum_{j=1}^{k} \frac{\alpha_j^{(i)}}{\Pr\left[G_i, \mathcal{A}_j^{(i)}\right]}\cdot\Pr\left[G_i, \mathcal{A}_j^{(i)} \mid X=x\right]$ is the scaling factor of λ_i that is determined by the form of $q_i(f)$. Observe that the term $\Pr[Y=1 \mid X=x] - 0.5$ is exactly the threshold for the unconstrained optimal classifier $f^\star$, and the remaining term $\sum_{i\in[p]} \lambda_i\cdot\psi_i(x)$ can be regarded as a threshold correction induced by λ.

Theorem 4.2 (**Solution characterization and computation for $q \in \mathcal{Q}_{\text{lin}}$**). *Given any parameters $\ell, u \in [0,1]^p$, there exist optimal Lagrangian parameters $\lambda^\star \in \mathbb{R}^p$ such that $\mathrm{I}[s_{\lambda^\star}(X) > 0]$ is an optimal fair classifier for Group-Fair. Moreover, $\lambda^\star$ can be computed in polynomial time as a solution to the following convex program:*

$$\lambda^\star = \arg\min_{\lambda\in\mathbb{R}^p} g(\lambda) = \arg\min_{\lambda\in\mathbb{R}^p} \mathbb{E}_{X\sim\mathfrak{I}}\left[|s_\lambda(X)|\right] + \sum_{i\in[p]}\left(\alpha_0^{(i)} - u_i\right)\lambda_i + \sum_{i\in[p]}(u_i-\ell_i)\cdot\max\{0,\lambda_i\}. \tag{4}$$

This theorem asserts that Group-Fair can be solved efficiently (up to an arbitrary accuracy): first compute the optimal Lagrangian parameters $\lambda^\star$ via (4) and then output the fair classifier $\mathrm{I}[s_{\lambda^\star}(X) > 0]$. The running time depends on how fast we can solve Program (4). Since $u_i - \ell_i \ge 0$, Program (4) is convex and we can apply standard convex optimization algorithms, e.g., the stochastic subgradient method [8] to compute an ε-approximate λ such that $g(\lambda) \le g(\lambda^\star) + \varepsilon$ in $\tilde{O}(k^2p/\varepsilon^2)$ time.

The proof of this theorem reduces Group-Fair to an unconstrained optimization problem by the Lagrangian principle (Appendix B.1). We then derive (4) as the dual program to Group-Fair and show that $\lambda^\star$ is an optimal solution to (4) (Appendix B.2). Consequently, Theorem 4.2 leads to an algorithm Group-Fair$\left(\mathfrak{I}, q^{\mathfrak{I}}, \ell, u\right)$ that computes an optimal fair classifier for the Group-Fair program. Theorem 4.2 can also be directly extended to $\mathcal{F} = \{0,1\}^{\mathcal{X}\times[p]}$ by replacing X to (X,Z) everywhere.

Menon and Williamson [46, Algorithm 1] also propose an algorithmic framework for fair classification with respect to statistical rate and true positive rate using the Lagrangian principle. However, they only analyzed the characterization but did not show how to compute the optimal Lagrangian parameters. Our approach can be naturally applied to their setting for computing the optimal Lagrangian parameters; see Appendix F for details.

Theorem 4.2 can be generalized to $q \in \mathcal{Q}_{\text{linf}}$. The key observation is that we can rewrite the fairness constraint $\ell_i \le q_i(f)$ as

$$\alpha_0^{(i)} + \sum_{j=1}^{k}\alpha_j^{(i)}\cdot\Pr_{\mathfrak{I}}\left[f=1 \mid G_i, \mathcal{A}_j^{(i)}\right] \ge \ell_i\cdot\left(\beta_0^{(i)} + \sum_{j=1}^{l}\beta_j^{(i)}\cdot\Pr_{\mathfrak{I}}\left[f=1 \mid G_i, \mathcal{B}_j^{(i)}\right]\right).$$

By rearranging, the above inequality is expressible as a linear constraint $a^\top f + b \le 0$. This also holds for $q_i(f) \le u_i$, which implies that Group-Fair is a linear program of f. Hence, introducing fairness constraints can handle predictive parity with $q \in \mathcal{Q}_{\text{linf}}$, but the prior work can not – due to the fact that the constraint $q_i(f) \ge \tau\cdot q_j(f)$ may not be convex in general. [2]

Similar to $q \in \mathcal{Q}_{\text{lin}}$, we also apply the Lagrangian principle. The only difference is that we need to introduce two regularization parameters ν_i and ζ_i respectively for constraints $\ell_i \le q_i(f)$ and

[2] On a cursory look, this reduction may seem related to linear-fractional programs. However, the *objective function* of linear-fractional programming is a ratio of linear functions and this results in a simple reduction from it to linear programming (see https://en.wikipedia.org/wiki/Linear-fractional_programming). In our reduction, the constraints are linear-fractional and it does not seem easy to reduce to a single linear program.

