Fair Classification with Noisy Protected Attributes: A Framework with Provable Guarantees

L. Elisa Celis [1] **Lingxiao Huang** [2] **Vijay Keswani** [1] **Nisheeth K. Vishnoi** [3]

Abstract

We present an optimization framework for learning a fair classifier in the presence of noisy perturbations in the protected attributes. Compared to prior work, our framework can be employed with a very general class of linear and *linear-fractional* fairness constraints, can handle multiple, *non-binary* protected attributes, and outputs a classifier that comes with provable guarantees on *both* accuracy and fairness. Empirically, we show that our framework can be used to attain either statistical rate or false positive rate fairness guarantees with a minimal loss in accuracy, even when the noise is large, in two real-world datasets.

1. Introduction

Fair classification has been a topic of intense study due to the growing importance of addressing social biases in automated prediction. Consequently, a host of fair classification algorithms have been proposed that learn from data (Bellamy et al., 2018a; Zafar et al., 2017b; Zhang et al., 2018; Menon & Williamson, 2018b; Goel et al., 2018; Celis et al., 2019; Hardt et al., 2016; Fish et al., 2016; Pleiss et al., 2017; Woodworth et al., 2017; Dwork et al., 2018).

Fair classifiers need metrics that capture the extent of similarity in performance for different groups. The performance of a classifier f for group z can be defined in many ways and a general definition that captures the most common group performance metrics in literature is the following: given events ξ and ξ' (that depend on classifier and/or class label Y), the performance for group z can be quantified as $\Pr[\xi \mid \xi', z]$ (Defn 2.1). For instance, we get group-specific statistical rate (a linear metric) by setting $\xi := (f{=}1)$ and $\xi' := \emptyset$ and group-specific false discovery rate (a "linear-fractional metric" (Celis et al., 2019)) by setting $\xi := (Y{=}0)$ and $\xi' := (f{=}1)$. Then, given a performance function, fairness constraints in classification impose a feasible classifier to have similar performance for all groups, by constraining the performance difference to be within τ of each other either multiplicatively or additively. For any fixed group performance function, multiplicative constraints imply additive constraints and, hence, are traditionally studied (Calders & Verwer, 2010; Zafar et al., 2017a;b; Menon & Williamson, 2018a; Celis et al., 2019) (see also Remark 2.2).

The choice of fairness metric depends on the context and application. For instance, in a lending setting, statistical rate metric can capture the disparity in loan approval rate across gender/race (Comptroller, 2010). In a recidivism assessment setting, false positive rate metric is more relevant as it captures the disparity in proportion to defendants falsely assigned high-risk across racial groups (Angwin et al., 2016a). In other settings, e.g., healthcare, where the costs associated with positive classification are large, false discovery rate is alternately employed to assess the disparity in proportion to the treated patients who didn't require treatment across protected attribute types (Srivastava et al., 2019).

Most of the aforementioned fair classification algorithms crucially assume that one has access to the protected attributes (e.g., race, gender) for training and/or deployment. Data collection, however, is a complex process and may contain recording and reporting errors, unintentional or otherwise (Saez et al., 2013). Cleaning the data also requires making difficult and political decisions along the way, yet is often necessary especially when it comes to questions of race, gender, or identity (Nobles, 2000). Further, information about protected attributes may be missing entirely (Data et al., 2004), or legally prohibited from being used directly, as in the case of lending applications for non-mortgage products in US (Federal Reserve Bank, 1993). In such cases, protected attributes can be predicted from other data, however, we know that this process is itself contains errors and biases (Muthukumar et al., 2018; Buolamwini & Gebru, 2018). The above scenarios raise a challenge for existing fair classifiers as they may not achieve the same fairness as they would if the data were perfect. This raises the question of learning fair classifiers in the presence of noisy protected attributes, and has attracted recent attention (Awasthi et al., 2020; Lamy et al., 2019; Wang et al., 2020).

[1]Department of Statistics and Data Science, Yale University, USA [2]Tsinghua University, China [3]Department of Computer Science, Yale University, USA. Correspondence to: L. Elisa Celis <elisa.celis@yale.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Our contributions. We study the setting of "flipping noises" where a protected type $Z = i$ may be flipped to $\hat{Z} = j$ with some known fixed probability H_{ij} (Definition 2.3). We present an optimization framework for learning a fair classifier that can handle: 1) **flipping noises** in the train, test, and future samples, 2) multiple, **non-binary** protected attributes, and 3) multiple fairness metrics, including the general class of **linear-fractional** metrics (e.g., statistical parity, false discovery rate) in a multiplicative sense. Our framework can learn a near-optimal fair classifier on the underlying dataset with high probability and comes with provable guarantees on both accuracy and fairness.

We implement our framework using the logistic loss function (Freedman, 2009) and examine it on **Adult** and **COMPAS** datasets (Section 4). We consider sex and race as the protected attribute and generate noisy datasets varying flipping noise parameters. For **COMPAS** dataset, the race protected attribute is non-binary. We use statistical rate, false positive rate, and false discovery rate fairness metrics, and compare against natural baselines and existing noise-tolerant fair classification algorithms (Lamy et al., 2019; Awasthi et al., 2020; Wang et al., 2020). The empirical results show that, for most combinations of dataset and protected attribute (both binary and non-binary), our framework attains better fairness than an unconstrained classifier, with a minimal loss in accuracy. Further, in most cases, the fairness-accuracy tradeoff of our framework, for statistical and false positive rate, is also better than the baselines and other noise-tolerant fair classification algorithms, which either do not always achieve high fairness levels or suffer a larger loss in accuracy for achieving high fairness levels compared to our framework (Table 1). For false discovery rate (linear-fractional metric), our approach has better fairness-accuracy tradeoff than baselines for **Adult** dataset and similar tradeoff as the best-performing baseline for **COMPAS** dataset (Table 2).

Techniques. Our framework starts by designing *denoised* constraints to achieve the desired fairness guarantees which take into account the noise in the protected attribute (Program **DFair**). The desired fairness is governed using an input parameter $\tau \in [0, 1]$. The key is to estimate each group-specific performance on the underlying dataset, which enables us to handle non-binary protected attributes. Concretely, we represent a group-specific performance as a ratio and estimate its numerator and denominator separately, which enables us to handle linear-fractional constraints. Subsequently, we show that an optimizer f^{Δ} of our program is provably **both** approximately optimal and fair on the underlying dataset (Theorem 3.3) with high probability under a mild assumption that an optimizer $f^{\star}$ of the underlying program (Program **TargetFair**) has a non-trivial lower bound on the group-specific prediction rate (Assumption 1). The constraints in our program enable us to capture the range of alteration in the probability of any classifier prediction for different protected attribute types due to flipping noises and, consequently, allow us to provide guarantees on f^{Δ} (Theorem 3.3). The guarantee on accuracy uses the fact that an optimal fair classifier $f^{\star}$ for the underlying uncorrupted dataset is *likely* to be feasible for Program **DFair** as well, which ensures that the empirical risk of f^{Δ} is less than $f^{\star}$ (Lemma 3.6). The guarantee on the fairness of f^{Δ} is attained by arguing that classifiers that considerably violate the desired fairness guarantee are infeasible for Program **DFair** with high probability (Lemma 3.8). The key technical idea is to discretize the space of unfair classifiers by carefully chosen multiple ε-nets with different violation degrees to our denoised program, and upper bound the capacity of the union of all nets via a VC-dimension bound.

