Fair Classification with Adversarial Perturbations

L. Elisa Celis
Yale University

Anay Mehrotra
Yale University

Nisheeth K. Vishnoi
Yale University

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 [50], credit lending [21], and predictive policing [34]. 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 [32, 26, 28, 17, 63, 62, 48, 24, 27, 1, 13]. 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 protected 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 maximizes accuracy constrained to being fair with respect to the given fairness metric (and Z) [8]. To specify fairness constraints, these approaches need protected attributes of training data to be known.

However, protected attributes can be erroneous for various reasons; there could be uncertainties during data collection or data cleaning process [20, 52], or the attributes could be strategically misreported [46]. 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, 66, 35]. In these cases, protected attributes can be "imputed" [18, 36, 16], but this can also introduce errors [12]; further, imputation by machine-learning-based methods is known to be fragile to imperceptible changes in the inputs [29] and to have correlated errors across samples [49]. 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., [16, 7].

Towards addressing this problem, several recent works have developed fair classification algorithms for various models of errors in the protected attributes. [44] consider an extension of the "mutually contaminated learning model" [53] where, instead of observing samples from the "true" joint distribution, distributions of observed group-conditional distributions are stochastic mixtures of their true counterparts. [6] consider a binary protected attribute and Bernoulli perturbations that are

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

independent of the labels (and of each other). [14] consider the setting where each sample's protected attribute is independently flipped to a different value with a known probability. [59] 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 data containing independent draws of both the true and perturbed protected attributes. In their distributionally robust 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, [40] study fair classification under the Malicious noise model [56, 39] 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 a 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 [31] to the fair classification setting and we refer to it as the η-Hamming model. Unlike 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 data. 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 D.2. Taking these perturbed samples, a fairness metric , and a desired fairness threshold τ as input, the goal is to learn a classifier f with the maximum accuracy with respect to the true distribution $\mathcal{D}$ subject to having a fairness value, $_{D}(f)$, of at least τ with respect to the true distribution $\mathcal{D}$.

Our contributions. We present an optimization framework (Definition 4.1) that outputs fair classifiers for the η-Hamming model and 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 [13]. The framework provably outputs a classifier whose accuracy is within 2η of the accuracy of $f^\star$ and which violates the fairness constraint by at most $O(\eta/\lambda)$ additively (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 [14] 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 give a non-trivial guarantee on *both* accuracy and fairness value of the output classifier (Theorem 4.4). Moreover, we prove the near-tightness of our framework's guarantee under Assumption 1: No algorithm can guarantee to output a classifier with accuracy closer than η to that of $f^\star$ and any algorithm that violates the fairness constraint by less than $\eta/(20\lambda)$ additively has an accuracy at most $19/20$ (Theorems 4.5 and A.21). Finally, we also extend our framework's guarantees to the Nasty Sample Noise model (Supplementary Material A.1.5). The Nasty Sample Noise model is a generalization of the η-Hamming model, which was studied by [11] in the context of PAC learning (without any fairness considerations), where the adversary can choose any η-fraction of the samples, and can arbitrarily perturb both their labels and features.

We implement our framework for logistic loss function with linear classifiers and evaluate its performance on COMPAS [3], Adult [23], 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 several baselines [44, 6, 59, 14, 40] with statistical rate and false-positive rate as fairness metrics.[1] On the synthetic dataset, we compare against a method developed for fair classification under stochastic perturbations [14] and demonstrate the comparative strength of the η-Hamming model; our results show that [14]'s framework achieves a significantly lower accuracy than our framework for the same statistical rate. Empirical results on COMPAS and Adult show that the classifier output by our framework can attain better statistical rate and false-positive rate than the accuracy maximizing classifier on the true distribution, with a small loss in accuracy. Further, our framework has a similar (or better) fairness-accuracy trade-off compared to all baselines we consider in a variety of settings, and is not dominated by any other approach (Figure 1 and Figures 7 and 8 in Supplementary Material E.2).

[1] Let $q_\ell(f, \mathsf{SR})$ (respectively $q_\ell(f, \mathsf{FPR})$) be the fraction of positive predictions (respectively false-positive predictions) by f in the ℓ-th protected group. f's statistical rate (respectively false-positive rate) is the ratio of the minimum value to the maximum value of $q_\ell(f, \mathsf{SR})$ (respectively $q_\ell(f, \mathsf{FPR})$) over all protected groups.

Techniques. The starting point of our optimization framework (Definition 4.1) is the "standard" optimization program for fair classification in the *absence* of any perturbations: Given a fairness metric and a desired fairness threshold τ as input, find $f^\star \in \mathcal{F}$ that maximizes the accuracy on the *given data* $\widehat{S}$ constrained to a fairness value at least τ on the given data. However, when $\widehat{S}$ is given to us by an η-Hamming adversary, this standard program, which imposes the fairness constraints with respect to the perturbed data $\widehat{S}$, may output a classifier with an accuracy/fairness-value worse than that of $f^\star$ when measured with respect to $\mathcal{D}$. But, observe that the difference in accuracies of a classifier when measured with respect to the given data $\widehat{S}$ and data sampled from $\mathcal{D}$ is at most η. Thus, if $f^\star \in \mathcal{F}$ is feasible for the standard optimization program, this observation (used twice) implies that the accuracy of the output classifier measured with respect to $\mathcal{D}$ is within 2η of the accuracy of $f^\star$ measured with respect $\mathcal{D}$ (Equation (8)). However, without any modifications, the classifier output by the standard optimization program could still have a fairness value much lower than τ with respect to $\mathcal{D}$ (see Example A.27). To bypass this, we introduce the notion of s-stability that allows us to lower bound the fairness value of a classifier with respect to $\mathcal{D}$ given its fairness value on $\widehat{S}$. Roughly, $f \in \mathcal{F}$ is said to be s-stable with respect to a fairness metric if for any $\widehat{S}$ that is generated by an η-Hamming adversary, the ratio of fairness value of f with respect to $\mathcal{D}$ and with respect to $\widehat{S}$ is between s and $1/s$ (see Definition 4.7). It follows that any s-stable classifier that has fairness value $\tau' > 0$ with respect to $\widehat{S}$, has fairness value at least $s \cdot \tau'$ with respect to $\mathcal{D}$. Hence, an optimization program that ensures that all feasible classifiers are s-stable (for a suitable choice of s) and have fairness value at least $\tau' > 0$ with respect to $\widehat{S}$, comes with a guarantee that any feasible classifier has a fairness value at least $s \cdot \tau'$ (with respect to $\mathcal{D}$). If such an optimization program could further ensure that $f^\star$ is feasible for it, then by arguments presented above, the classifier output by this optimization program would satisfy required guarantees on both fairness and accuracy (Lemma 4.9). The issue is that, to directly enforce s-stability, one needs to compute the fairness values of classifiers with respect to $\mathcal{D}$, but this is not possible in the absence of samples from $\mathcal{D}$. We overcome this by present a "proxy" constraint on the classifier (Equation (5)) that involves only $\widehat{S}$ and ensures that any classifier that satisfies it is s-stable. Moreover, $f^\star$ satisfies this constraint under Assumption 1. Overall, modifying Program (2) to include this constraint (Equation (5)) with a suitable value of s, and setting an appropriate fairness threshold τ so that $f^\star$ remains feasible, leads us to our framework.

2 Related work

In this section, we situate this paper in relation to lines of work which also consider fair classification with perturbed protected attributes; additional related work (e.g., on fair classification in the absence of protected attributes) are presented in Supplementary Material D.1.