$q_i(f) \le u_i$. Then similar to (3), for any regularization parameters $\nu, \zeta \in \mathbb{R}^p_{\ge 0}$, we define an instance-dependent threshold function

$$s_{\nu,\zeta}(x) := \Pr[Y = 1 \mid X = x] - 0.5 + \sum_{i\in[p]} \nu_i \cdot \left(\sum_{j\in[k]} \frac{\alpha_j^{(i)}}{\pi_j^{(i)}} \cdot \eta_j^{(i)}(x) - \ell_i \sum_{j\in[l]} \frac{\beta_j^{(i)}}{\omega_j^{(i)}} \cdot \xi_j^{(i)}(x) \right) + \zeta_i \cdot \left(- \sum_{j\in[k]} \frac{\alpha_j^{(i)}}{\pi_j^{(i)}} \cdot \eta_j^{(i)}(x) + u_i \sum_{j\in[l]} \frac{\beta_j^{(i)}}{\omega_j^{(i)}} \cdot \xi_j^{(i)}(x) \right),$$

which consists of the term $\Pr[Y = 1 \mid X = x] - 0.5$ that is the threshold for the unconstrained optimal classifier $f^\star$, and threshold correction terms induced by ν and ζ. Then we prove the following theorem which indicates that $q \in \mathcal{Q}_{\text{linf}}$ can also be solved efficiently by first computing the optimal Lagrangian parameters $\nu^\star, \zeta^\star$ and then outputting the fair classifier $\mathrm{I}[s_{\nu^\star,\zeta^\star}(X) > 0]$.

Theorem 4.3 (**Solution characterization and computation for $q \in \mathcal{Q}_{\text{linf}}$**). *Suppose $\mathcal{F} = \{0,1\}^{\mathcal{X}}$ and $q \in \mathcal{Q}_{\text{linf}}$. Given any parameters $\ell_i, u_i \in [0,1]$ $(i \in [p])$, there exists $\nu^\star, \zeta^\star \in \mathbb{R}^p_{\ge 0}$ such that $\mathrm{I}[s_{\nu^\star,\zeta^\star}(x) > 0]$ is an optimal fair classifier for Group-Fair. Moreover, we can compute the optimal Lagrangian parameters $\nu^\star$ and $\zeta^\star$ in polynomial time as a solution of the following convex program:*

$$(\nu^\star, \zeta^\star) = \arg \min_{\nu,\zeta \in \mathbb{R}^p_{\ge 0}} \mathbb{E}_X \left[|s_{\nu,\zeta}(X)| \right] + \sum_{i\in[p]} \nu_i \cdot \left(\alpha_0^{(i)} - \ell_i \beta_0^{(i)} \right) + \sum_{i\in[p]} \zeta_i \cdot \left(-\alpha_0^{(i)} + u_i \beta_0^{(i)} \right).$$

4.3 Algorithm for ρ-Fair

We proceed to designing an algorithm that handles the fairness metric ρ_q. In real-world settings, instead of knowing $\mathfrak{I}$, we only have N samples $\{(x_i, z_i, y_i)\}_{i\in[N]}$ drawn from $\mathfrak{I}$. To handle this, we use the idea inspired by [46, 49]: estimate $\mathfrak{I}$ by $\widehat{\mathfrak{I}}$, e.g., via Gaussian Naive Bayes or logistic regression on samples, and then compute a classifier based on $\widehat{\mathfrak{I}}$ by solving a family of Group-Fair programs as stated in Theorem 4.1 ; see Algorithm 1. By Theorems 4.1 and 4.2, the running time of Algorithm 1 is polynomial in N.

Analyzing Algorithm 1. Intuitively, if $\widehat{\mathfrak{I}}$ is close to $\mathfrak{I}$, then the quality of f in both accuracy and fairness should be comparable to an optimal fair classifier for ρ-Fair under $\mathfrak{I}$. Define

$$\kappa := 2 \max_{i\in[p], f\in\mathcal{F}} \left| q_i^{\widehat{\mathfrak{I}}}(f) - q_i^{\mathfrak{I}}(f) \right|$$

as the error introduced in $q^{\mathfrak{I}}$ when replacing $\mathfrak{I}$ by $\widehat{\mathfrak{I}}$. Let $d_{TV}(\mathfrak{I}, \widehat{\mathfrak{I}})$ denote the total variation distance between $\widehat{\mathfrak{I}}$ and $\mathfrak{I}$.

Theorem 4.4 (**Quantification of the output classifier**). *Let $f^\star$ be a fair classifier minimizing the prediction error $\Pr_{\mathfrak{I}}[f \ne Y]$ subject to the relaxed τ-rule:*

$$\min_{i\in[p]} q_i^{\mathfrak{I}}(f) \ge \tau \cdot \max_{i\in[p]} q_i^{\mathfrak{I}}(f) + \kappa.$$

Then Algorithm 1 outputs a classifier f such that

(1) $\Pr_{\mathfrak{I}}[f \ne Y] \le \Pr_{\mathfrak{I}}\left[f^\star \ne Y\right] + 2 \cdot d_{TV}(\widehat{\mathfrak{I}}, \mathfrak{I})$;
(2) $\min_{i\in[p]} q_i^{\mathfrak{I}}(f) \ge \tau \cdot \max_{i\in[p]} q_i^{\mathfrak{I}}(f) - \varepsilon - \kappa$.