Related work. (Lamy et al., 2019) consider binary protected attributes, linear fairness metrics including statistical rate (SR) and equalized odds constraints (Donini et al., 2018). They give a provable algorithm that achieves an approximate optimal fair classifier by down-scaling the "fairness tolerance" parameter in the constraints to adjust for the noise. In contrast, our approach estimates the altered form of fairness metrics in the noisy setting, and hence, can also handle linear-fractional metrics and non-binary attributes. Awasthi et al. (2020) study the performance of the equalized odds post-processing method of Hardt et al. (2016) for a single noisy binary protected attribute. However, their analysis assumes that the protected attributes of test/future samples are uncorrupted. Our framework, instead, can handle multiple, non-binary attributes and noise in test/future samples. Wang et al. (2020) propose two robust optimization approaches to solve the noisy fair classification problem. The first one is to solve a min-max distributionally robust optimization (DRO) problem, which guarantees to find a classifier that satisfies fairness constraints on the underlying dataset. However, there is no provable guarantee on the accuracy of the learned classifier. For the second one, by proposing an iterative procedure to solve the arising min-max problem, they can only guarantee a stochastic classifier that is near-optimal w.r.t. accuracy and near-feasible w.r.t. fairness constraints on the underlying dataset *in expectation*, but not with high probability. However, their iterative procedure relies on a minimization oracle, which is not always computationally tractable and their practical algorithm does not share the guarantees of their theoretical algorithm for the output classifier. In contrast, our denoised fairness program ensures that the optimal classifier is deterministic, near-optimal w.r.t. **both** accuracy and near-feasible w.r.t. fairness constraints on the underlying dataset, with high probability. Additionally, we define performance disparity across protected attribute values as the ratio of the "performance" for worst and best-performing groups (multiplicative constraints), while existing works (Lamy et al., 2019; Awasthi

et al., 2020; Wang et al., 2020) define the disparity using the additive difference across the protected attribute values (additive constraints); see Remark D.1. See Section A for other related work omitted here due to space constraints.

2. The Model

Let $\mathcal{D} = \mathcal{X} \times [p] \times \{0,1\}$ denote the underlying domain ($p \geq 2$ is an integer). Each sample (X, Z, Y) drawn from $\mathcal{D}$ contains a protected attributes Z, a class label $Y \in \{0,1\}$, and non-protected features $X \in \mathcal{X}$. Here, we discuss a single protected attribute, $Z = [p]$, and generalize our model and results to multiple protected attributes in Section D. [1] We assume that X is a d-dimensional vector, for a given $d \in \mathbb{N}$, i.e., $\mathcal{X} \subseteq \mathbb{R}^d$. Let $S = \{s_a = (x_a, z_a, y_a) \in \mathcal{D}\}_{a\in[N]}$ be the (underlying, uncorrupted) dataset. Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ denote a family of all possible allowed classifiers. Given a loss function $L : \mathcal{F} \times \mathcal{D} \to \mathbb{R}_{\geq 0}$, the goal of unconstrained classification is to find a classifier $f \in \mathcal{F}$ that minimizes *the empirical risk* $\frac{1}{N}\sum_{a\in[N]} L(f, s_a)$.

Fair classification and fairness metrics. We consider the problem of classification for a general class of fairness metrics. Let D denote the empirical distribution over S, i.e., selecting each sample s_a with probability $1/N$.

Definition 2.1 (Linear/linear-fractional group performance functions (Celis et al., 2019)) *Given a classifier $f \in \mathcal{F}$ and $i \in [p]$, we call $q_i(f)$ the group performance of $Z = i$ if $q_i(f) = \Pr_D[\xi(f) \mid \xi'(f), Z = i]$ for some events $\xi(f), \xi'(f)$ that might depend on the choice of f. If ξ' does not depend on the choice of f, q is said to be **linear**; otherwise, q is said to be **linear-fractional**.*

At a high level, a classifier f is considered to be fair w.r.t. q if $q_1(f) \approx \cdots \approx q_p(f)$. Definition 2.1 is general and contains many fairness metrics. For instance, if $\xi := (f = 1)$ and $\xi' := \emptyset$, we have $q_i(f) = \Pr_D[f = 1 \mid Z = i]$ which is linear and called the statistical rate. If $\xi := (Y = 0)$ and $\xi' := (f = 1)$, we have $q_i(f) = \Pr_D[Y = 0 \mid f = 1, Z = i]$ which is linear-fractional and called the false discovery rate. See Table 1 in (Celis et al., 2019) for a comprehensive set of special cases. Given a group performance function q, we define $\Omega_q : \mathcal{F} \times \mathcal{D}^\star \to [0,1]$ to be $\Omega_q(f, S) := \min_{i\in[p]} q_i(f) / \max_{i\in[p]} q_i(f)$ as a specific fairness metric, where $\mathcal{D}^\star$ is the collection of all datasets S on the domain $\mathcal{D}$. Then we define the following fair classification problem: Given a group performance functions q and a threshold $\tau \in [0,1]$, the goal is to learn an (approximate) optimal fair classifier $f \in \mathcal{F}$ of the following program:

$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{a\in[N]} L(f, s_a) \quad s.t. \quad \Omega_q(f, S) \geq \tau. \tag{TargetFair}$$

For instance, we can set q to be the statistical rate and $\tau = 0.8$ to encode the 80% disparate impact rule (Biddle, 2006). Note that $\Omega_q(f, S) \geq 0.8$ is usually non-convex for certain q. Often, one considers a convex function as an estimate of $\Omega_q(f, S)$, for instance $\Omega_q(f, S)$ is formulated as a covariance-type function in (Zafar et al., 2017b), and as the weighted sum of the logs of the empirical estimate of favorable bias in (Goel et al., 2018).

Remark 2.2 (Multiplicative v.s. additive fairness constraints) *We note that the fairness constraints mentioned above (Ω_q) are **multiplicative** and appear in (Calders & Verwer, 2010; Zafar et al., 2017a;b; Menon & Williamson, 2018a; Celis et al., 2019). Multiplicative fairness constraints control disparity across protected attribute values by ensuring that the ratio of the "performance" for the worst and best-performing groups are close. In contrast, related prior work for noisy fair classification (Lamy et al., 2019; Awasthi et al., 2020; Wang et al., 2020) usually consider **additive** fairness constraints, i.e., of the form $\Omega'_q(f, S) := \max_{i\in[p]} q_i(f) - \min_{i\in[p]} q_i(f) \leq \tau'$ for some $\tau' \in [0,1]$ (difference instead of ratio). Specifically, letting $\tau' = 0$ in the additive constraint is equivalent to letting $\tau = 1$ in the multiplicative constraint with respect to the same group performance function q. Note that **multiplicative implies additive**, i.e., given $\tau \in [0,1]$, we have that $\Omega(f, S) \geq \tau$ implies that $\Omega'(f, S) \leq 1 - \tau$. However, the converse is not true: for instance, given arbitrary small $\tau' > 0$, we may learn a classifier $f^\star$ under additive constraints such that $\min_{i\in[p]} q_i(f^\star) = 0$ and $\max_{i\in[p]} q_i(f^\star) = \tau'$; however, such $f^\star$ violates the 80% rule (Biddle, 2006) that is equivalent to $\Omega_q(f, S) \geq 0.8$.*

Noise model. If S is observed, we can directly use Program **TargetFair**. However, as discussed earlier, the protected attributes in S may be imperfect and we may only observe a noisy dataset $\widehat{S}$ instead of S. We consider the following noise model on the protected attributes (Lamy et al., 2019; Awasthi et al., 2020; Wang et al., 2020).

