Fair Classification with Adversarial Perturbations

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study fair classification in the presence of an omniscient adversary that, given an η, is allowed to choose an arbitrary η-fraction of the training samples and arbitrarily perturb their protected attributes. The motivation comes from settings in which protected attributes can be incorrect due to strategic misreporting, malicious actors, or errors in imputation; and prior approaches that make stochastic or independence assumptions on errors may not satisfy their guarantees in this adversarial setting. Our main contribution is an optimization framework to learn fair classifiers in this adversarial setting that comes with provable guarantees on accuracy and fairness. Our framework works with multiple and non-binary protected attributes, is designed for the large class of linear-fractional fairness metrics, and can also handle perturbations besides protected attributes. We prove near-tightness of our framework's guarantees for natural hypothesis classes: no algorithm can have significantly better accuracy and any algorithm with better fairness must have lower accuracy. Empirically, we evaluate the classifiers produced by our framework for statistical rate on real-world and synthetic datasets for a family of adversaries.

1 Introduction

It is increasingly common to deploy classifiers to assist in decision-making in applications such as criminal recidivism [40], credit lending [21], and predictive policing [30]. Hence, it is imperative to ensure that these classifiers are fair with respect to protected attributes such as gender and race. Consequently, there has been extensive work on approaches for fair classification [29, 24, 26, 18, 49, 48, 38, 23, 25, 1, 14]. At a high level, a classifier f is said to be "fair" with respect to a protected attribute Z if it has a similar "performance" with respect to a given metric on different groups defined by Z. Given a fairness metric and a hypothesis class $\mathcal{F}$, fair classification frameworks consider the problem of finding a classifier $f^\star \in \mathcal{F}$ that has the optimal accuracy while subject to being fair with respect to the given fairness metric (and Z) [9]. To specify the fairness constraints, these approaches need the protected attributes of the training data to be known.

However, the protected attributes can be erroneous for various reasons; there could be uncertainties during the data collection or data cleaning process [20, 41], or the attributes could be strategically misreported [37]. Further, protected attributes may be missing entirely, as is often the case for racial and ethnic information in healthcare [20] or when data is scraped from the internet as with many image datasets [22, 50, 31]. In these cases, protected attributes can be "imputed" [19, 32, 17], but this can also introduce errors [13]; imputation is known to be fragile to imperceptible changes in the inputs [27] and to have correlated errors across samples [39]. Perturbations in protected attributes, regardless of origin, have been shown to have adverse effects on fair classifiers, affecting their performance on both accuracy and fairness metrics; see e.g., [17, 8].

Towards addressing this problem, several recent works have developed fair classification algorithms for various models of errors in the protected attributes. [35] consider an extension of the "mutually contaminated learning model" [42] where, instead of observing samples from the "true" joint

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

distribution, the distribution of observed group-conditional distributions are stochastic mixtures of their true counterparts. [7] consider a binary protected attribute and Bernoulli perturbations that are independent of the labels (and of each other). [15] consider the setting where each sample's protected attribute is independently flipped to a different value with a known probability. [47] considers two approaches to deal with perturbations. In their "soft-weights" approach, they assume perturbations follow a fixed distribution and one has access to an auxiliary dataset containing independent draws of both the true and perturbed protected attributes. In their distributionally robust (DR) approach, for each protected group, its feature and label distributions in the true data and the perturbed data are a known total variation distance away from each other. Finally, in an independent work, [34] study fair classification under the Malicious noise model [46, 33] in which a fraction of the training samples are chosen uniformly at random, and can then be perturbed arbitrarily.

Our perturbation model. We extend this line of work by studying fair classification under the following worst-case adversarial perturbation model: Given an $\eta > 0$, after the training samples are independently drawn from an true distribution $\mathcal{D}$, the adversary with unbounded computation power sees all the samples and can use this information to choose any η-fraction of the samples and perturb their protected attributes arbitrarily. This model is a straightforward adaptation of the perturbation model of [28] to the fair classification setting and we refer to it as the η-Hamming model. Unlike the perturbation models studied before, this model can capture settings where the perturbations are strategic or arbitrarily correlated as can arise in the data collection stage or during imputation of the protected attributes, and in which the errors cannot be "estimated" using auxiliary datasets. In fact, under this perturbation model, the classifiers outputted by prior works can violate the fairness constraints by a large amount or have an accuracy that is significantly lower than the accuracy of $f^\star$; see Section 5 and Supplementary Material H. Taking these perturbed samples as input, the goal is to learn a classifier that satisfies a given set of fairness constraints with minimal loss to accuracy, where accuracy and fairness are measured with respect to true distribution $\mathcal{D}$.

Our contributions. Our main contribution is an optimization framework (Definition 4.1) to learn fair classifiers for the η-Hamming model, which comes with provable guarantees on accuracy and fairness (Theorem 4.3). Our framework works for multiple and non-binary protected attributes, and the large class of linear-fractional fairness metrics (that capture most fairness metrics studied in the literature); see Definition 3.1 and [14]. The framework provably outputs a classifier whose error is at most 2η larger than the error of $f^\star$ and that additively violates the fairness constraint by at most $O(\eta/\lambda)$ (Theorem 4.3), under the mild assumption that the "performance" of $f^\star$ on each protected group is larger than a known constant $\lambda > 0$ (Assumption 1).

Assumption 1 is drawn from the work of [15] for fair classification with stochastic perturbations. While it is not clear if the assumption is necessary in their model, we show that Assumption 1 is necessary for fair classification in the η-Hamming model: If λ is not bounded away from 0, then no algorithm can achieve a non-trivial guarantee on *both* accuracy and fairness (Theorem 4.4). Moreover, we prove the near-tightness of our framework's guarantee under Assumption 1: No algorithm can guarantee an accuracy closer than η to that of $f^\star$ and any algorithm that additively violates the fairness constraint by less than $\eta/(20\lambda)$ must have at least $1/20$ error; Theorems 4.5 and D.1. Finally, we also extend our framework's guarantees to the Nasty Sample Noise model (Remark 2 in Section 4). The Nasty Sample Noise model is a generalization of the η-Hamming model, that was studied by [12] in the context of PAC learning (without any fairness considerations), can choose any η-fraction of the samples, and can arbitrarily perturb both their labels and features.

We implement our approach using the logistic loss function with linear classifiers and evaluate its performance on COMPAS [4] and a synthetic dataset (Section 5). We generate perturbations of these datasets admissible in the η-Hamming model and compare the performance of our approach to key baselines [35, 7, 47, 15, 34] with statistical rate (SR) and false-positive rate (FPR) as fairness metrics. On the synthetic dataset we compare against a method developed for fair classification under stochastic perturbations [15] and demonstrate the comparative strength of the η-Hamming model; the results show that [15]'s framework achieves a significantly lower accuracy than our framework for the same SR. Empirical results on COMPAS show that our framework can attain better fairness than the unconstrained classifier, with a minimal loss in accuracy. Further, our framework has a similar (or better) fairness-accuracy trade-off compared to all baselines we consider (Figure 1 and Figure 7) in a variety of settings, and is not dominated by any other approach.