Algorithm 1: An algorithm for ρ-Fair

Input : Samples $\{(x_i, z_i, y_i)\}_{i\in[N]}$ from distribution $\mathfrak{I}$, a linear-fractional group performance function $q^{\mathfrak{I}} \in \mathcal{Q}_{\text{linf}}$, a fairness parameter $\tau \in [0,1]$ and an error parameter $\varepsilon \in [0,1]$.

Output : A classifier $f \in \mathcal{F}$.

1 Compute an estimated distribution $\widehat{\mathfrak{I}}$ (e.g., via Gaussian Naive Bayes) on $\{(x_i, z_i, y_i)\}_{i\in[N]}$.
2 $T \leftarrow \lceil \tau/\varepsilon \rceil$. For each $t \in [T]$, $a_t \leftarrow (t-1)\cdot\varepsilon$ and $b_t \leftarrow t \cdot \varepsilon/\tau$.
3 For each $t \in [T]$, let $f_t \leftarrow$ Group-Fair$\left(\widehat{\mathfrak{I}}, q^{\widehat{\mathfrak{I}}}, \{\ell_j = a_t\}_{j\in[p]}, \{u_j = b_t\}_{j\in[p]}\right)$.
4 Return $f \leftarrow \arg\min_{f_t} \Pr_{\widehat{\mathfrak{I}}}[f_t \ne Y]$.

We defer the proof to Appendix C.1. The key is to show $f^\star$ is feasible for ρ-Fair under $\widehat{\mathfrak{I}}$. This can be inferred by the assumption that $\min_{i\in[p]} q_i^{\widehat{\mathfrak{I}}}(f^\star) \ge \tau \cdot \max_{i\in[p]} q_i^{\widehat{\mathfrak{I}}}(f^\star) + \kappa$ and the definition of κ. Then we prove by Theorem 4.1 that

(1) $\Pr_{\widehat{\mathfrak{I}}}[f \ne Y] \le \Pr_{\widehat{\mathfrak{I}}}\left[f^\star \ne Y\right]$ and
(2) $\min_{i\in[p]} q_i^{\widehat{\mathfrak{I}}}(f) \ge \tau \cdot \max_{i\in[p]} q_i^{\widehat{\mathfrak{I}}}(f) - \varepsilon$.

To account for the error when going from $\widehat{\mathfrak{I}}$ to $\mathfrak{I}$, the terms $2 \cdot d_{TV}(\widehat{\mathfrak{I}}, \mathfrak{I})$ and κ are introduced.

Note that $f^\star$ is only an approximately optimal fair classifier for ρ-Fair due to the additional error κ. Assume the optimal fair classifier for ρ-Fair is f^0. Since we do not have access to $\mathfrak{I}$ (only to $\widehat{\mathfrak{I}}$), it is unknown whether f^0 satisfies the τ-rule with respect to $\widehat{\mathfrak{I}}$. Hence, we can only compare the performance of the output f to $f^\star$, instead of the optimal classifier f^0. If the number of samples N is large, we can expect that $\widehat{\mathfrak{I}}$ and $\mathfrak{I}$ are close, and hence $\kappa, d_{TV}(\widehat{\mathfrak{I}}, \mathfrak{I})$ are small. Then the performance of f is close to $f^\star$ over $\mathfrak{I}$. Specifically, if $\widehat{\mathfrak{I}} = \mathfrak{I}$, we have $\kappa = d_{TV}(\widehat{\mathfrak{I}}, \mathfrak{I}) = 0$. The output classifier f then satisfies the properties of Theorem 4.1 with $\beta = 1$, which implies that f is an approximately optimal fair classifier for ρ-Fair.

Remark 4.5. *For the fairness metric δ_q (introduced in Remark 2.3), we can also design an algorithm similar to Algorithm 1. We only need to modify Line 2 by $L := \lceil \frac{1+\tau}{\varepsilon} \rceil$ (recall $\tau \in [-1, 0]$), $a_i := (i-1)\cdot\varepsilon$ and $b_i := i \cdot \varepsilon - \tau$. The quantification of the output cf is similar to Theorem 4.4. The main differences are*

$$f^\star := \arg \min_{f\in\mathcal{F}: \delta_q(f) \ge \tau + \kappa} \Pr_{\mathfrak{I}}[f \ne Y],$$

and the output f satisfies that 1) $\Pr_{\mathfrak{I}}[f \ne Y] \le \Pr_{\mathfrak{I}}\left[f^\star \ne Y\right] + 2 \cdot d_{TV}(\mathfrak{I}, \widehat{\mathfrak{I}})$; 2) $\delta_q(f) \ge \tau - \varepsilon - \kappa$. The details are discussed in Appendix D.