[44] give a framework which comes with provable guarantees on the accuracy and fairness value of output classifiers for a binary protected attribute and either statistical rate or equalized-odds fairness metrics. [6] identify conditions on the distribution of perturbations under which the post-processing algorithm of [32] improves the fairness value of the accuracy-maximizing classifier with respect to equalized-odds on the true distribution with a binary protected attribute. [59] consider a non-binary protected attribute. In their "soft-weights" approach, they give provable guarantees on the accuracy (with respect to $f^\star$) and fairness value of the output classifier *in expectation* and in their distributionally robust approach, they give provable guarantees on the fairness value of the output classifiers.[2] [14] give provable guarantees on the accuracy and fairness value of output classifiers for multiple non-binary protected attributes and the class of linear-fractional metrics. All of the aforementioned works [44, 6, 59, 14] consider stochastic perturbation models, which are weaker than the model considered in this paper. Further, compared to [44, 6], our approach (and that of [14]) can handle multiple categorical protected attributes and multiple linear-fractional metrics (which include statistical rate and can ensure equalized-odds constraints). Compared to [6, 59], our work (and those of [44, 14]) give provable guarantees on the accuracy (with respect to $f^\star$) and fairness value of output classifiers *with high probability*. In another related work, [40] 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 value with respect to the true-positive rate fairness metric. They study the Malicious noise model, which can modify a uniformly randomly selected subset of samples arbitrarily; this is weaker than the Nasty Sample Noise model [11, 4], and hence, than the model considered in this paper. Further, our

[2] Supplementary Material D.2.3 gives an example where [59]'s distributionally robust approach outputs a classifier whose accuracy is arbitrarily close to $1/2$.

framework works without the realizable assumption (i.e., in the agnostic setting), can handle multiple and non-binary protected attributes, and can ensure fairness with respect to multiple linear-fractional metrics (which include 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 [39, 2, 11, 15, 5]; see also [4]. In particular, [11] study PAC learning (without fairness constraints) under the Nasty Sample Noise model. They use the empirical risk minimization framework (see, e.g., [54]) run on the perturbed samples to output a classifier. 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 [11]'s framework is guaranteed to output a classifier with high accuracy, it can perform poorly on fairness metrics; see Section 5 and Supplementary Material D.2.1.

3 Model

Let the data domain be $D := \mathcal{X} \times \{0,1\} \times [p]$, 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 $\mathcal{F} \subseteq \{0,1\}^{X \quad [p]}$ be a hypothesis class of binary classifiers. For $f \in \mathcal{F}$, let $\mathsf{Err}_D(f) := \mathsf{Pr}_{(X,Y,Z) \quad D}[f(X,Z) \neq Y]$ denote f's predictive error on draws from $\mathcal{D}$. In the vanilla classification problem, the learner $\mathcal{L}$'s goal is to find a classifier with minimum error: $\mathsf{argmin}_{f 2 F} \mathsf{Err}_D(f)$. 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 [13]**). *Given $f \in \mathcal{F}$ and two events $\mathcal{E}(f)$ and $\mathcal{E}^0(f)$, that can depend on f, define the performance of f on $Z = \ell$ ($\ell \in [p]$) as $q_\ell(f) := \mathsf{Pr}_D[\mathcal{E}(f) \mid \mathcal{E}^0(f), Z = \ell]$. If $\mathcal{E}^0$ 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 $\mathcal{E} := (f = 1)$ and $\mathcal{E}^0 := \emptyset$, we get statistical rate (a linear metric).[3] For $\mathcal{E} := (f = 1)$ and $\mathcal{E}^0 := (Y = 0)$, we get false-positive rate (also a linear metric). For $\mathcal{E} := (Y = 0)$ and $\mathcal{E}^0 := (f = 1)$, we get false-discovery rate (a linear-fractional metric). Given a performance metric q, the corresponding fairness metric is defined as

$$\quad _D(f) := \frac{\mathsf{min}_{\ell 2 [p]}\, q_\ell(f)}{\mathsf{max}_{\ell 2 [p]}\, q_\ell(f)}. \tag{1}$$

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

$$\mathsf{min}_{f 2 F} \mathsf{Err}_D(f) \quad \text{s.t.,} \quad \quad _D(f) \geq \tau. \tag{2}$$

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 $N \in \mathsf{N}$ samples from $\mathcal{D}$ *exactly once*. On receiving the request, A draws N independent samples $S := \{(x_i, y_i, z_i)\}_{i 2 [N]}$ 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 $\widehat{S} := \{(x_i, y_i, \widehat{z}_i)\}_{i 2 [N]}$. 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 (2).

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 $\mathcal{F} \subseteq \{0,1\}^{X \quad [p]}$ with perturbation rate $\eta \in [0,1]$ and confidence $\delta \in (0,1)$ if for all*

[3] We overload the notation f to denote both the classifier as well as its prediction, and the terms, statistical rate and false-positive rate, to refer to both the linear/linear-fractional metric q and the resulting fairness metric $\quad$.

- *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 $\widehat{S} := A(S)$, $\mathcal{L}$ outputs $f \in \mathcal{F}$ that satisfies $\mathrm{Err}_D(f) - \mathrm{Err}_D(f^\star) \leq \varepsilon$ and $\quad_D(f) \geq \tau - \nu$, where $f^\star$ is the optimal solution of Program (2) (i.e., $f^\star := \mathrm{argmin}_{f \in \mathcal{F}} \mathrm{Err}_D(f)$, s.t., $\quad_D(f) \geq \tau$).

Given a 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 a predictive error at most ε smaller than that of $f^\star$, with probability at least $1-\delta$. Like PAC learning [56], 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). *Given a hypothesis class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X} \quad [p]}$, a fairness metric $\quad$, 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 $\varepsilon, \nu \in (0,1)$.*

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 (2). The main difficulty is that, unlike Program (2), 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 A.27 in Supplementary Material A.5). 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 (2), $f^\star \in \mathcal{F}$, is feasible for our framework. To ensure this, we require Assumption 1 that also appeared in [14].

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

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 $\mathcal{E}^0$. We expect λ to be a non-vanishing positive constant in applications. For example, if q is statistical rate, the minority protected group makes at least 20% of the population (i.e., $\min_{\ell \in [p]} \Pr_D[Z = \ell] \geq 0.2$), 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 [14]). 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 $\quad$ and corresponding events $\mathcal{E}$ and $\mathcal{E}^0$ (as in Definition 3.1), a perturbation rate $\eta \in [0,1]$, and constants $\lambda, \quad \in (0,1]$, we define the error-tolerant program for perturbed samples $\widehat{S}$, whose empirical distribution is $\widehat{D}$, as*

$$\min_{f \in \mathcal{F}} \quad \mathrm{Err}_{\widehat{D}}(f), \tag{ErrTolerant} \quad (3)$$

$$\text{s.t.,} \quad (f, \widehat{S}) \geq \tau \cdot \left(\frac{1 - (\eta + \quad)/\lambda}{1 + (\eta + \quad)/\lambda} \right)^2, \tag{4}$$

$$\forall\, \ell \in [p],\ \Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell] \geq \lambda - \eta - \quad . \tag{5}$$

acts as a relaxation parameter in Program (ErrTolerant), which can be fixed in terms of the other parameters; see Theorem 4.3. Equation (4) ensures all feasible classifiers satisfy fairness constraints with respect to the perturbed samples $\widehat{S}$. Equation (5) 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 (4) and (5) 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 $\mathcal{F}_A := \{\{a \in A \mid f(a) = 1\} \mid f \in \mathcal{F}\}$. 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 of an optimal solution f_{ET} of Program (ErrTolerant) for any hypothesis class $\mathcal{F}$ with a finite VC dimension using $O(\mathsf{VC}(\mathcal{F}))$ 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 $\nu > 8\eta\tau/(\lambda \quad 2\eta)$, and confidence parameters $\delta \in (0, 1)$ with probability at least $1 - \delta$, the optimal solution $f_{\mathsf{ET}} \in \mathcal{F}$ of Program (ErrTolerant) with parameters η, λ, and $\quad := O\left(\min\{\varepsilon - 2\eta, \nu - 8\eta\tau/(\lambda \quad 2\eta), \lambda - 2\eta\}\right)$, and $N = \mathrm{poly}(d, 1/\quad, \log(p/\delta))$ perturbed samples from the η-Hamming model satisfies $\mathsf{Err}_D(f_{\mathsf{ET}}) - \mathsf{Err}_D(f^\star) \le \varepsilon$ and $\quad_D(f_{\mathsf{ET}}) \ge \tau - \nu$.*