Definition 2.3 (Flipping noises) *Let $H \in [0,1]^{p\times p}$ be a stochastic matrix with $\sum_{j\in[p]} H_{ij} = 1$ and $H_{ii} > 0.5$ for any $i \in [p]$. Assume each protected attribute $Z = i$ ($i \in [p]$) is observed as $\widehat{Z} = j$ with probability H_{ij}, for any $j \in [p]$.*

Note that H can be **non-symmetric**. The assumption that $H_{ii} > 0.5$ ensures that the total flipping probability of each protected attribute is strictly less than a half. Consequently,

[1] One can also consider combinations of demographic groups by treating the intersections (e.g., of race and sex) as disjoint groups and using our program accordingly.

H is a diagonally-dominant matrix, which is always non-singular (Horn & Johnson, 2012). We will use the entries of H to design our fairness constraints; this case of when H is known and the flipping noise is i.i.d. arises in important real-world applications, such as the randomized response model (including local differential privacy (Duchi et al., 2013)). Additionally, we also discuss extensions to other noise models in Section 3.2.

Due to noise, directly applying the same fairness constraints on $\widehat{S}$ may introduce bias on S and, hence, modifications to the constraints are necessary; see Section F for a discussion.

Remark 2.4 (Limitation of Definition 2.3) *In certain applications, we may not know H explicitly, and can only estimate them by, say, finding a small appropriate sample of the data for which ground truth is known (or can be found), and computing estimates for H accordingly (Kallus et al., 2020). For instance, H could be inferred from prior data that contains both true and noisy (or proxy) protected attribute values; e.g., existing methods, such as Bayesian Improved Surname Geocoding method (Elliott et al., 2009), employ census data to construct conditional race membership probability models given surname and location. In the following sections, we assume H is given. For settings in which the estimates of H may not be accurate, we analyze the influences of the estimation errors at the end of Section 3.2 and empirically in Appendix E.*

Problem 1 (Fair classification with noisy protected attributes) *Given a group performance functions q, a threshold $\tau \in [0,1]$, and a noisy dataset $\widehat{S}$ with noise matrix H, the goal is to learn an (approximate) optimal fair classifier $f \in \mathcal{F}$ of Program* **TargetFair**.

3. Framework and Theoretical Results

We show how to learn an approximately fair classifier w.h.p. for Problem 1 (Theorem 3.3). This result is generalized to multiple protected attributes/fairness metrics in Section D. The approach is to design *denoised fairness constraints* over $\widehat{S}$ (Definition 3.1) that estimate the underlying constraints of Program **TargetFair**, and solve the constrained optimization problem (Program **DFair**). Let $f^\star \in \mathcal{F}$ denote an optimal classifier of Program (**TargetFair**). Our result relies on a natural assumption on $f^\star$.

Assumption 1 *There exists a constant $\lambda \in (0, 0.5)$ such that $\min_{i\in[p]} \Pr_D[\xi(f^\star), \xi'(f^\star), Z = i] \geq \lambda$.*

Note that λ is a lower bound for $\min_{i\in[p]} q_i(f^\star)$. In many applications we expect this assumption to hold. For instance, $\lambda \geq 0.1$ if there are at least 20% of samples with $Z = i$ and $\Pr_D[\xi(f^\star), \xi'(f^\star) \mid Z = i] \geq 0.5$ for each $i \in [p]$. In practice, exact λ is unknown but we can set λ according to the context. This assumption is not strictly necessary, i.e., we can simply set $\lambda = 0$, but the scale of λ decides certain capacity of classifiers that we do not want to learn, which affects the performance (success probability) of our approaches; see Remark 3.4.

3.1. Our Optimization Framework

Let $\widehat{D}$ denote the empirical distribution over $\widehat{S}$. Let $\widehat{u}(f) := \left(\Pr\left[\xi(f), \xi'(f), \widehat{Z} = i\right]\right)_{i\in[p]}$ and $\widehat{w}(f) := \left(\Pr\left[\xi'(f), \widehat{Z} = i\right]\right)_{i\in[p]}$. If D and $\widehat{D}$ are clear from the context, we denote $\Pr_{D,\widehat{D}}[\cdot]$ by $\Pr[\cdot]$. Let $M := \max_{i\in[p]} \|(H^\top)_i^{-1}\|_1$ where $(H^\top)_i^{-1}$ means first invert H^T and then take the i-th row. Define the denoised fairness constraints and the induced program as follows.

Definition 3.1 (Denoised fairness constraints) *Given a classifier $f \in \mathcal{F}$, for $i \in [p]$ let $\Gamma_i(f) := \frac{(H^\top)_i^{-1}\widehat{u}(f)}{(H^\top)_i^{-1}\widehat{w}(f)}$. Let $\delta \in (0,1)$ be a fixed constant and $\tau \in [0,1]$ be a threshold. We define our denoised fairness program to be*

$$\begin{aligned} &\min_{f\in\mathcal{F}} \frac{1}{N} \sum_{a\in[N]} L(f, \widehat{s}_a) \ \ \textit{s.t.} \\ &(H^\top)^{-1}\widehat{u}(f) \geq (\lambda - M\delta)\mathbf{1}, \\ &\min_{i\in[p]} \Gamma_i(f) \geq (\tau - \delta) \cdot \max_{i\in[p]} \Gamma_i(f). \end{aligned} \tag{DFair}$$

δ is used as a relaxation parameter depending on the context where a larger value of δ increases the probability of learning an (approximate) fair classifier by Program (**DFair**) (Remark 3.4). By definition, we can regard M as a metric that measures how noisy H is. Note that the term $(\lambda - M\delta)$ does not need to be nonnegative. Intuitively, as diagonal elements H_{ii} increases, eigenvalues of H increase, and hence, M decreases. Also note that $M \geq 1$ since M is at least the largest eigenvalue of H^{-1} and H is a non-singular stochastic matrix whose largest eigenvalue is 1. Intuitively, $\Gamma_i(f)$ is designed to estimate $\Pr[\xi(f) \mid \xi'(f), Z = i]$: its numerator approximates $\Pr[\xi(f), \xi'(f) \mid Z = i]$ and its denominator approximates $\Pr[\xi'(f), Z = i]$. For the denominator, since Definition 2.3 implies $\Pr\left[\widehat{Z} = j \mid \xi'(f), Z = i\right] \approx H_{ij}$, we can estimate $\Pr[\xi'(f), Z = i]$ by a linear combination of $\Pr\left[\xi'(f), \widehat{Z} = j\right]$, i.e., $(H^\top)_i^{-1}\widehat{w}(f)$. Similar intuition is behind the estimate of the numerator $(H^\top)_i^{-1}\widehat{u}(f)$. Due to how Γ_is are chosen, the first constraint is designed to estimate Assumption 1, and the last constraint is designed to estimate $\Omega_q(f, S) \geq \tau$. Recall that the λ parameter in Assumption 1 crucially affects the success probability of learning an approximate optimal classifier (Theorem 3.3). Thus, by incorporating Assumption 1 as a constraint in Program (**DFair**), we ensure that the classifier returned by Program (**DFair**) satisfies it.

This design ensures that an optimal fair classifier $f^\star$ satisfies our denoised constraints w.h.p. (Lemma 3.6), and hence, is a feasible solution to Program **DFair**. Consequently, the empirical risk of an optimal classifier f^Δ of Program **TargetFair** is at most that of $f^\star$. The main difficulty is to prove that f^Δ achieves fairness on the underlying dataset S, since an unfair classifier may also satisfy our denoised constraints and is output as a feasible solution of Program (**DFair**). To handle this, we show all unfair classifiers that are infeasible for Program **TargetFair** should violate our denoised constraints (Lemma 3.8). For this, we verify that the probability of each unfair classifier being feasible is exponentially small, and bound certain "capacity" of unfair classifiers (Definition B.5) using Assumption 1.