Techniques. The starting point to our optimization framework (Definition 4.1) is Program (1) for fair classification in the absence of perturbations. The accuracy guarantee of our framework comes by ensuring that $f^\star \in \mathcal{F}$, which is an optimal solution for Program (1), is feasible for our framework. However, without modifications, classifiers with higher accuracy than that of $f^\star$ and much lower fairness (with respect to the true distribution $\mathcal{D}$) can also be feasible for our framework. This is because our framework can only impose fairness constraints with respect to the perturbed samples $\widehat{S}$; and the ratio of a classifier's fairness with respect to $\widehat{S}$ and with respect to $\mathcal{D}$ can be arbitrarily small (see Example G.2). To address this, we introduce the notion of s-stability (Definition 4.7). Roughly, $f \in \mathcal{F}$ is said to be s-stable with respect to a fairness metric Ω if, for all η-Hamming perturbations, the ratio of fairness of f (as measured Ω) on the true distribution $\mathcal{D}$ and the perturbed samples $\widehat{S}$ is between s and $1/s$. From this definition, it follows that any s-stable classifier that has fairness τ' with respect to $\widehat{S}$ (which is ensured by Condition (3)), has fairness at least $\tau' \cdot s$ with respect to $\mathcal{D}$. Thus, if we could ensure that all feasible solutions of our framework are s-stable (for a suitable s) and that $f^\star$ is feasible for our framework, then the classifier output by our framework would satisfy the required guarantees (Lemma 4.9). However, it is not possible to enforce s-stability in the absence of true samples S. Instead, we give a condition (4) (which $f^\star$ satisfies under Assumption 1) that can be verified with access to the perturbed samples $\widehat{S}$, and show that any classifier that satisfies (4) with respect to $\widehat{S}$ is s-stable (Lemma 4.8).

2 Related work

In this section, we present the key related works. We defer other related work (e.g., fair classification in the absence of protected attributes) to Supplementary Material A due to space constraints.

As discussed in Section 1, the perturbation model considered in this paper is stronger than those considered in [35, 7, 47, 15, 34]. There are several other distinctions between this paper and prior work. [35, 7] consider binary protected attributes, while our approach (and that of [47, 15]) can handle multiple categorical protected attributes. [7] consider equalized-odds (EO) fairness constraints and [35] consider SR and EO fairness constraints. In contrast, our approach (and that of [15]) works with multiple linear-fractional metrics (which include SR and can ensure EO fairness constraints). [7] identify conditions on the distribution of perturbations under which the equalized-odds post-processing algorithm of [29] improves the fairness of the unconstrained optimal. [47], in their DR approach, give provable guarantees on the fairness of the output classifiers[1] and in their "soft-weights" approach, give provable guarantees on the accuracy (with respect to $f^\star$) and fairness of the output classifier *in expectation*. In contrast, our work and those of [35, 15] give provable guarantees on the accuracy (with respect to $f^\star$) and fairness of output classifiers *with high probability*. The Malicious noise model studied by [34], which can modify a uniformly randomly selected subset of samples arbitrarily, is weaker than the Nasty Sample Noise model [12, 5], and hence, than the model considered in this paper. [34] give an algorithm for a binary protected attribute which, under the realizable assumption (i.e., assuming there exists a classifier with perfect accuracy), outputs a classifier with guarantees on accuracy and fairness with respect to true-positive rate. In contrast, we give a framework for the stronger Nasty Sample Noise model that works without the realizable assumption (i.e., in the agnostic setting), can handle multiple and non-binary protected attributes, and can ensure fairness for any linear-fractional metrics (which includes true-positive rate).

Another line of work has studied PAC learning in the presence of adversarial (and stochastic) perturbations in the data, without considerations of fairness [33, 2, 12, 16, 6]; see also [5]. In particular, [12] study PAC learning (without fairness constraints) under the Nasty Sample Noise model. They use the empirical risk minimization framework (ERM) (see, e.g., [43]) run on the perturbed samples to output a classifier. In contrast, our framework Program (ErrTolerant) finds empirical risk minimizing classifiers that satisfy fairness constraints on the perturbed data, and that are also "stable" for the given fairness metric. While both frameworks show that the accuracy of the respective output classifiers is within 2η of the respective optimal classifiers when the data is unperturbed, the optimal classifiers can be quite different. For instance, while [12]'s framework is guaranteed to output a classifier with high accuracy, it can perform poorly on fairness metrics; see Section 5 and Example H.1.

[1] Remark H.5 in Supplementary Material H gives an example where [47]'s DR approach has an accuracy arbitrarily close to $1/2$.

3 Model

Let the data domain be $D \coloneqq \mathcal{X} \times \{0,1\} \times [p]$, where $\mathcal{X}$ is the set of non-protected features, $\{0,1\}$ is the set of binary labels, and $[p]$ is the set of p protected attributes. Let $\mathcal{D}$ be a distribution over D. Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}\times[p]}$ be a hypothesis class of binary classifiers. For $f \in \mathcal{F}$, let $\mathrm{Err}_{\mathcal{D}}(f) \coloneqq \Pr_{(X,Y,Z)\sim\mathcal{D}}[f(X,Z) \neq Y]$ denote f's error on samples drawn from $\mathcal{D}$. In the vanilla classification problem, the goal of the learner $\mathcal{L}$ is to find a classifier with minimum error: $\operatorname{argmin}_{f\in\mathcal{F}} \mathrm{Err}_{\mathcal{D}}(f)$. In the fair classification problem, the learner is restricted to pick classifiers that have a "similar" performance conditioned on $Z = \ell$ for all $\ell \in [p]$. We consider the following class of metrics.

Definition 3.1 (Linear/linear-fractional metrics [14]). *Given $f \in \mathcal{F}$ and two events $\mathcal{E}(f)$ and $\mathcal{E}'(f)$, that can depend on f, define the performance of f on $Z = \ell$ ($\ell \in [p]$) as $q_\ell(f) \coloneqq \Pr_{\mathcal{D}}[\mathcal{E}(f) \mid \mathcal{E}'(f), Z = \ell]$. If $\mathcal{E}'$ depends on f, then $q_\ell(f)$ is said to be linear-fractional, otherwise linear.*

Definition 3.1 captures most of the performance metrics considered in the literature. For instance, for $\mathcal{E} \coloneqq (f = 1)$ and $\mathcal{E}' \coloneqq \emptyset$, we get statistical rate (a linear metric).[2] For $\mathcal{E} \coloneqq (Y = 0)$ and $\mathcal{E}' \coloneqq (f = 1)$, we get false-discovery rate. Given a performance metric q, the corresponding fairness metric is defined as

$$\Omega_{\mathcal{D}}(f) \coloneqq \min_{\ell\in[p]} q_\ell(f) / \max_{\ell\in[p]} q_\ell(f).$$

When $\mathcal{D}$ is the empirical distribution over samples S, then we use $\Omega(f, S)$ to denote $\Omega_{\mathcal{D}}(f)$. The goal of the fair classification problem, given a fairness metric Ω and a threshold $\tau \in (0,1]$, is to (approximately) solve the following:

$$\min_{f\in\mathcal{F}} \mathrm{Err}_{\mathcal{D}}(f) \quad \text{s.t.,} \quad \Omega_{\mathcal{D}}(f) \geq \tau. \tag{1}$$

If samples from $\mathcal{D}$ are available, then one could try to solve this program. However, as discussed in Section 1, we do not have access to the *true* protected attribute Z, but instead only see a perturbed version, $\widehat{Z} \in [p]$, generated by the following adversary.