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 $\nu = O((\eta\,\tau)/\lambda)$. 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 B.2 in Supplementary Material B.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 A.1.5. Finally, Program (ErrTolerant) only requires an estimate of one parameter, λ. (Since η is known, τ is fixed by the user, and can be set in terms of the other parameters.) If for each $\ell \in [p]$, we also have estimates of $\lambda_\ell := \mathsf{Pr}_D[\mathcal{E}(f^\star), \mathcal{E}^0(f^\star), Z = \ell]$ and $\gamma_\ell := \mathsf{Pr}_D[\mathcal{E}^0(f^\star), Z = \ell]$, then we can use this information to "tighten" Program (ErrTolerant) to the following program:

$$\begin{aligned} \min_{f \in F} \quad & \mathsf{Err}_{\widehat{D}}(f), \qquad \text{(ErrTolerant+)} \quad (6) \\ \text{s.t.,} \quad & (f, \widehat{S}) \ge \tau \cdot s, \\ & \forall\, \ell \in [p],\ \mathsf{Pr}_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell] \ge \lambda_\ell - \eta - \quad . \end{aligned}$$

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

$$\min_{\eta_1, \eta_2, \ldots, \eta_p \quad 0} \ \min_{\ell, k \in [p]} \frac{1 - \eta_\ell/\lambda_\ell}{1 + (\eta_k - \eta_\ell)/\gamma_\ell} \cdot \frac{1 + (\eta_\ell - \eta_k)/\gamma_k}{1 + \eta_\ell/\lambda_k}, \quad \text{s.t.,} \quad \sum_{\ell \in [p]} \eta_\ell \le \eta + \quad . \qquad (7)$$

If the classifiers in $\mathcal{F}$ do not use the protected attributes for prediction, then we can show that Program (ErrTolerant+) has a fairness guarantee of $(1 - s) + 4\eta\tau/(\lambda \quad 2\eta)$ (which is always smaller than $8\eta\tau/(\lambda \quad 2\eta)$) and an accuracy guarantee of 2η. We prove this result in Supplementary Material B.2. Thus, in applications where one can estimate $\lambda_1, \ldots, \lambda_p$ and $\gamma_1, \ldots, \gamma_p$, Program (ErrTolerant+) offers better fairness guarantee than Program (ErrTolerant) (up to constants).

The proof of Theorem 4.3 appears in Supplementary Material A.1.

As for computing f_{ET}, note that in general, Program (ErrTolerant) is a nonconvex optimization problem. In our simulations, we use the standard solver SLSQP in SciPy [57] to heuristically find f_{ET}; see Supplementary Material E.1. Theoretically, for any arbitrarily small $\alpha > 0$, the techniques from [13] can be used to find an $f \in \mathcal{F}$ that has the optimal objective value for Program (ErrTolerant) and that additively violates its fairness constraint (4) by at most α by solving a set of $O(1/(\lambda\alpha))$ convex programs; details appear in Supplementary Material C.

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 $\mathcal{F} \subseteq \{0, 1\}^{X \quad [p]}$ 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% 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 [11, 39], and holds for common hypothesis classes such as decision-trees and SVMs (Remark A.28 in Supplementary Material A.5). 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 A.2.

Theorem 4.5 (**Fairness guarantee of Theorem 4.3 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 fairness guarantee $\nu < \eta/(20\lambda) - O(\eta^2/\lambda^2)$, must have a poor error bound, of at least $\varepsilon \geq 1/4 - 2\eta/5 \geq 1/20$, to (ε, ν)-learn any $\mathcal{F}$ that satisfies a mild assumption. When η/λ is small, this shows that any learner with a fairness guarantee $\nu = o(\eta/\lambda)$ must have an error guarantee at least $1/4 - 2\eta/5 \gg 2\eta$. Thus, Theorem 4.5 shows that one cannot improve the fairness guarantee in Theorem 4.3 by more than a constant amount without deteriorating the error guarantee from 2η to $1/4 - 2\eta/5$. Like Theorem 4.4, the condition on $\mathcal{F}$ in Theorem 4.5 avoids "simple" hypothesis classes and holds for common hypothesis (Remark A.28 in Supplementary Material A.5). Finally, complementing our accuracy guarantee, we prove that for any $\varepsilon < \eta$, no algorithm can (ε, ν)-learn any hypothesis classes $\mathcal{F}$ satisfying mild assumptions (Theorem A.21 in Supplementary Material A.4); its proof appears in Supplementary Material A.4. Thus, the accuracy guarantee in Theorem 4.5 is optimal up to a factor of 2. The proof of Theorem 4.5 appears in Supplementary Material A.3.

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 bound on the accuracy of f_{ET}. This step relies on Lemma 4.6.

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

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

The proof of Lemma 4.6 follows from generalization bounds for bounded functions (e.g., see [54]) and because the η-Hamming model perturbs at most $\eta \cdot N$ samples. Let g be the 0-1 loss (i.e., $g(\widetilde{y}, y, z) := \mathtt{I}[\widetilde{y} \neq y]$), 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: $|\mathsf{Err}_D(f) - \mathsf{Err}(f, \widehat{S})| \leq \quad + \eta$. Thus, intuitively, minimizing $\mathsf{Err}(f, \widehat{S})$ could be a good strategy to minimize $\mathsf{Err}_D(f)$. 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

$$\mathsf{Err}_D(f_{\mathsf{ET}}) \leq \mathsf{Err}(f_{\mathsf{ET}}, \widehat{S}) + \eta + \quad \leq \mathsf{Err}(f^\star, \widehat{S}) + \eta + \quad \leq \mathsf{Err}_D(f^\star) + 2(\eta + \quad). \tag{8}$$

Step 2: Lower bound on 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 $f \in \mathcal{F}$: $|\ \ _D(f) - \ \ (f, \widehat{S})| \leq O(\eta/\lambda)$. Then ensuring that f has a high fairness on $\widehat{S}$ implies that it also has high fairness on S (up to an $O(\eta/\lambda)$ factor). However, such a bound does not hold for any $\mathcal{F}$ satisfying mild assumptions (see Example A.27). The first idea is to prove a similar (in fact, stronger multiplicative) bound on a specifically chosen *subset* of $\mathcal{F}$ (consisting of "stable" classifiers). Toward this, we define:

Definition 4.7. *A classifier $f \in \mathcal{F}$ is said to be s-stable for fairness metric , if for all adversaries $A \in \mathcal{A}(\eta)$ and confidence parameters $\delta \in (0,1)$, given $\mathrm{polylog}(1/\delta)$ samples $S \sim \mathcal{D}$, with probability at least $1 - \delta$, it holds that $\ \ _D(f)/\ \ (f,\widehat{S}) \in [s, 1/s]$, where $\widehat{S} := A(S)$.*

If an s-stable classifier f has fairness τ on $\widehat{S}$, then it has a fairness at least $\tau \cdot s$ on $\mathcal{D}$ with high probability. 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, $(\cdot, \widehat{S}) \geq \tau/s$, must have a fairness at least τ on $\mathcal{D}$ with high probability. The key idea is coming up such constraints.