3.2. Main theorem: Performance of Program DFair

Our main theorem shows that solving Program **DFair** leads to a classifier that does not increase the empirical risk (compared to the optimal fair classifier) and only slightly violates the fairness constraint. Before we state our result, we need the following definition that measures the complexity of $\mathcal{F}$.

Definition 3.2 (VC-dimension of $(S, \mathcal{F})$ (Har-peled, 2011)) *Given a subset $A \subseteq [N]$, we define $\mathcal{F}_A := \{\{a \in A : f(s_a) = 1\} \mid f \in \mathcal{F}\}$ to be the collection of subsets of A that may be shattered by some $f \in \mathcal{F}$. The VC-dimension of $(S, \mathcal{F})$ is the largest integer t such that there exists a subset $A \subseteq [N]$ with $|A| = t$ and $|\mathcal{F}_A| = 2^t$.*

Suppose $\mathcal{X} \subseteq \mathbb{R}^d$ for some integer $d \geq 1$. If $\mathcal{F} = \{0,1\}^{\mathcal{X}}$, we observe that the VC-dimension is $t = N$. Several commonly used families $\mathcal{F}$ have VC-dimension $O(d)$, including linear threshold functions (Har-peled, 2011), kernel SVM and gap tolerant classifiers (Burges, 1998). Using this definition, the main theorem in this paper is as follows.

Theorem 3.3 (Performance of Program DFair) *Suppose the VC-dimension of $(S, \mathcal{F})$ is $t \geq 1$. Given any flipping noise matrix $H \in [0,1]^{p \times p}$, $\lambda \in (0, 0.5)$ and $\delta \in (0,1)$, let $f^\Delta \in \mathcal{F}$ denote an optimal fair classifier of Program **DFair**. With probability at least $1 - O(pe^{-\frac{\lambda^2\delta^2 n}{60000M^2} + t\ln(50M/\lambda\delta)})$, we have*

$$\frac{1}{N}\sum_{a\in[N]} L(f^\Delta, s_a) \leq \frac{1}{N}\sum_{a\in[N]} L(f^\star, s_a) \text{ and}$$

$$\Omega_q(f^\Delta, S) \geq \tau - 3\delta.$$

Theorem 3.3 indicates that f^Δ is an approximate fair classifier for Problem 1 with an exponentially small failure probability to the data size n. A few remarks are in order.

Remark 3.4 *Observe that the success probability depends on $1/M$, δ, λ and the VC-dimension t of $(S, \mathcal{F})$. If $1/M$ or δ is close to 0, i.e., the protected attributes are very noisy or there is no relaxation for $\Omega_q(f, S) \geq \tau$ respectively, the success probability guarantee naturally tends to be 0. Next, we discuss the remaining parameters λ and t.*

Discussion on λ. *Intuitively, the success probability guarantee tends to 0 when λ is close to 0. For instance, consider q to be the statistical rate (Eq. (1)). Suppose there is only one sample s_1 with $Z = 1$ for which $f^\star(s_1) = 1$, i.e., $\Pr_D[f^\star = 1, Z = 1] = 1/N$ and, therefore, $\lambda \leq 1/N$. To approximate $f^\star$, we may need to label $f(s_1) = 1$. However, due to the flipping noises, it is likely that we can not find out the specific sample s_1 to label $f(s_1) = 1$, unless we let the classifier prediction be $f = 1$ for all samples, which leads to a large empirical risk (see discussion in Section F.1). In other words, the task is tougher for smaller values of λ.*

Discussion on t. *The success probability also depends on t which captures the complexity of $\mathcal{F}$. Suppose $\mathcal{X} \subseteq \mathbb{R}^d$ for some integer $d \geq 1$. The worst case is $\mathcal{F} = \{0,1\}^{\mathcal{X}}$ with $t = N$, which takes the success probability guarantee to 0. On the other hand, if the VC-dimension does not depend on N, e.g., only depends on $d \ll N$, the failure probability is exponentially small on N. For instance, if $\mathcal{F}$ is the collection of all linear threshold functions, i.e., each classifier $f \in \mathcal{F}$ has the form $f(s_a) = \mathbf{I}[\langle x_a, \theta\rangle \geq r]$ for some vector $\theta \in \mathbb{R}^d$ and threshold $r \in \mathbb{R}$. We have $t \leq d+1$ for an arbitrary dataset S (Har-peled, 2011).*

Remark 3.5 *The learned classifier f^Δ guaranteed by our theorem is **both** approximately fair and optimal w.h.p. This is in contrast to learning a stochastic classifier $\tilde{f} \sim \Lambda$ over $\mathcal{F}$, that is in expectation near-optimal for both accuracy and fairness, e.g.,*

$$\mathbb{E}_{\tilde{f}\sim\Lambda}\left[\frac{1}{N}\sum_{a\in[N]} L(\tilde{f}, s_a)\right] \leq \frac{1}{N}\sum_{a\in[N]} L(f^\star, s_a), \text{ and}$$

$$\mathbb{E}_{f\sim\Lambda}\left[\Omega_q(\tilde{f}, S)\right] \geq \tau - 3\delta.$$

For instance, suppose $f_1, f_2 \in \mathcal{F}$ such that the empirical risk of f_1 is $\frac{3}{2N}\sum_{a\in[N]} L(f^\star, s_a)$ and $\Omega(f_1, S) = \tau/2$, while the empirical risk of f_2 is $\frac{1}{2N}\sum_{a\in[N]} L(f^\star, s_a)$ and $\Omega(f_2, S) = 3\tau/2$. If Λ is uniform over f_1 and f_2, it satisfies the above two inequalities, But, neither of f_is is near-optimal for accuracy and fairness.

Estimation errors. In practice, we can use prior work on noise parameter estimation (Menon et al., 2015; Liu & Tao, 2015; Northcutt et al., 2017) to obtain estimates of H, say H'. The scale of estimation errors also affects the performance of our denoised program. In Section C in the Supplementary Material, we provide a technical discussion on the effect of the estimation errors on the performance. Concretely, we consider a specific setting that $p = 2$ and q is the statistical rate. Define $\zeta := \max_{i,j\in[p]} |H_{ij} - H'_{ij}|$

to be the additive estimation error. We show there exists constant $\alpha > 0$ such that $\Omega_q(f^\Delta, S) \geq \tau - 3\delta - \zeta\alpha$ holds. Compared to Theorem 3.3, the estimation errors introduce an additive $\zeta\alpha$ error term for the fairness guarantee of our denoised program.

Extension to other noise models. Our framework can also be extended to other non-independent noise models; e.g. settings for which concentration bounds in Inequalities (4) and (5) in Section B.1 hold. These include negative associated or negative dependent noise. We could also extend to some non-identical noise settings, e.g., if the noise is identical across (large) subgroups of a group (as opposed to across the entire group) by treating each subgroup disjointly. Extensions to other noise models, however, seems non-trivial and an interesting future direction.