Remark 4.6. *Since the distribution $\widehat{\mathfrak{I}}$ is constructed via samples from $\mathfrak{I}$, we can study the number of samples required such that $\widehat{\mathfrak{I}}$ and $\mathfrak{I}$ are* close *enough, i.e.,*

$$d_{TV}(\mathfrak{I}, \widehat{\mathfrak{I}}) \le \varepsilon/2, \qquad \kappa = 2 \max_{i\in[p], f\in\mathcal{F}} \left| q_i^{\widehat{\mathfrak{I}}}(f) - q_i^{\mathfrak{I}}(f) \right| \le \varepsilon. \quad (5)$$

Note that given the inequality $d_{TV}(\mathfrak{I}, \widehat{\mathfrak{I}}) \le \varepsilon/2$, learning $\widehat{\mathfrak{I}}$ is exactly a classic distribution learning problem in which the sample complexity is bounded under a certain assumption model of $\mathfrak{I}$, e.g., mixtures of a constant number of Gaussian distributions [48]. We refer interested readers to the survey [19] for distribution estimation techniques.

For the second inequality, when the constraint $q_i^{\mathfrak{I}}(f)$ is linear, we can use Chernoff bound to show that given $O(N \ln p/\alpha)$ samples from the underlying distribution $\mathfrak{I}$, there exists an algorithm that computes an estimated distribution $\widehat{\mathfrak{I}}$ such that

$$\kappa = 2 \max_{i\in[p], f\in\mathcal{F}} \left| q_i^{\widehat{\mathfrak{I}}}(f) - q_i^{\mathfrak{I}}(f) \right| \le \varepsilon,$$

where $\alpha := \min_{i\in[p]} \Pr_{\mathfrak{I}}[G_i, \mathcal{E}']$. The formal statement and discussion on sample complexity are presented in Appendix C.2.

5 EMPIRICAL EVALUATION

5.1 Experimental Setup

We compare the empirical performance of our algorithm against the state-of-the-art techniques for fair classification on three datasets that are commonly used to evaluate the fairness of algorithms.

5.1.1 Algorithms and Benchmarks. We compare three versions of Algorithm 1, which provide fair classification results with respect to different fairness metrics:

- Subject to τ_{sr}-fair (**Algo 1-SR**), i.e., fairness constraint with respect to the statistical rate;
- Subject to τ_{fdr}-fair (**Algo 1-FDR**), i.e., fairness constraint with respect to the false discovery rate (which is a kind of predictive parity constraint);
- Subject to τ_{sr}-fair and τ_{fdr}-fair (**Algo 1-SR+FDR**), i.e., fairness constraints with respect to both the statistical rate and the false discovery rate.

We benchmark our approach against four state-of-the-art algorithms:

- **COV** developed in [60] aims to constrain statistical rate (τ_{sr});
- **SHIFT** developed in [34] designed to ensure equalized odds (constrain τ_{fpr} and τ_{tpr});
- **FPR-COV** and **FNR-COV** presented in [59], aim to eliminate disparate mistreatment (control the ratios τ_{fpr} and τ_{fnr}).
- **REDUCTION** developed in [3], designed to constrain statistical parity and equalized odds (constrain τ_{sr}, τ_{fpr}, τ_{tpr}).[3]

We select **COV**, **FPR-COV**, **FNR-COV** and **SHIFT** to compare against because they are state-of-the-art algorithms for their respective fairness metrics. We also select **REDUCTION** to compare against because it provides a different meta-algorithm which works for a subset of the fairness metrics we consider.

5.1.2 Measurements. Let D denote the empirical distribution over the testing set. Given a group performance function q, we denote γ_q to be the fairness metric ρ_q under the empirical distribution D. For instance, given a classifier f,

$$\gamma_{\text{sr}}(f) := \min_{i\in[p]} \Pr_D[f = 1 \mid Z = i] \,/ \max_{i\in[p]} \Pr_D[f = 1 \mid Z = i].$$

γ_q represents the fairness of the output classifier over the testing set, while τ_q represent the input fairness constraint desired with respect to the underlying distribution. Ideally $\gamma_q \ge \tau_q$. However this may not always be satisfied in practice if the estimated distribution $\widehat{\mathfrak{I}}$ is not a good fit for the underlying distribution $\mathfrak{I}$. Hence we report γ_q as this is the output fairness obtained by the classifier. For completeness, we also report the correspondence between the output fairness γ_q and the input constraint τ_q.

5.1.3 Datasets. We conduct our experiments on the following three datasets, which are commonly used for benchmarking in the fairness literature:

- **Adult**: This is an income dataset [18], which records the demographics of 45222 individuals, along with a binary label indicating whether the income of an individual is greater than 50k USD. We use the pre-processed dataset[4] that was also used by Zafar et al. [60]. We take gender to be the sensitive attribute, which is binary in the dataset.
- **German**: This dataset [18], records the attributes corresponding to around 1000 individuals with a label indicating positive or negative credit risk. We use the pre-processed dataset[5] provided by Friedler et al [28]. We take gender to be the sensitive attribute, which is binary in the dataset.
- **COMPAS**: This dataset [4], compiled by Propublica, is a list of demographic data of criminal offenders along with a risk score. We refer the reader to [43] for more details on how the data was analysed and compiled. For our experiment, we use the following features for classification : 'sex', 'age', 'race', 'juvenile felony count', 'decile score', 'juvenile misdemeanor count', 'other juvenile charges count', 'priors count', 'days in jail', 'charge degree', and try to predict the 'is recid' label, which represents whether individuals recidivated within two years or not. We take race as the sensitive attribute, and consider the subset of the data corresponding to individuals for which the race attribute is either black or white.