Lemma 4.8. *Any classifier $f \in \mathcal{F}$ that satisfies $\min_{\ell \in [p]} \Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell] \geq \lambda + \eta + \ $, is $\left(\frac{1 - (\eta + \ )/\lambda}{1 + (\eta + \ )/\lambda}\right)^2$-stable for fairness metric (defined by events $\mathcal{E}$ and $\mathcal{E}^0$).*

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

Lemma 4.9. *If the following conditions hold then, $\mathsf{Err}_D(f_{\mathsf{ET}}) \quad \mathsf{Err}_D(f^\star) \quad 2\eta$ and $\ \ _D(f_{\mathsf{ET}}) \quad \tau \quad O(\eta/\lambda)$: (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 (C3) satisfy $(f, \widehat{S}) \geq \tau \cdot (1 - O(\eta/\lambda))$.*

Thus, it suffices to find error-tolerant constraints that satisfy conditions (C1) to (C3). Condition (C3) can be satisfied by adding the constraint $(\cdot, \widehat{S}) \geq \tau^0$, for $\tau^0 = \tau \cdot (1 - O(\eta/\lambda))$. From Lemma 4.8, condition (C2) follows by using the constraint in $\min_{\ell 2[p]} \Pr_D [\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell] \geq \lambda^0$, for $\lambda^0 \geq (\lambda)$. It remains to pick τ^0 and λ^0 such that condition (C1) also holds. The tension in setting τ^0 and λ^0 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 $\tau^0 := \tau \cdot (\frac{1 \ (\eta+ \ )/\lambda}{1+(\eta+ \ )/\lambda})^2$ and $\lambda^0 := \lambda - \eta -$ 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 [39, Theorem 1] and [11, 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, under some 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 these required guarantees. 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 with high probability $\widehat{S}$ is distributed according to iid samples from $\mathcal{D}_k$. Thus $\mathcal{L}$, who only sees $\widehat{S}$, with high probability, 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 the fairness and accuracy requirements are not vacuously satisfied.) Thus, for every $\mathcal{L}$ there is a distribution in $\mathcal{D}_1, \ldots, \mathcal{D}_m$ for which $\mathcal{L}$ outputs a bad classifier. (Note that even if the learner is randomized, it must fail with probability at least $1/m$.) Finally, the assumptions on $\mathcal{F}$ ensure that condition (C2) is satisfiable. For instance, if $\mathcal{F}$ has less than m hypothesis, then condition (C2) cannot be satisfied.

The key idea in the proofs is to come up with distributions satisfying the above conditions. [39, 11] 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. Our constructions are specific to the statistical rate fairness metric. However, one can still apply the general approach outlined above to other fairness metrics by constructing a suitable set of distributions. Full details appear in Supplementary Materials A.2 and A.3.

5 Empirical results

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

Metrics and baselines. The selection of an appropriate fairness metric is context-dependent and beyond the scope of this work [55]; 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 statistical rate under stochastic perturbations: **LMZV** [44] and **CHKV** [14]. **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** [40], which controls for true-positive rate (TPR) in the presence of a Malicious adversary, and **AKM** [6] that is the post-processing method of [32] and controls for equalized-odds fairness constraints. We also compare against the optimal unconstrained classifier, **Uncons;** this is the same as [11]'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 as the fairness metric with additional baselines and using the Adult data [23] in Supplementary Material E.

Implementation details. We use a randomly generated 70-30 train (S) test (T) split of the data, and generate the perturbed data $\widehat{S}$ 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 data T. **Err-Tol** is given the perturbation rate η and uses the SLSQP solver in SciPy [57] to solve Program (ErrTolerant+). 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 pertur-

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

	acc (η=0%)	SR (η=0%)	acc (η=3%)	SR (η=3%)	acc (η=5%)	SR (η=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)

bation rates: for each $\ell \in [p]$, $\eta_\ell := \mathsf{Pr}_D[\widehat{Z} \neq Z \mid Z = \ell]$, and **KL** is given η and for each $\ell \in [p]$, the probability $\mathsf{Pr}_D[Z = \ell, Y = 1]$; 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 $\gamma_\ell = \lambda_\ell := \mathsf{Pr}_{\widehat{D}}[Z = \ell]$, 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 statistical rate 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 statistical rate. The adversary's hope is that choosing samples far from the decision boundary would (falsely) give the appearance of a high statistical rate on $\widehat{S}$. This would make a fair classification framework output 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 poorer performance.

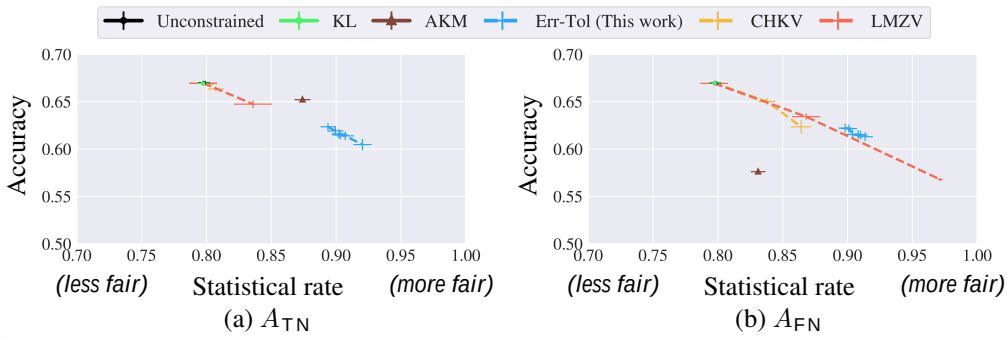
Simulation on synthetic data. We first show empirically that perturbations by the η-Hamming adversary can be prohibitively disruptive for methods that attempt to correct for stochastic noise. We consider synthetic data with 1,000 samples from two equally-sized protected groups; each sample has a binary protected attribute, two continuous features $x_1, x_2 \in \mathsf{R}$, 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.[4]

Results. The fairness and statistical rate averaged over 50 iterations are reported in Table 1 as a function of the perturbation η. At $\eta = 0$, both **CHKV** and **Err-Tol** nearly-satisfy the fairness constraint ($\mathsf{SR} \geq 0.79$) and have a near-perfect accuracy ($\mathsf{acc} \geq 0.98$). 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 ($\mathsf{acc} \geq 0.99$ and $\mathsf{SR} \geq 0.79$) 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 adversary. The COMPAS data in [9] contains 6,172 samples with 10 binary features and a label that is 1 if the individual did not recidivate and 0 otherwise; the statistical rate of **Uncons** on COMPAS is 0.78. We take gender (coded as binary) as the protected attribute, and set the fairness constraint on the statistical rate 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 statistical rate less than 0.78 on $\widehat{S}$.