η-Hamming model. Given an $\eta \in [0,1]$, let $\mathcal{A}(\eta)$ denote the set of all adversaries in the η-Hamming model. Any adversary $A \in \mathcal{A}(\eta)$ is a randomized algorithm with *unbounded* computation resources that knows the true distribution $\mathcal{D}$ and the algorithm of the learner $\mathcal{L}$. In this model, the learner $\mathcal{L}$ queries A for $N \in \mathbb{N}$ samples from $\mathcal{D}$ *exactly once*. On receiving the request, A draws N independent samples $S \coloneqq \{(x_i, y_i, z_i)\}_{i\in[N]}$ from $\mathcal{D}$, then A uses its knowledge of $\mathcal{D}$ and $\mathcal{L}$ to choose an arbitrary $\eta \cdot N$ samples ($\eta \in [0,1]$) and perturb their protected attribute arbitrarily to generate $\widehat{S} \coloneqq \{(x_i, y_i, \widehat{z}_i)\}_{i\in[N]}$. Finally, A gives these perturbed samples $\widehat{S}$ to $\mathcal{L}$.

Learning model. Given $\widehat{S}$ and the η, the learner $\mathcal{L}$ would like to (approximately) solve Program (1).

Definition 3.2 ((ε,ν)-learning). *Given bounds on error $\varepsilon \in (0,1)$ and constraint violation $\nu \in (0,1)$, a learner $\mathcal{L}$ is said to (ε,ν)-learn a hypothesis class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}\times[p]}$ with perturbation rate $\eta \in [0,1]$ and confidence parameter $\delta \in (0,1)$ if for all*

- *distributions $\mathcal{D}$ over $\mathcal{X} \times \{0,1\} \times [p]$ and*
- *adversaries $A \in \mathcal{A}(\eta)$,*

there exists a threshold $N_0(\varepsilon,\nu,\delta,\eta) \in \mathbb{N}$, such that with probability at least $1-\delta$ over the draw of $N \geq N_0(\varepsilon,\nu,\delta,\eta)$ iid samples $S \sim \mathcal{D}$, given η and the perturbed samples $\widehat{S} \coloneqq A(S)$, $\mathcal{L}$ outputs $f \in \mathcal{F}$ that satisfies $\mathrm{Err}_{\mathcal{D}}(f) - \mathrm{Err}_{\mathcal{D}}(f^\star) \leq \varepsilon$ and $\Omega_{\mathcal{D}}(f) \geq \tau - \nu$, where $f^\star$ is the optimal solution of Program (1) ($f^\star \coloneqq \operatorname{argmin}_{f\in\mathcal{F}} \mathrm{Err}_{\mathcal{D}}(f)$, s.t., $\Omega_{\mathcal{D}}(f) \geq \tau$).

Given finite number of perturbed samples, Definition 3.2 requires the learner to output a classifier that violates the fairness constraints additively by at most ν and that has an error at most ε smaller than that of $f^\star$, with probability at least $1-\delta$. Like PAC learning [46], for a given hypothesis class $\mathcal{F}$, Definition 3.2 requires the learner to succeed on all distributions $\mathcal{D}$.

Problem 1 (Fair classification with adversarial perturbations in protected attributes). *Given a hypothesis class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}\times[p]}$, a fairness metric Ω, a threshold $\tau \in [0,1]$, a perturbation rate $\eta \in [0,1]$, and perturbed samples $\widehat{S}$, the goal is to (ε,ν)-learn $\mathcal{F}$ for small $\varepsilon, \nu \in (0,1)$.*

[2] We overload the notation f to denote both the classifier as well as its prediction.

4 Theoretical results

In this section, we present our results on learning fair classifiers under the η-Hamming model. Our optimization framework (Program (ErrTolerant)) is a careful modification of Program (1). The main difficulty is that, unlike Program (1), it only has access to the perturbed samples $\widehat{S}$, and the ratio of a classifier's fairness with respect to the true distribution $\mathcal{D}$ and with respect to $\widehat{S}$ can be arbitrarily small (see Example G.2). To overcome this, our framework ensures that all feasible classifiers are "stable" (Definition 4.7). Then, as mentioned in Section 1, imposing the fairness constraint on $\widehat{S}$ guarantees (approximate) fairness on the true distribution $\mathcal{D}$. The accuracy guarantee follows by ensuring that the optimal solution of Program (1), $f^\star \in \mathcal{F}$, is feasible for our framework. To ensure this, we require Assumption 1 that also appeared in [15].

Assumption 1. *There is a known constant $\lambda > 0$ such that $\min_{\ell \in [p]} \Pr_{\mathcal{D}}[\mathcal{E}(f^\star), \mathcal{E}'(f^\star), Z = \ell] \geq \lambda$.*

It can be shown that this assumption implies that λ is also a lower bound on the performances $q_1(f^\star), \ldots, q_p(f^\star)$ that depend on $\mathcal{E}$ and $\mathcal{E}'$. We expect λ to be a non-vanishing positive constant in applications. For example, if q is SR, the minority protected group makes at least 20% of the population (i.e., $\min_{\ell \in [p]} \Pr_{\mathcal{D}}[Z = \ell] \geq 0.2$), and for all $\ell \in [p]$, $\Pr[f^\star = 1 \mid Z = \ell] \geq 1/2$, then $\lambda \geq 0.1$. In practice, λ is not known exactly but it can be set based on the context (e.g., see Section 5 and [15]). We show that Assumption 1 is necessary for the η-Hamming model (see Theorem 4.4).

Definition 4.1 (Error-tolerant program). *Given a fairness metric Ω and corresponding events $\mathcal{E}$ and $\mathcal{E}'$ (as in Definition 3.1), a perturbation rate $\eta \in [0, 1]$, and constants $\lambda, \Delta \in (0, 1]$, we define the error-tolerant program for perturbed samples $\widehat{S}$, whose empirical distribution is $\widehat{D}$, as*

$$\min_{f \in \mathcal{F}} \quad \mathrm{Err}_{\widehat{D}}(f), \tag{ErrTolerant; 2}$$

$$\text{s.t.,} \quad \Omega(f, \widehat{S}) \geq \tau \cdot \left(\frac{1-(\eta+\Delta)/\lambda}{1+(\eta+\Delta)/\lambda}\right)^2 \text{ and} \tag{3}$$

$$\forall\, \ell \in [p],\ \Pr_{\widehat{D}}\left[\mathcal{E}(f), \mathcal{E}'(f), \widehat{Z} = \ell\right] \geq \lambda - \eta - \Delta. \tag{4}$$

Δ acts as a relaxation parameter in Program (ErrTolerant), which can be fixed in terms of the other parameters; see Theorem 4.3. Equation (3) ensures all feasible classifiers satisfy fairness constraints with respect to the perturbed samples $\widehat{S}$. Equation (4) ensures that all feasible classifiers are $(1 - O(\eta/\lambda))$-stable (see Definition 4.7). As mentioned in Section 1, this suffices to ensure that all feasible classifiers are fair with respect to S. Finally, to ensure the accuracy guarantee the thresholds in the RHS of Equations (3) and (4) are carefully tuned to ensure that $f^\star$ is feasible for Program (ErrTolerant); see Lemma 4.9. We refer the reader to the proof overview of Theorem 4.3 at the end of this section for further discussion of Program (ErrTolerant).