5.1.4 Implementation Details. We perform five repetitions, in which we divide the dataset uniformly at random into training (70%) and testing (30%) sets and report the average statistics of the above algorithms. In Algorithm 1, we set the error parameter ε to 0.01, and fit the estimated distribution $\widehat{\mathfrak{I}}$ using Gaussian Naive Bayes using SciPy [23]. For each dataset we run **Algo 1-SR** and **Algo 1-FDR** for $\tau \in \{0.1, 0.2, \ldots, 1.0\}$, and plot the resulting γ_{sr} and accuracy. We solve the optimization problem using Gradient Descent methods.

5.2 Results

5.2.1 Accuracy vs Output Fairness on the Adult Datset. Fig. 1 summarizes the tradeoff between the accuracy and the observed fairness γ_{sr} with respect to the statistical rate. The red points represents the mean value of γ_{sr} and accuracy of Algo 1 for different input values of τ_{sr}, with the error bars representing the standard deviations respectively. For other algorithms, we report only the point with largest mean γ_{sr} value and the axes of the ellipse around the point are the standard deviations of the fairness and accuracy respectively. We observe that **Algo 1-SR** can achieve higher γ_{sr} than than other methods. However, this gain in fairness comes at a loss; accuracy is decreasing in γ_{sr} for **Algo 1-SR** (albeit always above 75%). Even for lower values of γ_{sr} (corresponding to weaker constraints

[3] We use the code provided by the authors (https://github.com/Microsoft/fairlearn) which uses the Least squares classifier as the base classifier.

[4] https://github.com/mbilalzafar/fair-classification

[5] https://github.com/algofairness/fairness-comparison

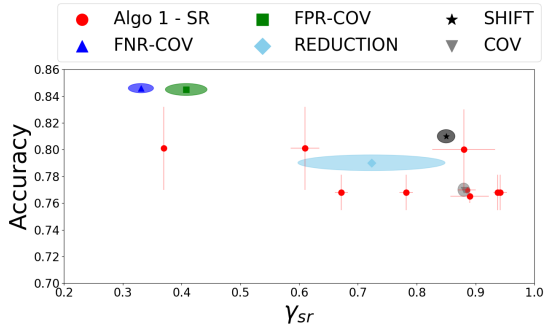

Figure 1: Acc. vs. γ_{sr} for Adult dataset. For Algo 1-SR, we plot the mean value of accuracy and observed fairness γ_{sr} for different values of input τ_{sr}. For other methods, we plot the datapoint with largest mean γ_{sr} and the ellipse around it represents the standard deviation. Algo 1-SR can achieve better fairness with respect to SR than any other method, albeit at a loss to accuracy.

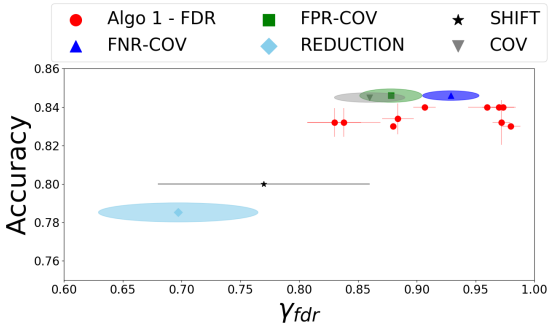

Figure 2: Acc. vs. γ_{fdr} for Adult dataset. For Algo 1-FDR, we plot the mean value of accuracy and observed fairness γ_{fdr} for different values of input τ_{fdr}. For other methods, we plot the datapoint with largest mean γ_{fdr} and the ellipse around it represents the standard deviation. Algo 1-FDR achieves better fairness with respect to FDR than any other method and the loss in accuracy is small.

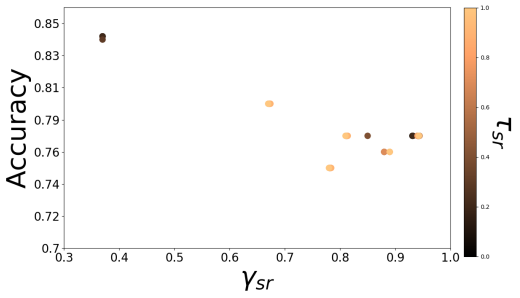

Figure 3: Acc. vs. γ_{sr}. Algo 1-SR for different values of input τ_{sr} on Adult dataset.

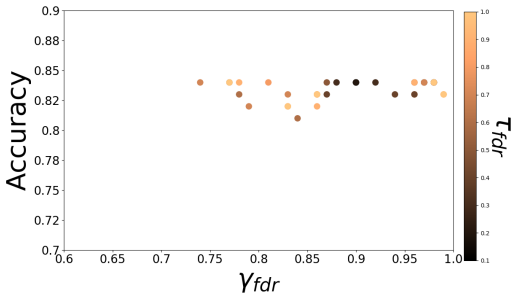

Figure 4: Acc. vs. γ_{fdr}. Algo 1-FDR for different values of input τ_{fdr} on Adult dataset.