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 are reported in Figure 1. For both adversaries, **Err-Tol** attains a better statistical rate than the unconstrained classifier (**Uncons**) for a small trade-off in accuracy. For adversary A_{TN} (Figure 1(a)), **Uncons** has statistical rate (0.80) and accuracy (0.67). In contrast,

[4] We also attempted to compare against **AKM**, **KL**, and **LMZV**. But they did not converge to $f^\star$ even on the unperturbed synthetic data, and hence, we did not include these results as it would be an unfair comparison.



(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.

Err-Tol achieves high statistical rate (0.92) with a trade-off in accuracy (0.60). In comparison, **AKM** has a higher accuracy (0.65) but a lower statistical rate (0.87), and other baselines have an even lower statistical rate (≤ 0.84) with accuracy comparable to **AKM.** For adversary A_{FN} (Figure 1(b)), **Uncons** has statistical rate (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 statistical rate (0.83) and accuracy (0.58). **LMZV** achieves the highest statistical rate (0.97) with a natural reduction in accuracy to (0.57). In this case, **Err-Tol** has similar accuracy to statistical rate trade-off as **LMZV,** but achieves a lower maximum statistical rate (0.91). Meanwhile, **Err-Tol** has a significantly higher statistical rate trade-off than **CHKV** at the same accuracy. We further evaluate our framework under stochastic perturbations in Supplementary Material E (specifically, against the perturbation model of [14]) and observe similar statistical rate and accuracy trade-offs as approaches [14, 44] tailored for stochastic perturbations.

Remark 5.1 (**Range of fairness parameters in the simulation**). *Among baselines,* **AKM, KL,** *and* **Uncons** *do not take the desired-fairness value as input, so they appear as points in Figure 1. For all other methods (***CHKV, Err-Tol,** *and* **LMZV***), we vary the fairness parameters starting from the tightest constraints (i.e., $\tau = 1$ and $\delta_L = 0$) and relax the constraints until all algorithms' achieved statistical rate matches the achieved statistical rate of the unconstrained classifier (this happens around $\tau = 0.7$ and $\delta_L = 0.1$). We do not relax the fairness parameters further because the resulting problem is equivalent to the unconstrained classification problem. (This is because the unconstrained classifier, which has the highest accuracy, satisfies the fairness constraints for $\tau \leq 0.7$ and $\delta_L \geq 0.1$).*

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 the class of linear-fractional fairness metrics. 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. That said, in a few applications, e.g., the randomized response procedure [60], where the perturbations are independent across samples and identically distributed according to a *known* distribution, frameworks for fair classification with stochastic perturbations can perform better. Further, like existing frameworks, our framework's 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 to avoid negative social implications in a misguided attempt to do good [45].

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.

Acknowledgements. This research was supported in part by an NSF CAREER Award (IIS-2045951), a J.P. Morgan Faculty Award, and an AWS MLRA Award.

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. COMPAS recidivism risk score data and analysis, 2016.

[4] Peter Auer. *Learning with Malicious Noise*, pages 1086–1089. Springer New York, New York, NY, 2016.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? In *FORC*, volume 156 of *LIPIcs*, pages 3:1–3:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

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

[12] 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.

[13] 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.

[14] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Fair classification with noisy protected attributes. In *ICML*, volume 120 of *Proceedings of Machine Learning Research*. PMLR, 2021.

[15] 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.

[16] 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.

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

[18] 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.

[19] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *AIES*, pages 91–98. ACM, 2019.

[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] Dua Dheeru and E Karra Taniskidou. UCI machine learning repository, 2017. `http://archive.ics.uci.edu/ml`.

[24] 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.

[25] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268. ACM, ACM, 2015.

[26] 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.

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

[28] 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.

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

[30] Maya R. Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *CoRR*, abs/1806.11212, 2018.

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

[32] 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.

[33] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1934–1943. PMLR, 2018.

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

[35] 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.

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

[37] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication, 2009. IC4 2009.*, pages 1–6. IEEE, 2009.

[38] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, Oct 2012.

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

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

[41] D. Kraft. *A software package for sequential quadratic programming*. Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt Köln: Forschungsbericht. Wiss. Berichtswesen d. DFVLR, 1988.

[42] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*. International World Wide Web Conferences Steering Committee, 2018.

[43] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In *NeurIPS*, 2020.

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

[45] 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.

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

[47] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *KDD 2011*, pages 502–510. ACM, 2011.

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

[49] 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.

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

[51] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5684–5693, 2017.

[52] 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.

[53] 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.

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

[55] 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.

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

[57] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[58] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *FAccT*, pages 526–536. ACM, 2021.

[59] 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.

[60] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965. PMID: 12261830.

[61] Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *COLT 2017*, pages 1920–1953, 2017.

[62] 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.

[63] 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.

[64] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–325–III–333. JMLR.org, 2013.

[65] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.

[66] 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.

Contents

A Proofs

A.1 Proof of Theorem 4.3

Recall that we assume Assumption 1 holds with constant $\lambda > 0$ and the VC dimension of $\mathcal{F}$ is finite, say $d \in \mathbb{N}$. Our goal is to prove that for all perturbation rates $\eta \in (0, \lambda/2)$, fairness thresholds $\tau \in (0, 1]$, bounds on error $\varepsilon > 2\eta$ and bounds on constraint violation $\nu > 8\eta\tau/(\lambda - 2\eta)$, and confidence parameters $\delta \in (0, 1)$, given sufficiently many perturbed samples from an η-Hamming adversary, with probability at least $1 - \delta$, it holds that

$$\mathrm{Err}_D(f_{\mathrm{ET}}) - \mathrm{Err}_D(f^\star) \leq \varepsilon,$$
$$_D(f_{\mathrm{ET}}) \geq \tau - \nu,$$

where f_{ET} is an optimal solution of Program (ErrTolerant). In the proof, we set the following parameters

$$_0 := \min\left\{2\eta - \varepsilon, \frac{8\eta\tau}{\lambda - 2\eta} - \nu\right\} \quad \text{and} \quad \delta_0 := \frac{\delta}{2p+1}. \tag{9}$$

We set $:= _0 \cdot \frac{(\lambda - 2\eta)^2}{32\ 3}$ in Program (ErrTolerant) and require at least N samples (perturbed by an η-Hamming adversary), where N satisfies

$$N = \left(\frac{1}{^2 \cdot (\lambda - 2\eta)^4} \cdot \left(d \log\left(\frac{d}{^2 \cdot (\lambda - 2\eta)^4}\right) + \log\left(\frac{p}{\delta_0}\right)\right)\right). \tag{10}$$

Note that these values satisfy the requirements $= O(\min\{\varepsilon - 2\eta, \nu - 8\eta\tau/(\lambda - 2\eta), \lambda - 2\eta\})$ and $N = \mathrm{poly}(d, 1/\ , \log(p/\delta))$.

Remark A.1. *All probabilities and expectations in all proofs are with respect to the draw of (X, Y, Z). Given a distribution $\mathcal{P}$, we use $\Pr_P[\cdot]$ to denote $\Pr_{(X,Y,Z)\ P}[\cdot]$ and $\mathrm{E}_P[\cdot]$ to denote $\mathrm{E}_{(X,Y,Z)\ P}[\cdot]$. If $\mathcal{P}$ denotes the distribution of perturbed samples, then we use $\Pr_P[\cdot]$ to denote $\Pr_{(X,Y,\widehat{Z})\ P}[\cdot]$ and $\mathrm{E}_P[\cdot]$ to denote $\mathrm{E}_{(X,Y,\widehat{Z})\ P}[\cdot]$; The difference between the two will be clear from context.*

A.1.1 Preliminaries: Generalization bound

We use Lemma A.2 in the proof of Theorem 4.3. See [54, Section 28.1] for a proof.