Before presenting our result we require the definition of the Vapnik–Chervonenkis (VC) dimension.

Definition 4.2. *Given a finite set A, define the collection of subsets $\mathcal{F}_A \coloneqq \{\{a \in A \mid f(a) = 1\} \mid f \in \mathcal{F}\}$ .We say that $\mathcal{F}$ shatters a set B if $|\mathcal{F}_B| = 2^{|B|}$. The VC dimension of $\mathcal{F}$, $\mathrm{VC}(\mathcal{F}) \in \mathbb{N}$, is the largest integer such that there exists a set C of size $\mathrm{VC}(\mathcal{F})$ that is shattered by $\mathcal{F}$.*

Our first result bounds the accuracy and fairness metric of optimal solution f_{ET} of Program (ErrTolerant) for any hypothesis class $\mathcal{F}$ with a finite VC dimension using $O(\mathrm{VC}(\mathcal{F}))$ samples.

Theorem 4.3 (Main result). *Suppose Assumption 1 holds with constant $\lambda > 0$ and $\mathcal{F}$ has VC dimension $d \in \mathbb{N}$. Then, for all perturbation rates $\eta \in (0, \lambda/2)$, fairness thresholds $\tau \in (0, 1]$, bounds on error $\varepsilon > 2\eta$ and constraint violation $\nu > 8\eta\tau/(\lambda - 2\eta)$, and confidence parameters $\delta \in (0, 1)$ with probability at least $1 - \delta$, the optimal solution $f_{\mathrm{ET}} \in \mathcal{F}$ of Program (ErrTolerant) with parameters η, λ, and $\Delta \coloneqq O\left(\varepsilon - 2\eta, \nu - 8\eta\tau/(\lambda - 2\eta), \lambda - 2\eta\right)$, and $N = \mathrm{poly}(d, 1/\Delta, \log(p/\delta))$ perturbed samples from the η-Hamming model satisfies $\mathrm{Err}_{\mathcal{D}}(f_{\mathrm{ET}}) - \mathrm{Err}_{\mathcal{D}}(f^\star) \leq \varepsilon$ and $\Omega_{\mathcal{D}}(f_{\mathrm{ET}}) \geq \tau - \nu$.*

Thus, Theorem 4.3 shows that any procedure that outputs f_{ET}, given with a sufficiently large number of perturbed samples, (ε, ν)-learns $\mathcal{F}$ for any $\varepsilon > 2\eta$ and $\nu = O((\eta \cdot \tau)/\lambda)$. Theorem 4.3 can be extended to provably satisfy multiple linear-fractional metrics (at the same time) and work for multiple non-binary protected attributes; see Theorem E.1 in Supplementary Material E.1. Moreover, Theorem 4.3 also holds for the Nasty Sample Noise model. The proof of this result is implicit in the proof of Theorem 4.3; we present the details in Supplementary Material B.5. Finally, Program (ErrTolerant) only requires an estimate of one parameter, λ. (Since η is known, τ is fixed by the user, and Δ can be set in terms of the other parameters.) If for each $\ell \in [p]$, we also have estimates of

$\lambda_\ell \coloneqq \Pr_{\mathcal{D}}[\mathcal{E}(f^\star), \mathcal{E}'(f^\star), Z = \ell]$ and $\gamma_\ell \coloneqq \Pr_{\mathcal{D}}[\mathcal{E}'(f^\star), Z = \ell]$, then we can use this information to "tighten" Program (ErrTolerant) to the following program:

$$\begin{aligned} \min_{f \in \mathcal{F}} \quad & \mathrm{Err}_{\widehat{D}}(f), \qquad \text{(ErrTolerant+; 5)} \\ \text{s.t.,} \quad & \Omega(f, \widehat{S}) \geq \tau \cdot s \text{ and } \forall\, \ell \in [p],\ \Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}'(f), \widehat{Z} = \ell] \geq \lambda_\ell - \eta - \Delta. \end{aligned}$$

where the scaling parameter $s \in [0, 1]$ is the solution of the following optimization program

$$\min_{\eta_1, \eta_2, \ldots, \eta_p \geq 0}\ \min_{\ell, k \in [p]} \frac{1 - \eta_\ell/\lambda_\ell}{1 + (\eta_k - \eta_\ell)/\gamma_\ell} \cdot \frac{1 + (\eta_\ell - \eta_k)/\gamma_k}{1 + \eta_\ell/\lambda_k}, \quad \text{s.t.,} \quad \textstyle\sum_{\ell \in [p]} \eta_\ell \leq \eta + \Delta. \tag{6}$$

We can show that Program (ErrTolerant+) has a fairness guarantee of $(1 - s) + {}^{4\eta\tau}/_{(\lambda - 2\eta)}$ (which is always be smaller than ${}^{8\eta\tau}/_{(\lambda - 2\eta)}$) and an accuracy guarantee of 2η. We prove this result in Supplementary Material E.2. Thus, in applications where one can estimate $\lambda_1, \ldots, \lambda_p, \gamma_1, \ldots, \gamma_p$, Program (ErrTolerant+) offers better fairness guarantee than Program (ErrTolerant) (up to constants).

The proof of Theorem 4.3 appears in Supplementary Material B.

Impossibility results. We now present results complementing the guarantees of Theorem 4.3.

Theorem 4.4 (No algorithm can guarantee high accuracy *and* fairness without Assumption 1). *For all perturbation rates $\eta \in (0, 1]$, thresholds $\tau \in ({}^1/_2, 1)$, confidence parameters $\delta \in [0, {}^1/_2)$, and bounds on the error $\varepsilon \in [0, {}^1/_2)$ and constraint violation $\nu \in [0, \tau - {}^1/_2)$, if the fairness metric is statistical rate, then it is impossible to (ε, ν)-learn any hypothesis class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X} \times [p]}$ that shatters a set of 6 points of the form $\{x_A, x_B, x_C\} \times [2] \subseteq \mathcal{X} \times [p]$ for some distinct $x_A, x_B, x_C \in \mathcal{X}$.*

Suppose that $\tau = 0.8$, say to encode the 80% disparate impact rule [11]. Then, Theorem 4.4 shows that for any $\eta > 0$, any $\mathcal{F}$ satisfying the condition in Theorem 4.4 is not (ε, ν)-learnable for any $\varepsilon < {}^1/_2$ and $\nu < \tau - {}^1/_2 = {}^3/_{10}$. Intuitively, the condition on $\mathcal{F}$ avoids "simple" hypothesis classes. It is similar to the conditions considered by works on PAC learning with adversarial perturbations [12, 33], and holds for common hypothesis classes such as decision-trees and SVMs (Remark C.10). Thus, even if η is vanishingly small, without additional assumptions, any $\mathcal{F}$ satisfying mild assumptions is not (ε, ν)-learnable for any $\varepsilon < {}^1/_2$ and $\nu < {}^3/_{10}$, justifying Assumption 1. The proof of Theorem 4.4 appears in Supplementary Material C.1.