3.3. Proof of Theorem 3.3 for $p = 2$ and Statistical Rate

For ease of understanding, we consider a specific case in the main body: a binary sensitive attribute $Z \in \{0, 1\}$ and statistical rate constraints, i.e.,

$$\gamma(f, S) := \frac{\min_{i\in\{0,1\}} \Pr_D [f = 1 \mid Z = i]}{\max_{i\in\{0,1\}} \Pr_D [f = 1 \mid Z = i]} \geq \tau. \quad (1)$$

Consequently, we would like to prove $\gamma(f^\Delta, S) \geq \tau - 3\delta$ to obtain Theorem 3.3. The proof for the general Theorem 3.3 can be found in Section D. We denote $\eta_0 = H_{01}$ to be the probability that $\widehat{Z} = 1$ conditioned on $Z = 0$, and $\eta_1 = H_{10}$ to be the probability that $\widehat{Z} = 0$ conditioned on $Z = 1$. By Assumption 1, we have $\eta_0, \eta_1 < 0.5$. Combining with Definition 2.3, we have $H = \begin{bmatrix} 1-\eta_0 & \eta_0 \\ \eta_1 & 1-\eta_1 \end{bmatrix}$, which implies that $M = \frac{1}{1-\eta_0-\eta_1}$. Consequently, Assumption 1 is equivalent to $\min_{i\in\{0,1\}} \Pr_D [f^\star = 1, Z = i] \geq \lambda$, and for $i \in \{0, 1\}$,

$$\Gamma_i(f) := \frac{(1-\eta_{1-i}) \Pr\left[f = 1, \widehat{Z} = i\right] - \eta_{1-i} \Pr\left[f = 1, \widehat{Z} = 1-i\right]}{(1-\eta_{1-i})\widehat{\mu}_i - \eta_{1-i}\widehat{\mu}_{1-i}}.$$

We define the denoised statistical rate to be $\gamma^\Delta(f, \widehat{S}) := \min\left\{\frac{\Gamma_0(f)}{\Gamma_1(f)}, \frac{\Gamma_1(f)}{\Gamma_0(f)}\right\}$, and our denoised constraints become

$$\begin{cases} (1-\eta_{1-i}) \Pr\left[f=1, \widehat{Z}=i\right] - \eta_{1-i} \Pr\left[f=1, \widehat{Z}=1-i\right] \\ \geq (1-\eta_0-\eta_1)\lambda - \delta, \quad i \in \{0, 1\} \\ \gamma^\Delta(f, \widehat{S}) \geq \tau - \delta, \end{cases} \quad (2)$$

Proof overview. The proof of Theorem 3.3 relies on two lemmas: 1) The first shows that $f^\star$ is a feasible solution for Constraints (2) (Lemma 3.6). The feasibility of $f^\star$ for the first constraint of (2) is guaranteed by Assumption 1 and for the second constraint of (2) follows from the fact that $\Gamma_i(f)$ ($i \in \{0, 1\}$) is a good estimation of $\Pr[\xi(f), \xi'(f) \mid Z = i]$ by the Chernoff bound. 2) The second lemma shows that w.h.p. ($1 - F$ for small F), all unfair classifiers $f \in \mathcal{F}$ that are either not feasible for Program ConFair or violate Assumption 1, violate Constraint (2) (Lemma 3.8). Since the space of unfair classifiers is continuous, the main difficulty is to upper bound the (violating) probability F. Towards this, we first divide the collection of all unfair classifiers into multiple groups depending on how much they violate Constraint (2) (Definition 3.7). Then, for each group G_i, we construct an ε_i-net $\mathcal{G}_i$ (Definition B.3); ensuring no classifier $f \in \mathcal{G}_i$ violate Constraint (2) is sufficient to guarantee that no classifier in group G_i violate Constraint (2). Here, ε_i is chosen to depend on the degree of violation of G_i. Using Chernoff bounds, we show that the probability each unfair classifier on the net $\mathcal{G}_i$ is feasible to Constraint (2) is $\exp(-O(1-\eta_0-\eta_1)^2\lambda^2 n))$. Hence, as λ decreases, it is more likely that an unfair classifier is feasible for Constraint (2). To bound the total violating probability, it remains to bound the number of classifiers in the union of these nets (Definition B.5). The idea is to apply the relation between VC-dimension and ε-nets (Theorem B.4). Overall, constant 60000 in the success probability comes by setting the appropriate ε_1 to capture the capacity of unfair classifiers in Lemma 3.6.

The two lemmas imply that the empirical risk of f^Δ is guaranteed to be at most that of $f^\star$ and f^Δ must be fair over S (Theorem 3.3). Overall, the main technical contribution is to discretize the space of unfair classifiers by carefully chosen multiple ε-nets with different violation degrees to our denoised program, and upper bound the capacity of the union of all nets via a VC-dimension bound.

We now present the formal statements of the two main lemmas: Lemmas 3.6 and 3.8, and defer all proofs to Section B.

Lemma 3.6 (Relation between Program TargetFair and DFair) *Let $f \in \mathcal{F}$ be an arbitrary classifier and $\varepsilon \in (0, 0.5)$. With probability at least $1 - 2e^{-\varepsilon^2 n/6}$,*

$$(1-\eta_{1-i}) \Pr\left[f = 1, \widehat{Z} = i\right] - \eta_{1-i} \Pr\left[f = 1, \widehat{Z} = 1-i\right] \in (1-\eta_0-\eta_1) \Pr[f = 1, Z = i] \pm \varepsilon,$$

for $i \in \{0, 1\}$. Moreover, if $\min_{i\in\{0,1\}} \Pr[f = 1, Z = i] \geq \frac{\lambda}{2}$, then with probability at least $1 - 4e^{-\frac{\varepsilon^2(1-\eta_0-\eta_1)^2\lambda^2 n}{2400}}$,

$$\gamma^\Delta(f, \widehat{S}) \in (1 \pm \varepsilon)\gamma(f, S).$$

The first part of this lemma shows how to estimate $\Pr[f = 1, Z = i]$ ($i \in \{0, 1\}$) in terms of $\Pr\left[f = 1, \widehat{Z} = 0\right]$ and $\Pr\left[f = 1, \widehat{Z} = 1\right]$, which motivates the first constraint of (2). The second part of the lemma motivates the second constraint of (2). Then by Assumption 1, $f^\star$ is likely to be feasible for Program DFair. Consequently, f^Δ has empirical loss at most that of $f^\star$.

For our second main lemma, we first define the collection of classifiers that are expected to violate. Constraint (2).

Definition 3.7 (Bad classifiers) *Given a family $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$, we call $f \in \mathcal{F}$ a bad classifier if f belongs to at least one of the following sub-families:*

- $\mathcal{G}_0 := \{f \in \mathcal{F} : \min\{\Pr[f=1, Z=0], \Pr[f=1, Z=1]\} < \frac{\lambda}{2}$;
- *Let* $T = \lceil 232 \log\log \frac{2(\tau-3\delta)}{\lambda} \rceil$. *For* $i \in [T]$, *define*

$$\mathcal{G}_i := \left\{ f \in \mathcal{F} \setminus \mathcal{G}_0 : \gamma(f, S) \in \left[\frac{\tau-3\delta}{1.01^{2^{i+1}-1}}, \frac{\tau-3\delta}{1.01^{2^i-1}} \right) \right\}.$$

Intuitively, classifier $f \in \mathcal{G}_0$ is likely to violate the first of Constraint (2); and for $f \in \mathcal{G}_i$ for some $i \in [T]$ it is likely that $\gamma^{\Delta}(f, \widehat{S}) < \tau - \delta$. Thus, any bad classifier is likely to violate Constraint (2) (Lemma 3.6). Then we lower bound the total violating probability for all bad classifiers by the following lemma.

Lemma 3.8 (Bad classifiers are not feasible for Constraint (2)) *Suppose the VC-dimension of $(S, \mathcal{F})$ is t; then with probability at least*

$$1 - O\left(e^{-\frac{(1-\eta_0-\eta_1)^2\lambda^2\delta^2 n}{5000} + t\ln(\frac{50}{(1-\eta_0-\eta_1)\lambda\delta})} \right),$$

any bad classifier violates Constraint (2).

Theorem 3.3 for $p = 2$ and statistical rate is almost a direct corollary of Lemmas 3.6 and 3.8 (see Section B) except that we need to verify that any classifier violating Program (**DFair**) is a bad classifier in the sense of Definition 3.7.

4. Empirical Results

We implement our denoised program, for binary and non-binary protected attributes, and compare the performance with baseline algorithms on real-world datasets.