Lemma A.2 (Concentration of mean of bounded functions). *For any bounded function $g\colon \{0,1\} \times \{0,1\} \times [p] \to [0,1]$ and constants $, \delta_0 \in (0,1)$, given $N \geq \left(^2 \cdot \left(\mathrm{VC}(\mathcal{F}) \cdot \log\left(^{\mathrm{VC}(F)}/\ \right) + \log\left(1/\delta_0\right)\right)\right)$ samples S iid from $\mathcal{D}$, with probability at least $1 - \delta_0$, it holds that*

$$\forall\, f \in \mathcal{F}, \quad \left|\mathrm{E}_D\left[g(f(X,Z),Y,Z)\right] - \mathrm{E}_D\left[g(f(X,Z),Y,Z)\right]\right| \leq \quad ,$$

where D is the empirical distribution of S.

A.1.2 Step 1: Lower bound on the accuracy of f_{ET}

This step relies on Lemma A.3, which is the formal version of Lemma 4.6. We use Lemma A.3 in the proof of Lemma A.4 to lower bound the accuracy of f_{ET}.

Lemma A.3 (Bound on difference in means of bounded functions on $\mathcal{D}$ and on $\widehat{S}$). *For any bounded function $g\colon \{0,1\} \times \{0,1\} \times [p] \to [0,1]$ and constants $, \delta_0 \in (0,1)$, given $N \geq \left(^2 \cdot \left(\mathrm{VC}(\mathcal{F}) \cdot \log\left(^{\mathrm{VC}(F)}/\ \right) + \log\left(1/\delta_0\right)\right)\right)$ samples S iid from $\mathcal{D}$, and corresponding perturbed samples $A(S) := \{(x_i, y_i, \widehat{z}_i)\}_{i2[N]}$, with probability at least $1 - \delta$, it holds that*

$$\forall\, f \in \mathcal{F}, \quad \left|\mathrm{E}_{\widehat{D}}\left[g(f(X,\widehat{Z}),Y,\widehat{Z})\right] - \mathrm{E}_D\left[g(f(X,Z),Y,Z)\right]\right| \leq \quad + \eta,$$

where $\widehat{D}$ is the empirical distribution of $\widehat{S}$.

Proof. Let $S := \{(x_i, z_i, y_i)\}_{i2[N]}$ and $\widehat{S} := \{(x_i, \widehat{z}_i, y_i)\}_{i2[N]}$. Using the triangle inequality for absolute value, we have

$$\begin{aligned}
&\left|\mathrm{E}_{\widehat{D}}\left[g(f(X,\widehat{Z}),Y,\widehat{Z})\right] - \mathrm{E}_D\left[g(f(X,Z),Y,Z)\right]\right| \\
&\qquad \leq \left|\mathrm{E}_D\left[g(f(X,Z),Y,Z)\right] - \mathrm{E}_D\left[g(f(X,Z),Y,Z)\right]\right| \\
&\qquad + \left|\mathrm{E}_{\widehat{D}}\left[g(f(X,\widehat{Z}),Y,\widehat{Z})\right] - \mathrm{E}_D\left[g(f(X,Z),Y,Z)\right]\right|.
\end{aligned} \tag{11}$$

We can upper bound the first term in the RHS using Lemma A.2. In particular, we have that with probability at least $1-\delta$, for all $f \in \mathcal{F}$, it holds that

$$\left|\mathsf{E}_D\left[g(f(X,Z),Y,Z)\right] - \mathsf{E}_D\left[g(f(X,Z),Y,Z)\right]\right| \leq \quad . \tag{12}$$

Further, we can upper bound the second term in the RHS of Equation (11) for all $f \in \mathcal{F}$ as follows

$$\begin{aligned}
&\left|\mathsf{E}_{\widehat{D}}\left[g(f(X,\widehat{Z}),Y,\widehat{Z})\right] - \mathsf{E}_{\widehat{D}}\left[g(f(X,\widehat{Z}),Y,\widehat{Z})\right]\right| \\
&= \quad \frac{1}{N}\cdot\left|\sum_{i \in [N]} g(f(x_i,\widehat{z}_i),y_i,\widehat{z}_i) - g(f(x_i,z_i),y_i,z_i)\right| \\
&= \quad \frac{1}{N}\cdot\left|\sum_{i \in [N]:\ z_i \neq \widehat{z}_i} g(f(x_i,\widehat{z}_i),y_i,\widehat{z}_i) - g(f(x_i,z_i),y_i,z_i)\right| \\
&\qquad \text{(For all } i \in [N] \text{, where } z_i = \widehat{z}_i,\ g(f(x_i,\widehat{z}_i),y_i,\widehat{z}_i) = g(f(x_i,z_i),y_i,z_i)\text{.)} \\
&\leq \quad \frac{1}{N}\cdot\left|\sum_{i \in [N]:\ z_i \neq \widehat{z}_i} 1\right| \qquad \text{(Using that } g \text{ is bounded by 0 and 1)} \\
&\leq \quad \eta. \qquad (13)
\end{aligned}$$

Since with probability at least $1-\delta$, both Equation (12) and (13) hold for all $f \in \mathcal{F}$, substituting them in Equation (11) gives us the required bound. □

Lemma A.4. *If $f^\star$ is feasible for Program (ErrTolerant), then it holds that:*

$$\mathsf{Err}_D(f_{\mathsf{ET}}) - \mathsf{Err}_D(f^\star) \leq 2\eta + \quad .$$

Proof. Let g be the 0-1 loss (i.e., $g(\widetilde{y},y,z) := \mathtt{I}[\widetilde{y} \neq y]$), then for all $f \in \mathcal{F}$, Lemma A.3 shows that the error of f on samples drawn from $\mathcal{D}$ and samples in $\widehat{S}$ are close: $\left|\mathsf{Err}_D(f) - \mathsf{Err}(f,\widehat{S})\right| \leq \quad + \eta$. 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 A.3 we get that

$$\begin{aligned}
\mathsf{Err}_D(f_{\mathsf{ET}}) &\overset{\text{Lemma A.3}}{\leq} \mathsf{Err}(f_{\mathsf{ET}},\widehat{S}) + \eta + \\
&\leq \mathsf{Err}(f^\star,\widehat{S}) + \eta + \\
&\overset{\text{Lemma A.3}}{\leq} \mathsf{Err}_D(f^\star) + 2(\eta + \quad).
\end{aligned}$$

□

A.1.3 Step 2: Lower bound on the fairness of f_{ET}

In this step, we show that any $f \in \mathcal{F}$ feasible for Program (ErrTolerant) satisfies $\quad_D(f) \geq \tau - \nu$ (Lemma A.8). This relies on the notion of s-stability (Definition A.5) and the fact that any $f \in \mathcal{F}$ feasible for Program (ErrTolerant) is s-stable (for s a function of η and λ); Corollary A.7.