Theorem 4.5 (Fairness guarantee of Theorem 4.4 is optimal up to a constant factor). *For all perturbation rates $\eta \in (0, 1]$, confidence parameter $\delta \in [0, {}^1/_2)$, and a (known) constant $\lambda \in (0, {}^1/_4]$, if the fairness metric is statistical rate and $\tau = 1$, then given the promise that Assumption 1 holds with constant λ, for any bounds $\varepsilon < {}^1/_4 - {}^{2\eta}/_5$ and $\upsilon < {}^{\eta}/_{(10\lambda)} \cdot (1 - 4\lambda) - O\left({}^{\eta^2}/_{\lambda^2}\right)$ it is impossible to (ε, ν)-learn any hypothesis class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X} \times [p]}$ that shatters a set of 10 points of the form $\{x_A, x_B, x_C, x_D, x_E\} \times [2] \subseteq \mathcal{X} \times [p]$ for some distinct $x_A, x_B, x_C, x_D, x_E \in \mathcal{X}$.*

Suppose that $\lambda < {}^1/_8$ and $\eta < {}^1/_2$, then Theorem 4.5 shows that for any $\eta > 0$, any learner $\mathcal{L}$ that has a constraint violation bound $\nu < {}^{\eta}/_{(20\lambda)} - O({}^{\eta^2}/_{\lambda^2})$, must have a large error bound $\varepsilon \geq {}^1/_{20}$ to (ε, ν)-learn any $\mathcal{F}$ satisfying a mild assumption. When η/λ is small, this shows that any learner whose fairness guarantee more than a constant amount smaller than the fairness guarantee in Theorem 4.3, must have a significantly larger error guarantee. Like Theorem 4.4, the condition on $\mathcal{F}$ in Theorem 4.5 avoids "simple" hypothesis classes and holds for common hypothesis (Remark C.10). Finally, complementing our accuracy guarantee, we prove that for any $\varepsilon < \eta$, no algorithm can (ε, ν)-learnable any hypothesis classes $\mathcal{F}$ satisfying mild assumptions (Theorem D.1); its proof appears in Supplementary Material D. Thus, the accuracy guarantee in Theorem 4.5 is optimal up to constant factors. The proof of Theorem 4.5 appears in Supplementary Material C.2.

Proof overview of Theorem 4.3. We explain the key ideas behind Program (ErrTolerant) and how they connect with the proof of Theorem 4.3. Our goal is to construct error-tolerant constraints using perturbed samples $\widehat{S}$ such that the classifier f_{ET}, that has the smallest error on $\widehat{S}$ subject to satisfying these constraints, has accuracy 2η-close to that of $f^\star$ and that additively violates the fairness constraints by at most $O(\eta/\lambda)$.

Step 1: Lower bounding the accuracy of f_{ET}. This step relies on Lemma 4.6.

Lemma 4.6. *For any bounded function $g\colon \{0, 1\}^2 \times [p] \to [0, 1]$, $\delta, \Delta \in (0, 1)$, and adversaries $A \in \mathcal{A}(\eta)$, given $N = \mathrm{poly}({}^1/_\Delta, \mathrm{VC}(\mathcal{F}), \log 1/\delta)$ true samples $S \sim \mathcal{D}$ and corresponding perturbed samples $A(S) \coloneqq \{(x_i, y_i, \widehat{z}_i)\}_{i \in [N]}$, with probability at least $1 - \delta$, it holds that*

$$\forall\, f \in \mathcal{F}, \quad \left| \tfrac{1}{N} \textstyle\sum_{i \in [N]} g(f(x_i, \widehat{z}_i), y_i, \widehat{z}_i) - \mathbb{E}_{(X, Y, Z) \sim \mathcal{D}}\left[g(f(X, Z), Y, Z)\right] \right| \leq \Delta + \eta.$$

The proof of Lemma 4.6 follows from generalization bounds for bounded functions (e.g., see [43]) and because the η-Hamming model perturbs at most $\eta \cdot N$ samples. Let g be the 0-1 loss (i.e., $g(\widetilde{y}, y, z) \coloneqq \mathbb{I}[\widetilde{y} \neq y]$), then for all $f \in \mathcal{F}$, Lemma 4.6 shows that the error of f on samples drawn from $\mathcal{D}$ and samples in $\widehat{S}$ are close: $|\mathrm{Err}_{\mathcal{D}}(f) - \mathrm{Err}(f, \widehat{S})| \leq \Delta + \eta$. Thus, intuitively, minimizing $\mathrm{Err}(f, \widehat{S})$ could be a good strategy to minimize $\mathrm{Err}_{\mathcal{D}}(f)$. Then, if $f^\star$ is feasible for Program (ErrTolerant), we can bound the error of f_{ET}: Since f_{ET} is optimal for Program (ErrTolerant), its error on $\widehat{S}$ is at most the error of $f^\star$ on $\widehat{S}$. Using this and applying Lemma 4.6 we get that

$$\mathrm{Err}_{\mathcal{D}}(f_{\mathrm{ET}}) \leq \mathrm{Err}(f_{\mathrm{ET}}, \widehat{S}) + \eta + \Delta \;\leq\; \mathrm{Err}(f^\star, \widehat{S}) + \eta + \Delta \leq \mathrm{Err}_{\mathcal{D}}(f^\star) + 2(\eta + \Delta). \tag{7}$$

Step 2: Lower bounding the fairness of f_{ET}. One could try to bound the fairness of f_{ET} using the same approach as Step 1, i.e., show that for all $f \in \mathcal{F}$: $|\Omega_{\mathcal{D}}(f) - \Omega(f, \widehat{S})| \leq O(\eta/\lambda)$. Then ensuring that f has a high fairness on $\widehat{S}$ implies that it also has high fairness on S (up to $O(\eta/\lambda)$ factor). However, such a bound does not hold for any $\mathcal{F}$ satisfying mild assumptions (see Example G.2). The first idea is to prove a similar (in fact, stronger multiplicative) bound on a subset of $\mathcal{F}$. Toward this, we define:

Definition 4.7. *A classifier $f \in \mathcal{F}$ is said to be s-stable for fairness metric Ω, if for all adversaries $A \in \mathcal{A}(\eta)$, w.h.p. over draw of $S \sim \mathcal{D}$, it holds that $\Omega_{\mathcal{D}}(f)/\Omega(f,\widehat{S}) \in [s, 1/s]$, where $\widehat{S} \coloneqq A(S)$.*

If a s-stable classifier f has fairness τ on $\widehat{S}$, then it has a fairness at least $\tau \cdot s$ on $\mathcal{D}$ w.h.p. Thus, if we have a condition such that any feasible $f \in \mathcal{F}$ satisfying this condition is s-stable, then any classifier satisfying this condition and the fairness constraint, $\Omega(\cdot, \widehat{S}) \geq \tau/s$, must have a fairness at least τ on $\mathcal{D}$ w.h.p. The key idea is coming up such constraints.