Datasets. We perform simulations on the **Adult** (Asuncion & Newman, 2007) and **COMPAS** (Angwin et al., 2016b) benchmark datasets, as pre-processed in AIF360 toolkit (Bellamy et al., 2018b). The **Adult** dataset consists of rows corresponding to 48,842 individuals, with 18 binary features and a label indicating whether the income is greater than 50k USD or not. We use binary protected attributes sex ("male" (Z=1) vs "female" (Z=0)) and race ("White" (Z=1) vs "non-White" (Z=0)) for this dataset. The **COMPAS** dataset consists of rows corresponding to 6172 individuals, with 10 binary features and a label that takes value 1 if the individual does not reoffend and 0 otherwise. We take sex (coded as binary) and race (coded as non-binary - "African-American" (Z=1), "Caucasian" (Z=2), "Other" (Z=3)) to be the protected attributes.

Metrics and baselines. We implement our program using logistic loss with denoised constraints with respect to the statistical rate and false positive rate metrics; we refer to our algorithm with statistical rate constraints as **DLR-SR** and with false positive rate constraints as **DLR-FPR**. [2] To obtain computationally feasible formulations of our optimization problem (2), we expand the constraint on the fairness metrics by forming constraints on relevant (empirical) rates of all groups, and solve the nonconvex program using SLSQP; the details of the constraints are presented in Section E. We compare against state-of-the-art noise-tolerant fair classification algorithms: **LZMV** (Lamy et al., 2019), **AKM** (Awasthi et al., 2020), and **WGN+** (Wang et al., 2020). **LZMV** takes as input a parameter, ε_L, to control the fairness of the final classifier; for statistical rate, this parameter represents the desired absolute difference between the likelihood of positive class label across the two protected groups and **LZMV** is, therefore, the primary baseline for comparison with respect to statistical rate. We present the results of (Lamy et al., 2019) for different ε_L values.[3] **AKM**[4] and **WGN+**[5] are the primary baseline for comparison with respect to false positive rate metric. As discussed earlier, the algorithm **AKM** is the post-processing algorithm of Hardt et al. (2016). [6] For **WGN+**, we use the algorithm that employs soft-group assignments (Kallus et al., 2020) to form false positive rate constraints; it is the only prior algorithm that can handle non-binary noisy protected attributes and, hence, it is also the main baseline for the COMPAS dataset with race protected attribute.

Note that we test all methods in the setting of flipping noises in protected attribute. While the **LZMV** and **AKM** methods also address the flipping noise setting, they require less information about the noise model than our algorithm. **LZMV** only needs access to the sum of $\eta_0+\eta_1$ to construct their modified constraints (in case of binary protected attribute), and **AKM** do not require access to the noise parameters (they directly use the algorithm of (Hardt et al., 2016)). Our approach **DLR**, on the other hand, uses the entire noise matrix to design appropriate denoised constraints.

Additionally, we implement the baseline which minimizes the logistic loss with fairness constraints ($\tau = 0.9$) over the noisy protected attribute as described in Section F.2. When the fairness metric is the statistical rate, we will refer to this program as **LR-SR**, and when the fairness metric is the false

[2] We use the (noisy) protected attribute to construct the constraints, but not for classification. However, if necessary, the protected attribute can also be used as a feature for classification.

[3] github.com/AIasd/noise_fairlearn.

[4] github.com/matthklein/equalized_odds_under_perturbation.

[5] github.com/wenshuoguo/robust-fairness-code.

[6] Equalized odds fairness metric aims for parity w.r.t false positive and true positive rates. For clarity of presentation, we present the empirical analysis with respect to false positive rate only.

Table 1. The performance on accuracy and fairness metrics of all algorithms over the test datasets; we report the average and standard error (in parenthesis). When the protected attribute is binary, the fairness metrics (SR, FPR) are $\min_{i\in\{0,1\}} q_i(f)/\max_{i\in\{0,1\}} q_i(f)$. For the non-binary protected attribute (COMPAS-race), we report the performance for all groups; i.e., SR_j, FPR_j denote $q_j(f)/\max_{i\in[p]} q_i(f)$, for all $j\in[p]$. By definition, SR$=\min\{\text{SR}_j\}$ and FPR$=\min\{\text{FPR}_j\}$. The full accuracy-fairness tradeoffs when varying τ can be found in Section E in the supplementary material. For each dataset and protected attribute, the metrics of the method that achieves the largest sum of mean accuracy and mean statistical rate (one way to measure fairness-accuracy tradeoff) has also been colored in green, and the method that achieves the largest sum of mean accuracy and mean false positive rate has been colored in yellow. Our method **DLR** achieves the best tradeoff or is within one standard deviation of the best tradeoff, as measured in this manner, in 6 out of 8 settings.

	Adult						**COMPAS**									
	sex (binary)			race (binary)			sex (binary)			race (non-binary)						
	acc	SR	FPR	acc	SR	FPR	acc	SR	FPR	acc	SR_0	SR_1	SR_2	FPR_0	FPR_1	FPR_2
Unconstrained	**.80** (0)	.31 (.01)	.45 (.03)	**.80** (0)	.68 (.02)	.81 (.09)	**.67** (.01)	.78 (.04)	.70 (.08)	**.67** (0)	.66 (.02)	.96 (.01)	1.0 (.0)	.57 (.02)	1.0 (0)	.94 (.01)
LR-SR	.76 (.01)	.68 (.24)	.68 (.21)	.76 (.01)	.69 (.27)	.71 (.26)	.67 (.01)	.79 (.04)	.72 (.08)	.58 (.06)	.86 (.09)	.98 (.03)	.98 (.02)	.85 (.11)	.98 (.04)	.96 (.04)
LR-FPR	.76 (.01)	.82 (.21)	.78 (.25)	.76 (0)	.83 (.29)	.84 (.29)	.67 (.02)	.80 (.04)	.72 (.08)	.56 (.05)	.87 (.08)	.97 (.06)	.97 (.03)	.86 (.09)	.96 (.07)	.95 (.09)
LZMV ε_L=.01	.35 (.01)	**.99** (0)	**.99** (0)	.37 (.05)	**.98** (0)	**.99** (0)	.55 (.01)	**.98** (.04)	**.98** (.09)	-	-	-	-	-	-	-
LZMV ε_L=.04	.67 (.04)	.85 (.06)	**.99** (.01)	.77 (.03)	.79 (.10)	.85 (.09)	.58 (.01)	.94 (.02)	.94 (.03)	-	-	-	-	-	-	-
LZMV ε_L=.10	.78 (.02)	.69 (.09)	.79 (.11)	.80 (0)	.70 (.01)	.82 (.08)	.64 (.02)	.85 (.05)	.81 (.07)	-	-	-	-	-	-	-
AKM	.77 (0)	.66 (.05)	.89 (.04)	**.80** (0)	.72 (.02)	.90 (.08)	.66 (.01)	.83 (.04)	.77 (.09)	-	-	-	-	-	-	-
WGN+	.70 (.05)	.73 (.12)	.76 (.05)	.76 (.01)	.84 (.05)	.92 (.05)	.59 (.01)	.90 (.02)	.84 (.01)	.56 (.02)	.89 (.14)	.91 (.18)	.96 (.13)	.85 (.16)	.87 (.23)	.94 (.16)
DLR-SR τ=.7	.77 (.01)	.74 (.14)	.87 (.17)	.79 (.01)	.80 (.12)	.90 (.10)	.67 (.01)	.79 (.04)	.72 (.08)	.66 (.01)	.73 (.04)	.99 (.01)	1.0 (0)	.66 (.05)	1.0 (.0)	.92 (.03)
DLR-SR τ=.9	.76 (.01)	.85 (.15)	.80 (.12)	.76 (.01)	.88 (.18)	.90 (.19)	.63 (.04)	.86 (.05)	.83 (.08)	.55 (.04)	.91 (.06)	.97 (.04)	.97 (.03)	.89 (.09)	.97 (.04)	.93 (.1)
DLR-FPR τ=.7	.77 (.02)	.73 (.14)	.85 (.17)	.78 (.02)	.77 (.11)	.88 (.11)	.66 (.01)	.80 (.04)	.73 (.08)	.64 (.02)	.76 (.05)	.99 (.01)	.98 (.02)	.72 (.06)	1.0 (.0)	.89 (.06)
DLR-FPR τ=.9	.77 (.02)	.77 (.12)	.91 (.11)	.77 (.02)	.80 (.15)	.88 (.14)	.60 (.06)	.86 (.07)	.82 (.10)	.53 (.04)	**.92** (.06)	.97 (.06)	.95 (.06)	**.93** (.08)	.94 (.09)	.93 (.07)

positive rate, we will refer to it as **LR-FPR**. Finally, we also learn an unconstrained optimal classifier as a baseline.