Definition A.5 (**s-stability**). *Given a constant $s \in (0,1)$ and perturbation rate $\eta \in [0,1)$, a classifier $f \in \mathcal{F}$ is said to be s-stable for fairness metric $\quad$ with perturbation rate η, if for all adversaries $A \in \mathcal{A}(\eta)$ and confidence parameters $\delta_0 \in (0,1)$ given $N = \mathrm{polylog}(1/\delta_0)$ samples S iid from $\mathcal{D}$ and corresponding perturbed samples $\widehat{S} := A(S)$, with probability at least $1-\delta_0$ (over draw of $S \sim \mathcal{D}$), it holds that*

$$\frac{\quad_D(f)}{(f,\widehat{S})} \in \left[s, \frac{1}{s}\right].$$

If an s-stable classifier f has fairness τ on $\widehat{S}$, then it has a fairness at least $\tau \cdot s$ on $\mathcal{D}$ with high probability. Thus, if we have some constraint C such that any feasible $f \in \mathcal{F}$ satisfying constraint C is s-stable, then any classifier satisfying constraint C and the fairness constraint, $\quad(\cdot,\widehat{S}) \geq \tau/s$, must have a fairness at least τ on $\mathcal{D}$ with high probability. The key idea is coming up such a constraints. First, in Lemma A.6, we give such constraints which use the unperturbed protected attributes Z, later in Corollary A.7 we give constraints which only use the perturbed protected attributes $\widehat{Z}$.

Lemma A.6 (**Sufficient condition for a stable classifier**). *For each $\alpha \in (0,1)$, any classifier $f \in \mathcal{F}$ satisfying*

$$\min_{\ell \in [p]} \Pr_D\left[\mathcal{E}(f), \mathcal{E}^0(f), Z = \ell\right] \geq \alpha, \tag{14}$$

is $\left(\frac{1 \quad (\eta + \quad)/\alpha}{1+(\eta + \quad)/\alpha}\right)^2$-stable with respect to the fairness metric $\quad$ defined by events $\mathcal{E}$ and $\mathcal{E}^0$.

Proof of Lemma A.6. Let $\widehat{D}$ be the empirical distribution of $\widehat{S}$. Let J be the event that for all $\ell \in [p]$ and all $f \in \mathcal{F}$

$$\left|\Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell] - \Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z_i = \ell]\right| < \eta + \quad , \tag{15}$$

$$\left|\Pr_{\widehat{D}}[\mathcal{E}^0(f), \widehat{Z} = \ell] - \Pr_D[\mathcal{E}^0(f), Z_i = \ell]\right| < \eta + \quad . \tag{16}$$

Using Lemma A.3 with $g(\widetilde{y}, y, z) := \mathtt{I}[\mathcal{E}(\widetilde{y}), \mathcal{E}^0(\widetilde{y}), z = \ell]$, we get that with probability at least $1 - \delta_0$, Equation (15) holds for $f \in \mathcal{F}$ and a particular $\ell \in [p]$. Similarly, using Lemma A.3 with $g(\widetilde{y}, y, z) := \mathtt{I}[\mathcal{E}(\widetilde{y}), z = \ell]$, we get that with probability at least $1 - \delta_0$, Equation (16) holds for $f \in \mathcal{F}$ and a particular $\ell \in [p]$. Applying the union bound over all $\ell \in [p]$, we get that

$$\Pr[J] \geq 1 - 2p\delta_0. \tag{17}$$

Suppose event J holds. Then, for any $\ell, k \in [p]$, we have

$$\begin{aligned}
&\frac{\Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell]}{\Pr_{\widehat{D}}[\mathcal{E}^0(f), \widehat{Z} = \ell]} \cdot \frac{\Pr_{\widehat{D}}[\mathcal{E}^0(f), \widehat{Z} = k]}{\Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = k]} \\
&\overset{(15),(16)}{\geq} \quad \frac{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = \ell] - \eta -}{\Pr_D[\mathcal{E}^0(f), Z = \ell] + \eta +} \cdot \frac{\Pr_D[\mathcal{E}^0(f), Z = k] - \eta -}{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = k] + \eta +}.
\end{aligned} \tag{18}$$

Using Equation (14), we can lower bound the RHS of Equation (18) as follows

$$\begin{aligned}
&\frac{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = \ell] - \eta -}{\Pr_D[\mathcal{E}^0(f), Z = \ell] + \eta +} \cdot \frac{\Pr_D[\mathcal{E}^0(f), Z = k] - \eta -}{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = k] + \eta +} \\
&\overset{(14)}{\geq} \quad \frac{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = \ell]}{\Pr_D[\mathcal{E}^0(f), Z = \ell]} \cdot \frac{\Pr_D[\mathcal{E}^0(f), Z = k]}{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = k]} \cdot \left(\frac{1 - \frac{1}{\alpha} \cdot (\eta + \quad)}{1 + \frac{1}{\alpha} \cdot (\eta + \quad)}\right)^2.
\end{aligned} \tag{19}$$

Therefore, combining Equations (18) and (19), we have that conditioned on event J

$$\begin{aligned}
(f, \widehat{S}) \quad &= \quad \min_{\ell, k 2 [p]} \frac{\Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell]}{\Pr_{\widehat{D}}[\mathcal{E}^0(f), \widehat{Z} = \ell]} \cdot \frac{\Pr_{\widehat{D}}[\mathcal{E}^0(f), \widehat{Z} = k]}{\Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = k]} \\
&\qquad \text{(Using the definition of the fairness metric)} \\
&\overset{(18),(19)}{\geq} \left(\frac{1 - \frac{1}{\alpha} \cdot (\eta + \quad)}{1 + \frac{1}{\alpha} \cdot (\eta + \quad)}\right)^2 \cdot \min_{\ell, k 2 [p]} \frac{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = \ell]}{\Pr_D[\mathcal{E}^0(f), Z = \ell]} \cdot \frac{\Pr_D[\mathcal{E}^0(f), Z = k]}{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = k]} \\
&= \quad {}_D(f) \cdot \left(\frac{1 - \frac{1}{\alpha} \cdot (\eta + \quad)}{1 + \frac{1}{\alpha} \cdot (\eta + \quad)}\right)^2. \qquad \text{(Using the definition of the fairness metric)}
\end{aligned}$$

This completes the proof of the upper bound in Definition A.5. It remains to prove the lower bound in Definition A.5 The proof of the lower bound in Definition A.5 is analogous to the proof of the upper bound. For any $\ell, k \in [p]$, we have

$$\begin{aligned}
&\frac{\Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell]}{\Pr_{\widehat{D}}[\mathcal{E}^0(f), \widehat{Z} = \ell]} \cdot \frac{\Pr_{\widehat{D}}[\mathcal{E}^0(f), \widehat{Z} = k]}{\Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = k]} \\
&\overset{(15),(16)}{\leq} \quad \frac{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = \ell] + \eta +}{\Pr_D[\mathcal{E}^0(f), Z = \ell] - \eta -} \cdot \frac{\Pr_D[\mathcal{E}^0(f), Z = k] + \eta +}{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = k] - \eta -}.
\end{aligned} \tag{20}$$

Using Equation (14), we can upper bound the RHS of Equation (20) as follows

$$\begin{aligned}
&\frac{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = \ell] - \eta -}{\Pr_D[\mathcal{E}^0(f), Z = \ell] + \eta +} \cdot \frac{\Pr_D[\mathcal{E}^0(f), Z = k] - \eta -}{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = k] + \eta +} \\
&\overset{(14)}{\leq} \quad \frac{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = \ell]}{\Pr_D[\mathcal{E}^0(f), Z = \ell]} \cdot \frac{\Pr_D[\mathcal{E}^0(f), Z = k]}{\Pr_D[\mathcal{E}(f), \mathcal{E}^0(f), Z = k]} \cdot \left(\frac{1 + \frac{1}{\alpha} \cdot (\eta + \quad)}{1 - \frac{1}{\alpha} \cdot (\eta + \quad)}\right)^2.
\end{aligned} \tag{21}$$