Lemma 4.8. *Any classifier $f \in \mathcal{F}$ that satisfies $\min_{\ell \in [p]} \Pr_{\mathcal{D}}[\mathcal{E}(f), \mathcal{E}'(f), \widehat{Z} = \ell] \geq \lambda + \eta + \Delta$, is $(\frac{1-(\eta+\Delta)/\lambda}{1+(\eta+\Delta)/\lambda})^2$-stable for fairness metric Ω (defined by events $\mathcal{E}$ and $\mathcal{E}'$).*

Step 3: Requirements for the error-tolerant program. Building on Steps 1 and 2, we prove Lemma 4.9.

Lemma 4.9. *If the following conditions hold then, $\mathrm{Err}_{\mathcal{D}}(f_{\mathrm{ET}}) - \mathrm{Err}_{\mathcal{D}}(f^\star) \leq 2\eta$ and $\Omega_{\mathcal{D}}(f_{\mathrm{ET}}) \geq \tau - O(\eta/\lambda)$: (C1) $f^\star$ is feasible for Program (ErrTolerant), and all $f \in \mathcal{F}$ feasible for Program (ErrTolerant) are (C2) s-stable for $s = 1 - O(\eta/\lambda)$, and satisfy (C3) $\Omega(f, \widehat{S}) \geq \tau \cdot (1 - O(\eta/\lambda))$.*

Thus, it suffices to find error-tolerant constraints that satisfy conditions (C1) to (C3). Condition (C3) can be satisfied by adding the constraint $\Omega(\cdot, \widehat{S}) \geq \tau'$, for $\tau' = \tau \cdot (1 - O(\eta/\lambda))$. From Lemma 4.8, condition (C2) follows by using the constraint in $\min_{\ell \in [p]} \Pr_{\mathcal{D}}[\mathcal{E}(f), \mathcal{E}'(f), \widehat{Z} = \ell] \geq \lambda'$, for $\lambda' \geq \Theta(\lambda)$. It remains to pick τ' and λ' such that condition (C1) also holds. The tension in setting τ' and λ' is that if they are too large then condition (C1) does not hold and if they are too small then conditions (C2) and (C3) do not hold. In the proof we show that $\tau' \coloneqq \tau \cdot (\frac{1-(\eta+\Delta)/\lambda}{1+(\eta+\Delta)/\lambda})^2$ and $\lambda' \coloneqq \lambda - \eta - \Delta$ suffice to satisfy conditions (C1) to (C3) (this is where we use Assumption 1).

Overall the main technical idea is to identify the notion of s-stable classifiers and sufficient conditions for a classifier to be s-stable; combining these conditions with the fairness constraints on $\widehat{S}$, ensures that f_{ET} has high fairness on S, and carefully tuning the thresholds so that $f^\star$ is likely to be feasible for Program (ErrTolerant) ensures that f_{ET} has an accuracy close to $f^\star$.

Proof overviews of Theorems 4.4 and 4.5. Our proofs are inspired by [33, Theorem 1] and [12, Theorem 1] which consider PAC learning with adversarial corruptions. In both Theorems 4.4 and 4.5, for some $\varepsilon, \nu \in [0, 1]$, the goal is to show that given samples perturbed by an η-Hamming adversary, (possibly under additional assumptions), no learner $\mathcal{L}$ can output a classifier that has accuracy ε-close to the accuracy of $f^\star$ and that additively violates the fairness constraints by at most ν. Say a classifier $f \in \mathcal{F}$ is "good" if it satisfies the required conditions. The approach is to construct two or more distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$ that satisfy the following conditions: (C1) For any ℓ, k, given a iid draw S from $\mathcal{D}_\ell$, an η-Hamming adversary can add perturbations such that w.h.p. $\widehat{S}$ is distributed according to iid samples from $\mathcal{D}_k$. Thus $\mathcal{L}$, who only sees $\widehat{S}$, w.h.p., cannot identify the original distribution of S and is forced to output a classifier that is good for all $\mathcal{D}_1, \ldots, \mathcal{D}_m$. The next condition ensures that this is not possible. (C2) No classifier $f \in \mathcal{F}$ is good for all $\mathcal{D}_1, \ldots, \mathcal{D}_m$, and for each $\mathcal{D}_i$ ($i \in [m]$), there is at least one good classifier $f_i \in \mathcal{F}$. (The latter half ensures that requirements are not vacuously satisfied.) Thus, for all $\mathcal{L}$ there is some distribution in $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$ for which it outputs a bad classifier. (Note that even if the learner is randomized, it must fail with probability at least $1/m$.)

The key idea in the proofs is to come up with distributions satisfying the above conditions. [33, 12] follow the same outline in the context of PAC learning, however, as we also consider fairness constraints, our constructions end up being very different from their constructions. The assumptions on $\mathcal{F}$ ensure that condition (C2) is satisfiable. For example, if $\mathcal{F}$ has less than m hypothesis, then condition (C2) cannot be satisfied. Full details appear in Supplementary Material C.

5 Empirical results

We implement our approach using the logistic loss function with linear classifiers and evaluate its performance on real world and synthetic datasets.

Metrics and baselines. The selection of an appropriate fairness metric is context-dependent and beyond the scope of this work [45]; for illustrative purposes we (arbitrarily) consider the statistical rate (SR) and compare an implementation of our framework (Program (ErrTolerant+)), **Err-Tol**, with state-of-the-art fair classification frameworks for SR under stochastic perturbations: **LMZV** [35] and **CHKV** [15]. **LMZV** and **CHKV** take parameters $\delta_L, \tau \in [0, 1]$ as input; these parameters control the desired fairness, where decreasing δ_L or increasing τ increases the desired fairness. We also compare against **KL** [34], which controls for true-positive rate (TPR) in the presence of a Malicious adversary, and **AKM** [7] that is the post-processing method of [29] and controls for equalized-odds fairness constraints (EO). We also compare against the optimal unconstrained classifier, **Uncons**; this is the same as [12]'s algorithm for PAC-learning in the Nasty Sample Noise Model without fairness constraints. We provide additional comparisons using our framework with false-positive rate (FPR) as the fairness metric and additional baselines in Supplementary Material F.

Implementation details. We use a randomly generated 70-30 train (S) test (T) split of the data, and generate the perturbed dataset $\widehat{S}$ from S for a (known) perturbation rate η. We train each algorithm on $\widehat{S}$, and report the accuracy (acc) and statistical rate (SR) of the output classifiers on the (unperturbed) test dataset T. **Err-Tol** is given the perturbation rate η. To advantage the baselines in our comparison, we provide them with even more information as needed by their approaches: **LMZV** and **CHKV** are given group-specific perturbation rates: for each $\ell \in [p]$, $\eta_\ell := \Pr_D[\widehat{Z} \neq Z \mid Z = \ell]$, and **KL** is given η and for each $\ell \in [p]$, the probability $\Pr_D[Z = \ell, Y = 1]$; where D is the empirical distribution of S. **Err-Tol** implements Program (ErrTolerant+) which requires estimates of λ_ℓ and γ_ℓ for all $\ell \in [p]$. As a heuristic, we set $\gamma_\ell = \lambda_\ell := \Pr_{\widehat{D}}[Z = \ell]$, where $\widehat{D}$ is the empirical distribution of $\widehat{S}$. We find that these estimates suffice, and expect that a more refined approach would only improve the performance of **Err-Tol**.