τ_{sr}), the accuracy is worse than that of **COV** and **SHIFT**. This is likely due to the fact that we use a simple model for estimating the empirical distribution $\mathfrak{I}$, which will affect the overall accuracy of the algorithm (see Theorem 4.4); we expect that the performance would improve if we were to tune the fit (see also Section H.2). [6]

Similarly, Fig. 2 summarizes the tradeoff between the accuracy and the observed fairness γ_{fdr} with respect to the false discovery rate. The red points represents the mean value of γ_{fdr} and accuracy of Algo 1 for different input values of τ_{fdr}, while we report only the point with largest mean γ_{fdr} for the other algorithms. Here we observe that **Algo 1-FDR** can attain the highest observed fairness γ_{fdr} (for appropriate input values τ_{fdr}. Furthermore its accuracy, even for the highest fairness values, is comparable to that of other methods. Note that the overall fairness γ_{fdr} for all methods is higher than γ_{sr}; this is likely because the unconstrained optimal classifier for Algorithm 1 achieves $\gamma_{\text{fdr}} = 0.84$ (see Table 2), i.e., the **Adult** dataset is already relatively fair across genders with respect to FDR.

Quadrianto and Sharmanska [55] also provide a heuristic meta-algorithm for multiple fairness metrics. However, we were unable to compare against their approach directly due to the unavailability of their code online and our inability to replicate their results via our own implementation. Comparing against the raw numbers reported in their paper, they achieve 0.81 accuracy overall while the accuacy-difference (their metric of fairness) across groups is ~ 0.05 on the Adult dataset. **Algo 1-SR** can achieve a similar accuracy-difference (~ 0.05) for an overall accuracy of 0.80 and can achieve a smaller accuracy-difference (~ 0.02) for overall accuracy ~ 0.78.

5.2.2 Relationship Between τ and γ on the Adult Dataset. Empirically, we find that the observed fairness (γ) is almost always close to the target constraint. The output fairness and accuracy of the classifier against the input measure τ is depicted in Fig. 3 and Fig. 4. We plot all points from all the training/test splits in these figures.

5.2.3 Results on COMPAS and German datasets. A similar evaluation on the COMPAS [4] and German dataset [18] are presented in Appendix H, and we simply summarize the primary observations here. The performance of **Algo 1-FDR** with respect to other algorithms on German dataset is depicted in Figures 2 and 4 . From Fig 4, we observe that the classifier is able to satisfy the input fairness constraint almost every time, i.e., for almost all values of input τ_{fdr}, the observed fairness of the classifier, γ_{fdr}, is greater than or almost equal to τ_{fdr}. Furthermore, as shown in Fig 2, the maximum γ_{fdr}

[6] Note that the trade-offs in these figures appear non-monotone because they represent the average results for all five training-test splits of the dataset. Within each partition, they are monotone.

Table 2: The performance (mean and standard deviation in parens), of different fair classification alorithms with respect to accuracy and the fairness metrics from γ_q in Table 1 on the Adult dataset. We present the performance of an unconstrained optimal classifier for Algorithm 1 for comparison.

			Fairness Metrics									
		Acc.	γ_{sr}	γ_{fpr}	γ_{fnr}	γ_{tpr}	γ_{tnr}	γ_{ar}	γ_{fdr}	γ_{for}	γ_{ppr}	γ_{npr}
	Unconstrained	0.83 (0.00)	0.33 (0.03)	0.30 (0.02)	**0.87** (0.05)	0.86 (0.06)	0.94 (0.00)	0.86 (0.01)	**0.84** (0.07)	0.34 (0.03)	**0.93** (0.03)	0.87 (0.01)
This paper	**Algo 1-SR**	0.77 (0.01)	**0.89** (0.05)	0.51 (0.04)	0.55 (0.10)	0.81 (0.03)	0.82 (0.02)	**0.90** (0.02)	0.46 (0.03)	0.21 (0.04)	0.39 (0.04)	0.88 (0.00)
This paper	**Algo 1-FDR**	0.83 (0.00)	0.32 (0.04)	0.27 (0.05)	0.78 (0.07)	0.86 (0.06)	0.88 (0.01)	**0.89** (0.05)	**0.85** (0.03)	0.36 (0.03)	**0.93** (0.04)	0.89 (0.00)
This paper	**Algo 1-SR+FDR**	0.44 (0.13)	**0.84** (0.04)	**0.83** (0.09)	0.21 (0.27)	**0.96** (0.01)	0.36 (0.37)	0.48 (0.26)	0.70 (0.04)	0.15 (0.16)	0.34 (0.06)	**0.95** (0.03)
Baselines	**COV** [60]	0.79 (0.28)	0.83 (0.01)	0.63 (0.06)	0.27 (0.19)	0.76 (0.07)	0.79 (0.10)	0.81 (0.06)	0.55 (0.12)	0.10 (0.05)	0.44 (0.11)	0.86 (0.02)
Baselines	**FPR-COV** [59]	**0.85** (0.01)	0.41 (0.07)	0.39 (0.08)	**0.87** (0.10)	**0.91** (0.07)	0.94 (0.01)	**0.88** (0.01)	0.80 (0.08)	0.29 (0.05)	0.91 (0.04)	0.87 (0.02)
Baselines	**FNR-COV** [59]	**0.85** (0.01)	0.22 (0.05)	0.14 (0.04)	0.61 (0.09)	0.67 (0.10)	0.89 (0.01)	**0.88** (0.04)	0.80 (0.05)	**0.50** (0.05)	0.92 (0.02)	0.91 (0.01)
Baselines	**SHIFT** [34]	0.81 (0.01)	0.50 (0.11)	0.40 (0.16)	**0.90** (0.06)	0.84 (0.09)	0.98 (0.00)	0.83 (0.01)	**0.84** (0.06)	0.31 (0.02)	**0.96** (0.02)	0.82 (0.01)
Baselines	**REDUCTION** [3]	0.79 (0.00)	**0.86** (0.06)	0.69 (0.10)	0.74 (0.02)	0.43 (0.03)	**0.99** (0.00)	0.80 (0.01)	0.59 (0.11)	0.27 (0.03)	**0.91** (0.03)	0.79 (0.01)