Implementation details. We first shuffle and partition the dataset into a train and test partition (70-30 split). Given the training dataset S, we generate a noisy dataset $\widehat{S}$. For binary protected attributes, we use $\eta_0 = 0.3$ and $\eta_1 = 0.1$. For non-binary protected attributes, we use the noise matrix $H = \begin{bmatrix} 0.70 & 0.15 & 0.15 \\ 0.05 & 0.90 & 0.05 \\ 0.05 & 0.05 & 0.90 \end{bmatrix}$ (i.e., the minority group is more likely to contain errors, as would be expected in various applications (Nobles, 2000)). Our algorithms, as well as the baselines, have access to the known η and H values. We consider other choices of noise parameters, impact of error in estimates of noise parameter, and performance when protected attribute is partially predicted using non-protected features in Section E in the Supplementary Material. We train each algorithm on $\widehat{S}$ and vary the fairness constraints (e.g., the choice of $\tau \in [0.5, 0.95]$ in **DLR**) to learn the corresponding fair classifier and record its accuracy (acc) over the test dataset. We also record the fairness metric (statistical rate or false positive rate) γ of the classifier with respect to the true (non-noisy) version of the protected attributes in the test dataset. We perform 50 repetitions and report the mean and standard error of fairness and accuracy metrics across the repetitions. For COMPAS, we use λ=0.1 as a large fraction (47%) of training samples have class label 1, while for Adult, we use λ=0 as the fraction of positive class labels is small (24%). [7]

[7] Alternately, one could use a range of values for λ to construct multiple classifiers, and choose the one which satisfies the program constraints and has the best accuracy over a separate validation partition. We find that these λ are sufficient to obtain fair classifiers for the considered datasets.

Results. Table 1 summarizes the fairness and accuracy achieved by our methods and baseline algorithms over the Adult and COMPAS test datasets. The first observation is that our approach, **DLR-SR** and **DLR-FPR**, achieve higher fairness than the unconstrained classifier, showing its effectiveness in noise-tolerant fair classification. The extent of this improvement varies with the strength of the constraint τ, but comes with a natural tradeoff with accuracy.

For each dataset and protected attribute, we also highlight the method that achieves the largest sum of mean accuracy and mean fairness in Table 1. Note that this is just one way to measure the fairness-accuracy tradeoff and this measure can highlight approaches that achieve high fairness but low accuracy (or vice versa). The full results of all metrics in Table 1 should be taken into account to develop a more nuanced understanding of the performance of different algorithms and their fairness-accuracy tradeoffs. Nevertheless, when fairness-accuracy tradeoff is measured in this manner, our method **DLR** achieves the best tradeoff or is within one standard deviation of the best tradeoff in 6 out of 8 settings.

For Adult dataset, **DLR-SR** and **DLR-FPR** (with τ=0.9) can attain a higher fairness metric value than **LR-SR** and **LR-FPR** respectively, and perform similarly with respect to accuracy. The statistical rate-accuracy tradeoff of **DLR-SR**, for this dataset, is also better than **LZMV**, **AKM**, and **WGN+**; in particular, high statistical rate for Adult dataset using **LZMV** (i.e., $\geq$0.8) is achieved only with a relatively larger loss in accuracy (for example, with ε_L=0.01), whereas for **DLR-SR**, the loss in accuracy when using τ=0.9 is relatively small ($\sim$0.03) while the statistical rate

is still high ($\sim$0.85). With respect to false positive rate, **AKM** can achieve a high false positive rate for the Adult dataset ($\sim$0.90), while **WGN+** does not achieve high false positive rate when sex is the protected attribute. In comparison, **DLR-FPR** with τ=0.9 can also achieve a high false positive rate at a small loss of accuracy for both protected attributes, and the best false positive rate and accuracy of **DLR-FPR** and **AKM** are within a standard deviation of each other. Baseline **LZMV** attains a high false positive rate too for the Adult dataset, but the loss in accuracy is larger compared to **DLR-FPR**.

For the COMPAS dataset, with sex as protected attribute, **LZMV** (ε_L=0.01, 0.04) achieves high statistical rate and false positive rate, but at a large cost to accuracy. Meanwhile **DLR-SR** (τ=0.9) returns a classifier with SR $\sim$ 0.86 and FPR $\sim$ 0.83 and significantly better accuracy (0.63) than **LZMV**(ε_L=0.01, 0.04). Further, our algorithm can achieve higher fairness as well, at the cost of accuracy, using a larger input τ (e.g., τ=1; see Section E in the Supplementary Material). Note that in this case, the unconstrained classifier already has high fairness values. Hence, despite the noise in protected attribute, the task of fair classification is relatively easy and all baselines, as well as, our methods perform well for this dataset and protected attribute.

For the COMPAS dataset with non-binary race protected attribute, we also present the complete breakdown of relative performance for each protected attribute value in Table 1. Both **DLR-SR** and **DLR-FPR** (with $\tau = 0.9$) reduce the disparity between group-performances $q_j(f)$ and $\max_{i\in[p]}$ $\forall j \in [p]$, for SR and FPR metrics, to a larger extent compared to the unconstrained classifier, baselines and **WGN+**.

The tradeoff between the fairness metric and accuracy for all methods is also graphically presented in Section E in the Supplementary Material. Evaluation with respect to both metrics shows that our framework can handle binary and non-binary protected attributes, and attain close to the user-desired fairness metric values (as defined using τ). For $\tau = 0.9$, **DLR** can achieve high fairness (> 0.8), albeit at a cost to accuracy, in all considered settings. Comparison with baselines further shows that, unlike **AKM** and **WGN+**, our approach always returns classifiers with high fairness metrics values, and unlike **LZMV**, the loss in accuracy to achieve high fairness values is relatively small.

We also present the performance of our approach using false discovery rate (linear-fractional metric) constraints in Section E; in that setting, our approach has better fairness-accuracy tradeoff than baselines for **Adult** and similar tradeoff as the best-performing baseline for **COMPAS**.

5. Conclusion, Limitations & Future Work

In this paper, we study fair classification with noisy protected attributes. We consider flipping noises and propose a unified framework that constructs an approximate optimal fair classifier over the underlying dataset for multiple, non-binary protected attributes and multiple linear-fractional fairness constraints. Our framework outputs a classifier that is guaranteed to be both fair and accurate. Empirically, our denoised algorithm can achieve the high fairness values at a small cost to accuracy. Thus this work broadens the class of settings where fair classification techniques can be applied by working even when the information about protected attributes is noisy. [8]

Our framework can be applied to a wide class of fairness metrics, and hence may be suitable in many domains. However, it is not apriori clear which fairness metrics should be used in any given setting, and the answers will be very context-dependent; the effectiveness of our framework towards mitigating bias will depend crucially on whether the appropriate choice of features and parameters are selected. An ideal implementation of our framework would involve an active dialogue between the users and designers, a careful assessment of impact both pre and post-deployment. This would in particular benefit from regular public audits of fairness constraints, as well as ways to obtain and incorporate community feedback from stakeholders (Sassaman et al., 2020; Chancellor et al., 2019).