Adversaries. We consider two η-Hamming adversaries (which we call A_{TN} and A_{FN}); each one computes the "optimal fair classifier" $f^\star$, which has the highest accuracy (on S) subject to having SR at least τ on S. A_{TN} considers the set of all true negatives of $f^\star$ that have protected attribute $Z = 1$, selects the $\eta \cdot |S|$ samples that are furthest from the decision boundary of $f^\star$, and perturbs their protected attribute to $\widehat{Z} = 2$. A_{FN} is similar, except that it considers the set of false negatives of $f^\star$. Both adversaries try to increase the performance of $f^\star$ on $Z = 1$ in $\widehat{S}$ by removing the samples that $f^\star$ predicts as negative; thus, increasing $f^\star$'s SR. The adversary's hope is that choosing samples far from the decision boundary would (falsely) give the appearance of a high SR on $\widehat{S}$. This would make a fair classification algorithms select unfair classifiers with higher accuracy. Note that these are not intended to be "worst-case" adversaries; As **Err-Tol** comes with provable guarantees, we expect it to perform well against other adversaries while other approaches may have even worse performance. .

Simulation on synthetic data. We first show empirically that perturbations by the η-Hamming adversary can be prohibitively disruptive for methods that attempt to correct for stochastic noise. We consider a synthetic dataset with 1,000 samples from two equally-sized protected groups; each sample has a binary protected attribute, two continuous features $x_1, x_2 \in \mathbb{R}$, and a binary label. Conditioned on the protected attribute, (x_1, x_2) are independent draws from a mixture of 2D Gaussians (see Figure 4). This distribution and the labels are such that a) one group has a higher likelihood of a positive label than the other, and b) **Uncons** has a near-perfect accuracy ($> 99\%$) and a statistical rate of 0.8 on S. Similar to **Uncons**, we consider a fairness constraint of $\tau = 0.8$. Thus, in the absence of noise, this is an "easy case:" where **Uncons** satisfies the fairness constraints. We generate $\widehat{S}$ using A_{TN}, and compare against **CHKV**, which was developed for correcting stochastic perturbations.

Results. The fairness and statistical rate averaged over 50 iterations is reported in Table 1 as a function of the perturbation η. At $\eta = 0$, both **CHKV** and **Err-Tol** nearly-satisfy the fairness constraint (SR≥ 0.79) and have a near-perfect accuracy (acc≥ 0.98). However, as η increases, while **CHKV** retains the same statistical rate (~ 0.8), it loses a significant amount of accuracy ($\sim 20\%$). In contrast, **Err-Tol** has high accuracy and fairness (acc≥ 0.99 and SR≥ 0.79) for all η considered. Hence, this shows that stochastic approaches may fail to satisfy their guarantees under the η-Hamming model.

Simulations on real-world data. In this simulation, we show that our framework can outperform each baseline with respect to the accuracy-fairness trade-off under perturbations from the adversaries we consider, and does not under-perform compared to baselines under perturbations from either

Table 1: *Simulation on synthetic data:* We run **CHKV** and **Err-Tol** with $\tau = 0.8$ on a synthetic dataset and report their average accuracy and statistical rate (std. deviation in parenthesis). The result shows that prior approaches can fail to satisfy their guarantees under the η-Hamming model.

	acc ($\eta = 0\%$)	SR ($\eta = 0\%$)	acc ($\eta = 3\%$)	SR ($\eta = 3\%$)	acc ($\eta = 5\%$)	SR ($\eta = 5\%$)
Unconstrained	1.00 (.001)	.799 (.001)	1.00 (.000)	.799 (.002)	1.00 (.001)	.800 (.001)
CHKV (τ=.8)	1.00 (.001)	.800 (.002)	.859 (.143)	.787 (.015)	.799 (.139)	.795 (.049)
Err-Tol (τ=.8)	.985 (.065)	.800 (.001)	1.00 (.001)	.799 (.002)	.999 (.002)	.799 (.004)

adversary. The COMPAS data in [10] contains 6,172 samples with 10 binary features and a label that is 1 if the individual did not recidivate and 0 otherwise; the SR of **Uncons** on COMPAS is 0.78.) We take gender (coded as binary) to be the protected attribute, and set the fairness constraint on the SR to be $\tau = 0.9$ for **Err-Tol** and all baselines. We consider both adversaries A_{TN} and A_{FN}, and a perturbation rate of $\eta = 3.5\%$ as 3.5% is roughly the smallest value for η necessary to ensure that the optimal fair classifier $f^\star$ for $\tau = 0.9$ (on S) has a SR less than 0.78 on the perturbed data.

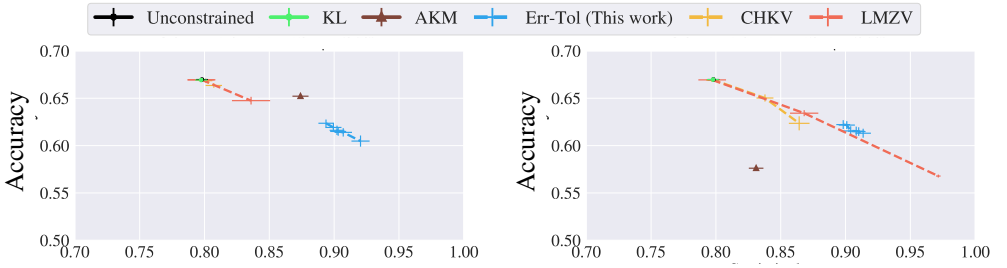
Results. The accuracy and statistical rate (SR) of **Err-Tol** and baselines for $\tau \in [0.7, 1]$ and $\delta_L \in [0, 0.1]$ and averaged over 100 iterations is reported in Figure 1. For both adversaries, **Err-Tol** attains a better SR than the unconstrained classifier (**Uncons**) for a small trade-off in accuracy. For adversary A_{TN} (Figure 1(a)), **Uncons** has SR (0.80) and accuracy (0.67). In contrast, **Err-Tol** achieves high SR (0.92) with a trade-off in accuracy (0.60). In comparison, **AKM** has a higher accuracy (0.65) but a lower SR (0.87), and other baselines have an even lower SR (≤ 0.84) with accuracy comparable to **AKM**. For adversary A_{FN} (Figure 1(b)), **Uncons** has SR (0.80) and accuracy (0.67), while **Err-Tol** has a significantly higher SR (0.91) and accuracy (0.61). This significantly outperforms **AKM** which has SR (0.83) and accuracy (0.58). **LMZV** achieves the highest SR (0.97) with a natural reduction in accuracy to (0.57). In this case, **Err-Tol** has similar accuracy to SR trade-off as **LMZV**, but achieves a lower maximum SR (0.91). Meanwhile, **Err-Tol** has a significantly higher SR trade-off than **CHKV** at the same accuracy. We further evaluate our framework under stochastic perturbations in Supplementary Material F (specifically, against the perturbation model of [15]) and observe similar statistical rate and accuracy trade-offs as approaches [15, 35] tailored for stochastic perturbations.