value achieved by **Algo 1-FDR** is around 0.99, while amongst other algorithms, the maximum achieved is around 0.85. Similarly for **Algo 1-SR**, whose results are presented in Figures 1 and 3, we see that for almost all values of input τ_{sr}, we satisfy the input fairness constraint (except when $\tau_{\text{sr}} \sim 1$, in which case observed $\gamma_{\text{sr}} \sim 0.98$).

Figures 6 and 8 depict the performance of **Algo 1-FDR** with respect to other algorithms on the COMPAS dataset. **Algo 1-FDR** achieves a maximum γ_{fdr} of around 0.99, while other algorithms are able to achieve γ_{fdr} value around 0.9. For lower values of input τ_{fdr}, we achieve similar accuracy as other methods (∼0.70), however for higher values of input τ_{fdr}, we incur a loss in accuracy (∼ 0.68). Similarly the performance of **Algo 1-SR** is presented in Figures 5 and 7. Once again, the input fairness constraint is almost always satisfied, and we achieve higher γ_{fdr} values than other algorithms.

5.2.4 Effect of Constraints on Other Fairness Metrics. We also examine the performance of our methods and the baselines with respect to other fairness metrics γ_q and report their mean and standard deviation. For **Algo 1**-SR, **Algo 1-SR+FDR**, **COV** and **REDUCTION**, we consider only classifiers corresponding to $\gamma_{\text{sr}} \geq 0.8$, while for **Algo 1-FDR**, **FPR-COV**, **FNR-COV** and **SHIFT**, we choose the classifier corresponding to $\gamma_{\text{fdr}} \geq 0.8$. Different methods are better at optimizing different fairness metrics – the key difference is that Algo 1 can optimize different metrics depending on the given parameters, whereas other methods do not have this flexibility; e.g., here we constrain fairness with respect to SR and FDR (for which the maximal values of γ_{sr} and γ_{fdr} are attained), but we could instead constrain with respect to any other q if desired. Interestingly, although **Algo 1-SR** and **Algo 1-FDR** do not achieve the highest accuracy overall, both have significantly higher accuracy parity than other methods ($\gamma_{\text{ar}} \approx 0.9$). Furthermore, we can consider multiple fairness constraints simultaneously; **Algo 1-SR+FDR** can achieve both $\gamma_{\text{sr}} > 0.7$ and $\gamma_{\text{fdr}} > 0.7$, while remaining methods can not ($\gamma_{\text{sr}} < 0.45$ or $\gamma_{\text{fdr}} < 0.55$). Unfortunately, this does come at a loss of accuracy, likely due to the difficulty of simultaneously achieving accuracy and multiple fairness metrics [14, 41].

6 CONCLUSION

We present an efficient meta-algorithm for classification with (non-convex) linear-fractional constraints. Linear-fractional constraints capture many existing fairness definitions in the literature, and thus our algorithm can be used to derive several old and new results for classification with fairness constraints. In particular, to the best of our knowledge, our framework is the first that works for predictive parity with provable guarantees, which addresses an open problem proposed in [60]. Empirical evaluation of our algorithm on real-world datasets shows that our algorithm almost always satisfies the fairness constraints and the loss in accuracy is small.

This paper opens several possible directions for future work. As observed in the empirical results (and predicted by Theorem 4.4), the performance of our framework depends on the quality of the estimated distribution $\widehat{\mathfrak{I}}$. It would be interesting to optimize the approach in this regard, either empirically or theoretically. We also believe it would be valuable to extend this framework to other commonly used loss functions (e.g., l_2-loss or AUC) and other classifiers (e.g., margin-based classifiers or score-based classifiers). It would be interesting to get bounds on sample complexity for classification with linear fractional constraints. Finally, while in this paper we consider fairness constraints introduced by the τ-rule, other fairness constraints such as AUC and correlation (see the survey [63]) might also be worth considering.