Our work leaves several interesting future directions. One is to theoretically consider other noise models for non-binary attributes that are not independent, e.g., settings where the noise follows a general mutually contaminated model (Scott et al., 2013) or when the noise on the protected type also depends on other features, such as, when imputing the protected attributes. Our framework can still be employed in these settings (e.g., given group prediction error rates); however, methods that take into account the protected attribute prediction model could potentially further improve the performance. There exist several works that also design fair classifiers with noisy labels (Blum & Stangl, 2020; Biswas & Mukherjee, 2021) and another direction is to consider joint noises over both protected attributes and labels. Our model is also related to the setting in which each protected attribute follows a known distribution; whether our methods can be adapted to this setting can be investigated as part of future work.

Acknowledgements

This research was supported in part by a J.P. Morgan Faculty Award and an AWS MLRA grant.

[8] Code available at github.com/vijaykeswani/Noisy-Fair-Classification.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. M. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pp. 60–69, 2018.

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

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. https://github.com/propublica/compas-analysis, 2016b.

Asuncion, A. and Newman, D. UCI machine learning repository. archive.ics.uci.edu/ml/index.php, 2007. University of California, Irvine, School of Information and Computer Sciences.

Awasthi, P., Kleindessner, M., and Morgenstern, J. Equalized odds postprocessing under imperfect group information. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, 2020.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018a.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018b. URL https://arxiv.org/abs/1810.01943.

Biddle, D. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

Biswas, A. and Mukherjee, S. Ensuring fairness under prior probability shifts. *AIES*, 2021.

Blum, A. and Stangl, K. Recovering from biased data: Can fairness constraints improve accuracy? *Symposium on the foundations of responsible computing, (FORC) 2020*, 2020.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, pp. 77–91, 2018.

Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

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

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

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 3992–4001, 2017.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 319–328, 2019.

Celis, L. E., Keswani, V., and Vishnoi, N. K. Data preprocessing to mitigate bias: A maximum entropy based approach. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1349–1359. PMLR, 2020. URL http://proceedings.mlr.press/v119/celis20a.html.

Chancellor, S., Guha, S., Kaye, J., King, J., Salehi, N., Schoenebeck, S., and Stowell, E. The relationships between data, power, and justice in cscw research. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pp. 102–105, 2019.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 339–348, 2019.

Comptroller. *Fair lending: Comptroller's Handbook*. 2010. URL https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/fair-lending/pub-ch-fair-lending.pdf.

Data, E., Ploeg, M. V., and Perrin, E. Eliminating health disparities: Measurement and data needs. *Washington (DC): National Academies Press (US)*, 2004.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pp. 429–438, 2013.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. D. M. Decoupled classifiers for group-fair and efficient machine learning. In *Fairness, Accountability, and Transparency in Machine Learning*, pp. 119–133, 2018.

Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., and Lurie, N. Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2):69–83, 2009.

Federal Reserve Bank, B. *Closing the Gap: A Guide to Equal Opportunity Lending*. Federal Reserve Bank of Boston, 1993. URL https://books.google.com/books?id=UkB5kgAACAAJ.

Fish, B., Kun, J., and Lelkes, Á. D. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining, 2016*, pp. 144–152. SIAM, 2016.

Freedman, D. A. *Statistical models: theory and practice*. cambridge university press, 2009.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In *Fairness, Accountability, and Transparency in Machine Learning*, 2019.

Goel, N., Yaghini, M., and Faltings, B. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Goh, G., Cotter, A., Gupta, M. R., and Friedlander, M. P. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, pp. 2415–2423, 2016.

Gordaliza, P., del Barrio, E., Gamboa, F., and Loubes, J. Obtaining fairness using optimal transport theory. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 2357–2365, 2019.

Gupta, M. R., Cotter, A., Fard, M. M., and Wang, S. Proxy fairness. *CoRR*, abs/1806.11212, 2018.

Har-peled, S. *Geometric Approximation Algorithms*. American Mathematical Society, USA, 2011. ISBN 0821849115.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing*, pp. 3315–3323, 2016.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *ICML 2018: Thirty-fifth International Conference on Machine Learning*, pp. 1929–1938, 2018.

Haussler, D. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69 (2):217–232, 1995.

Hoeffding, W. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pp. 409–426. Springer, 1994.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Huang, L. and Vishnoi, N. K. Stable and fair classification. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 2879–2890, 2019.

Jiang, H. and Nachum, O. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

Kallus, N., Mao, X., and Zhou, A. Assessing algorithmic fairness with unobserved protected class using data combination. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pp. 110, 2020.

Kamiran, F. and Calders, T. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pp. 1–6. IEEE, 2009.

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

Kraft, D. *A Software Package for Sequential Quadratic Programming*. Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt Köln: Forschungsbericht. Wiss. Berichtswesen d. DFVLR, 1988. URL https://books.google.com.hk/books?id=4rKaGwAACAAJ.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. H. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

Lamy, A., Zhong, Z., Menon, A. K., and Verma, N. Noise-tolerant fair classification. In *Advances in Neural Information Processing Systems*, pp. 294–306, 2019.

Lint, J. H. V. *Introduction to Coding Theory*. Springer-Verlag, Berlin, Heidelberg, 3rd edition, 1998. ISBN 3540641335.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Menon, A., Van Rooyen, B., Ong, C. S., and Williamson, B. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pp. 125–134, 2015.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107–118, 2018a.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency, FAT*, pp. 107–118, 2018b.

Muthukumar, V., Pedapati, T., Ratha, N. K., Sattigeri, P., Wu, C., Kingsbury, B., Kumar, A., Thomas, S., Mojsilovic, A., and Varshney, K. R. Understanding unequal gender classification accuracy from face images. *CoRR*, abs/1812.00099, 2018.

Nobles, M. Shades of citizenship: Race and the census in modern politics. *Bibliovault OAI Repository, the University of Chicago Press*, 2000.

Northcutt, C. G., Wu, T., and Chuang, I. L. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'17. AUAI Press, 2017.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. M., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pp. 5684–5693, 2017.

Roh, Y., Lee, K., Whang, S., and Suh, C. Fr-train: A mutual information-based approach to fair and robust training. In *International Conference on Machine Learning*, pp. 8147–8157. PMLR, 2020.

Saez, J. A., Galar, M., Luengo, J., and Herrera, F. Tackling the problem of classification with noisy data using multiple classifier systems: Analysis of the performance and robustness. *Information Sciences*, 247:1–20, 2013.

Sassaman, H., Lee, J., Irvine, J., and Narayan, S. Creating community-based tech policy: case studies, lessons learned, and what technologists and communities can do together. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 685–685, 2020.

Scott, C., Blanchard, G., and Handy, G. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on Learning Theory*, pp. 489–511, 2013.

Srivastava, M., Heidari, H., and Krause, A. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2459–2468, 2019.

Wang, H., Ustun, B., and Calmon, F. P. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 6618–6627, 2019.

Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M. R., and Jordan, M. I. Robust optimization for fairness with noisy protected groups. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Woodworth, B. E., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pp. 1920–1953, 2017.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017*, pp. 1171–1180, 2017a.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, pp. 962–970, 2017b.

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