6 Limitations and conclusion

This work extends fair classification to real-world settings where perturbations in the protected attributes may be correlated or affect arbitrary subsets of samples. We consider the η-Hamming model and give a framework that outputs classifiers with provable guarantees on both fairness and accuracy; this framework works for categorical protected attributes and a class of linear-fractional fairness constraints. We show near-tightness of our framework's guarantee and extend it to the Nasty Sample Noise model, which can perturb both labels and features. Empirically, classifiers produced by our framework achieve high fairness at a small cost to accuracy and outperform existing approaches.

Compared to existing frameworks for fair classification with stochastic perturbations, our framework requires less information about the perturbations. However, its efficacy will depend on an appropriate choice of parameters; e.g., an overly conservative λ can decrease accuracy and an optimistic λ can decrease fairness. A careful assessment both pre- and post-deployment would be important in order to avoid negative social implications in a misguided attempt to do good [36].

Finally, we note that discrimination is a systematic problem and our work only addresses one part of it; this work would be effective as one piece of a broader approach to mitigate and rectify biases.



(a) A_{TN} (b) A_{FN}

Figure 1: *Simulations on COMPAS data:* Perturbed data is generated using adversary A_{TN} (a) and A_{FN} (b) as described in Section 5 with $\eta = 3.5\%$. All algorithms are run on the perturbed data varying the fairness parameters ($\tau \in [0.7, 1]$ and $\delta_L \in [0, 0.1]$). The y-axis depicts accuracy and the x-axis depicts statistical rate (SR); both values are computed over the unperturbed test set. We observe that for both adversaries our approach **Err-Tol**, attains a better fairness than the unconstrained classifier with a natural trade-off in accuracy. Further, **Err-Tol** achieves a better fairness-accuracy trade-off than each baseline on at least one of (a) or (b). Error bars represent the standard error of the mean.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A Reductions Approach to Fair Classification. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 2018.

[2] Dana Angluin and Philip D. Laird. Learning from noisy examples. *Mach. Learn.*, 2(4):343–370, 1987.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. `https://github.com/propublica/compas-analysis`, 2016.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. COMPAS recidivism risk score data and analysis, 2016.

[5] Peter Auer. Learning with malicious noise. In *Encyclopedia of Algorithms*, pages 1086–1089. 2016.

[6] Peter Auer and Nicolo Cesa-Bianchi. On-line learning with malicious noise and the closure algorithm. *Annals of mathematics and artificial intelligence*, 23(1):83–99, 1998.

[7] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics*, pages 1770–1780. PMLR, 2020.

[8] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 32:15479–15488, 2019.

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[10] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018.

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

[12] Nader H Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

[13] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 2018.

[14] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *FAT*, pages 319–328. ACM, 2019.

[15] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Fair classification with noisy protected attributes. *CoRR*, abs/2006.04778, 2020.

[16] Nicolo Cesa-Bianchi, Eli Dichterman, Paul Fischer, Eli Shamir, and Hans Ulrich Simon. Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM (JACM)*, 46(5):684–719, 1999.

[17] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *FAT*, pages 339–348. ACM, 2019.

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

[19] Andrew J Coldman, Terry Braun, and Richard P Gallagher. The classification of ethnic status using name information. *Journal of Epidemiology & Community Health*, 42(4):390–395, 1988.

[20] N.R. Council, D.B.S.S. Education, C.N. Statistics, P.D.C.R.E. Data, E. Perrin, and M.V. Ploeg. *Eliminating Health Disparities: Measurement and Data Needs*. National Academies Press, 2004.

[21] Bill Dedman. The color of money. *Atlanta Journal-Constitution*, pages 1–4, 1988.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[23] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Mark D. M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *FAT*, volume 81 of *Proceedings of Machine Learning Research*, pages 119–133. PMLR, 2018.

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

[25] Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-discriminatory machine learning through convex fairness criteria. In *AAAI 2018*, 2018.

[26] Gabriel Goh, Andrew Cotter, Maya R. Gupta, and Michael P. Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2415–2423, 2016.

[27] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR (Poster)*, 2015.

[28] Richard W Hamming. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.

[29] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323, 2016.

[30] Mara Hvistendahl. Can "predictive policing" prevent crime before it happens. *Science Magazine*, 28, 2016.

[31] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[32] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. In *FAT**, page 110. ACM, 2020.

[33] Michael J. Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM J. Comput.*, 22(4):807–837, 1993.

[34] Nikola Konstantinov and Christoph H. Lampert. Fairness-aware learning from corrupted data. *CoRR*, abs/2102.06004, 2021.

[35] Alexandre Louis Lamy and Ziyuan Zhong. Noise-tolerant fair classification. In *NeurIPS*, pages 294–305, 2019.

[36] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[37] Elizabeth Luh. Not so black and white: Uncovering racial bias from systematically misreported trooper reports. *Available at SSRN 3357063*, 2019.

[38] Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *FAT 2018*, pages 107–118, 2018.

[39] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R Varshney. Understanding unequal gender classification accuracy from face images. *arXiv preprint arXiv:1812.00099*, 2018.

[40] Northpointe. Compas risk and need assessment systems. `http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf`, 2012.

[41] Catherine Saunders, Gary Abel, Anas El Turabi, Faraz Ahmed, and Georgios Lyratzopoulos. Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: Evidence from the english cancer patient experience survey. *BMJ open*, 3, 06 2013.

[42] Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on learning theory*, pages 489–511. PMLR, 2013.

[43] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[44] Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948.

[45] Megha Srivastava, Hoda Heidari, and Andreas Krause. 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*, pages 2459–2468, 2019.

[46] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[47] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya R. Gupta, and Michael I. Jordan. Robust optimization for fairness with noisy protected groups. In *NeurIPS*, 2020.

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

[49] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970. PMLR, 2017.

[50] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Theorems 4.3 to 4.5 in Section 4 and see Table 1 and Figure 1 in Section 5.
   (b) Did you describe the limitations of your work? [Yes] We discuss some limitations of our work in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Section 6 discusses the importance of identifying the right parameters and using this as only one piece of a larger framework for mitigating discrimination.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Yes, we discuss potential negative social impacts of our framework in Section 6.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All theorem statements list the assumptions they require. For example, Theorem 4.3 states Assumption 1.
   (b) Did you include complete proofs of all theoretical results? [Yes] The proof of Theorem 4.3 appears in Supplementary Material B. The proof of Theorems 4.4 and 4.5 appear in Supplementary Material C. The formal statement of Theorem D.1 and its proof appear in Supplementary Material D. The formal statements and proofs of extensions of Theorem 4.3 appear in Supplementary Material E. We also overview the proofs of the main theoretical results (Theorems 4.3 to 4.5) in Section 4.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We submitted the code in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The hyper-parameters and other implementation details appear in Supplementary Material F.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] E.g., Table 1 reports the standard deviations and Figure 1 reports the standard error of the mean.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We report the total computational resources used for this work in Supplementary Material F.1.4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use the preprocessed version of the COMPAS data [3] provided by [10]. We cite both [3] and [10] in Section 5.
   (b) Did you mention the license of the assets? [No] To the best of knowledge, the COMPAS data is not licensed.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include the code for our simulations in the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]