REFERENCES

[1] ACM. 2017. Statement on Algorithmic Transparency and Accountability. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.
[2] An Act. 1964. Civil Rights Act. *Title VII, Equal Employment Opportunities* (1964).
[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*. 60–69.
[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. https://github.com/propublica/compas-analysis.
[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May* (2016).
[6] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review* (2016).
[7] Richard Berk. 2009. The role of race in forecasts of violent crime. *Race and social problems* (2009).
[8] Stephen Boyd and Almir Mutapcic. 2008. Stochastic subgradient methods. *Lecture Notes for EE364b, Stanford University* (2008).
[9] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* 21, 2 (2010), 277–292.
[10] L Elisa Celis, Lingxiao Huang, and Nisheeth K Vishnoi. 2018. Multiwinner voting with fairness constraints. In *Proceedings of the Twenty-seventh International Joint Conference on Artificial Intelligence and the Twenty-third European Conference on Artificial Intelligence, IJCAI-ECAI*.
[11] L. Elisa Celis, Vijay Keswani, Amit Deshpande, Tarun Kathuria, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Fair and Diverse DPP-based Data Summarization. In *Proceedings of the 35th International Conference on Machine Learning, ICML*

2018.
[12] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. In *Proceedings of the fourty-fifth International Colloquium on Automata, Languages, and Programming ICALP.*
[13] L. Elisa Celis and Nisheeth K Vishnoi. 2017. Fair Personalization. In *Fairness, Accountability, and Transparency in Machine Learning.*
[14] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[15] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017.* 797–806.
[16] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies* (2015).
[17] Bill Dedman et al. 1988. The Color of Money. *Atlanta Journal-Constitution* (1988).
[18] Dua Dheeru and Efi Karra Taniskidou. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.
[19] Ilias Diakonikolas. 2016. Learning Structured Distributions. *Handbook of Big Data* 267 (2016).
[20] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc* (2016).
[21] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012.* ACM, 214–226.
[22] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Mark D. M. Leiserson. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *Fairness, Accountability, and Transparency in Machine Learning.* 119–133.
[23] ENTHOUGHT. 2018. SciPy. https://www.scipy.org/.
[24] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.* ACM, 259–268.
[25] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining.* SIAM, 144–152.
[26] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. 2016. 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. *Fed. Probation* 80 (2016), 38.
[27] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (Im) Possibility of Fairness. *arXiv preprint arXiv:1609.07236* (2016).
[28] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2018. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. *arXiv preprint arXiv:1802.04422* (2018).
[29] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-Discriminatory Machine Learning through Convex Fairness Criteria. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.*
[30] Sharad Goel, Justin M Rao, Ravi Shroff, et al. 2016. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics* 10, 1 (2016), 365–394.
[31] Gabriel Goh, Andrew Cotter, Maya R. Gupta, and Michael P. Friedlander. 2016. Satisfying Real-world Goals with Dataset Constraints. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016.* 2415–2423.
[32] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018.* 903–912.
[33] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.*
[34] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016.* 3315–3323.
[35] Mara Hvistendahl. 2016. Can "predictive policing" prevent crime before it happens. *Science Magazine* 28 (2016).
[36] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems.* 325–333.
[37] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on.* IEEE, 1–6.
[38] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.
[39] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *In Proceedings of Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012.* 35–50.
[40] Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. 2017. Meritocratic fairness for cross-population selection. In *International Conference on Machine Learning.*
[41] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS, 2017.* 43:1–43:23.
[42] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018.* International World Wide Web Conferences Steering Committee.
[43] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9 (2016).
[44] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2011.* ACM, 502–510.
[45] Susan Magarey. 2004. The sex discrimination act 1984. *Australian Feminist Law Journal* (2004).
[46] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018.* 107–118.
[47] Claire Cain Miller. 2015. Can an algorithm hire better than a human. *The New York Times* 25 (2015).
[48] Ankur Moitra and Gregory Valiant. 2010. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on.* IEEE, 93–102.
[49] Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal. 2014. On the Statistical Consistency of Plug-in Classifiers for Non-decomposable Performance Measures. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014.* 1493–1501.
[50] Arvind Narayanan. 2018. Tutorial: 21 Fairness Definitions and Their Politics. https://www.youtube.com/watch?v=jIXIuYdnyyk.
[51] Northpointe. 2012. Compas risk and need assessment systems. http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf.
[52] United States. Executive Office of the President and John Podesta. 2014. *Big data: Seizing opportunities, preserving values.* White House, Executive Office of the President.
[53] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.* ACM, 560–568.
[54] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017.* 5684–5693.
[55] Novi Quadrianto and Viktoriia Sharmanska. 2017. Recycling Privileged Learning and Distribution Matching for Fairness. In *Advances in Neural Information Processing Systems.* 677–688.
[56] WhiteHouse. 2016. *Big data: A report on algorithmic systems, opportunity, and civil rights.* Executive Office of the President.
[57] Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017.* 1920–1953.
[58] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *SSDBM.*
[59] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017.* 1171–1180.
[60] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017.* 962–970.
[61] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, Krishna P. Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017.* 228–238.
[62] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013.* 325–333.
[63] Indre Zliobaite. 2017. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.* (2017).