Therefore, combining Equations (18) and (19), we have that conditioned on event J

$$
\begin{aligned}
{}_D(f) \quad &= \quad \min_{\ell,k2[p]} \frac{\mathsf{Pr}_D[\mathcal{E}(f),\mathcal{E}^0(f),Z=\ell]}{\mathsf{Pr}_D[\mathcal{E}^0(f),Z=\ell]} \cdot \frac{\mathsf{Pr}_D[\mathcal{E}^0(f),Z=k]}{\mathsf{Pr}_D[\mathcal{E}(f),\mathcal{E}^0(f),Z=k]} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Using the definition of the fairness metric)} \\
&\overset{(18),(19)}{\geq} \left(\frac{1-\frac{1}{\alpha}\cdot(\eta+\quad)}{1+\frac{1}{\alpha}\cdot(\eta+\quad)}\right)^2 \cdot \min_{\ell,k2[p]} \frac{\mathsf{Pr}_{\widehat{D}}[\mathcal{E}(f),\mathcal{E}^0(f),\widehat{Z}=\ell]}{\mathsf{Pr}_{\widehat{D}}[\mathcal{E}^0(f),\widehat{Z}=\ell]} \frac{\mathsf{Pr}_{\widehat{D}}[\mathcal{E}^0(f),\widehat{Z}=k]}{\mathsf{Pr}_{\widehat{D}}[\mathcal{E}(f),\mathcal{E}^0(f),\widehat{Z}=k]} \\
&= \quad {}_D(f,\widehat{S}) \cdot \left(\frac{1-\frac{1}{\alpha}\cdot(\eta+\quad)}{1+\frac{1}{\alpha}\cdot(\eta+\quad)}\right)^2. \qquad \text{(Using the definition of the fairness metric)}
\end{aligned}
$$

□

Corollary A.7. *For all $\alpha \in (0,1)$, any classifier $f \in \mathcal{F}$ satisfying*

$$\min_{\ell 2[p]} \mathsf{Pr}_D\left[\mathcal{E}(f),\mathcal{E}^0(f),\widehat{Z}=\ell\right] \geq \alpha, \tag{22}$$

is $(\frac{1\ \ (\eta+\ \ )/(\alpha\ \ \eta\ \ )}{1+(\eta+\ \ )/(\alpha\ \ \eta\ \ )})^2$-stable with respect to the fairness metric defined by events $\mathcal{E}$ and $\mathcal{E}^0$.

Proof. Suppose event J defined in Lemma A.6 occurs. Then by the first condition of J, Equation (15), we have that $\left|\mathsf{Pr}_D\left[\mathcal{E}(f),\mathcal{E}^0(f),Z=\ell\right] - \mathsf{Pr}_{\widehat{D}}\left[\mathcal{E}(f),\mathcal{E}^0(f),\widehat{Z}=\ell\right]\right| \leq \eta+\quad$. Combining this with Equation (22), we get that $\mathsf{Pr}_{\widehat{D}}\left[\mathcal{E}(f),\mathcal{E}^0(f),Z=\ell\right] \geq \alpha-\eta-\quad$. Then repeating the proof of Lemma A.6 we get that, conditioned on J, the following holds

$$\left(\frac{1+(\eta+\quad)/(\alpha-\eta-\quad)}{1-(\eta+\quad)/(\alpha-\eta-\quad)}\right)^2 \leq \frac{{}_D(f)}{(f,\widehat{S})} \leq \left(\frac{1-(\eta+\quad)/(\alpha-\eta-\quad)}{1+(\eta+\quad)/(\alpha-\eta-\quad)}\right)^2$$

Since J occurs with probability at least $1-2\delta_0 > 1-\delta$, we get that f is $(\frac{1\ \ (\eta+\ \ )/(\alpha\ \ \eta\ \ )}{1+(\eta+\ \ )/(\alpha\ \ \eta\ \ )})^2$-stable. □

Setting α as $\lambda+\eta+\quad$, we recover Lemma 4.8 from Corollary A.7. Now, we are ready to prove the main result in this step: Lemma A.8.

Lemma A.8 (Any feasible solution of Program (ErrTolerant) is approximately fair). *For all $,\delta_0 \in (0,1)$, given $N \geq \left(\quad{}^2 \cdot \left(\mathsf{VC}(\mathcal{F}) \cdot \log\left(\mathsf{VC}(F)/\ \ \right) + \log\left(1/\delta_0\right)\right)$ iid samples S from $\mathcal{D}$, and corresponding perturbed samples $A(S) := \{(x_i,y_i,\widehat{z}_i)\}_{i2[N]}$, with probability at least $1-2p\delta_0$, any $f \in \mathcal{F}$ feasible for Program (ErrTolerant) satisfies*

$${}_D(f) \geq \tau - \frac{8\eta\tau}{\lambda-2\eta} - \quad 0.$$

Proof. Consider any classifier $f \in \mathcal{F}$ that is feasible for Program (ErrTolerant). f must satisfy

$$(f,\widehat{S}) \geq \tau \cdot \left(\frac{1-(\eta+\quad)/\lambda}{1+(\eta+\quad)/\lambda}\right)^2, \tag{23}$$

$$\forall\, \ell \in [p], \quad \mathsf{Pr}_{\widehat{D}}\left[\mathcal{E}(f),\mathcal{E}^0(f),\widehat{Z}=\ell\right] \geq \lambda-\eta-\quad. \tag{24}$$

Since f satisfies Equation (24), using Lemma 4.8, we get that f is $(\frac{1\ \ (\eta+\ \ )/(\lambda\ \ 2\eta\ \ 2\ \ )}{1+(\eta+\ \ )/(\lambda\ \ 2\eta\ \ 2\ \ )})^2$-stable. Thus, with high probability, it holds that

$$
\begin{aligned}
{}_D(f) \quad &\geq \quad (f,\widehat{S}) \cdot \left(\frac{1-(\eta+\quad)/(\lambda-2\eta-2\quad)}{1+(\eta+\quad)/(\lambda-2\eta-2\quad)}\right)^2 \qquad (25) \\
&\overset{(23)}{\geq} \quad \tau \cdot \left(\frac{1-(\eta+\quad)/\lambda}{1+(\eta+\quad)/\lambda}\right)^2 \cdot \left(\frac{1-(\eta+\quad)/(\lambda-2\eta-2\quad)}{1+(\eta+\quad)/(\lambda-2\eta-2\quad)}\right)^2.
\end{aligned}
$$

We can lower bound the RHS of the above equation using Fact A.9.

Fact A.9. *For all $\alpha \in (0,1]$ and $\eta,\quad \in [0,1]$, $\left(\frac{1\ \ \frac{1}{\alpha}(\eta+\ \ )}{1+\frac{1}{\alpha}(\eta+\ \ )}\right)^2 \geq \left(1-\frac{4}{\alpha}\cdot(\eta+\quad)\right)$.*

Using Fact A.9 twice we get that

$$
\begin{aligned}
_D(f) &\geq \tau \cdot \left(1 - \frac{4(\eta + \quad)}{\lambda}\right) \cdot \left(1 - \frac{4(\eta + \quad)}{\lambda - 2\eta - 2}\right) \\
&> \tau \cdot \left(1 - \frac{4(\eta + \quad)}{\lambda} - \frac{4(\eta + \quad)}{\lambda - 2\eta - 2}\right) \quad \text{(Using that } \lambda > 2\eta + 2 \quad \text{and } \lambda, \eta, \quad > 0) \\
&\geq \tau \cdot \left(1 - \frac{8(\eta + \quad)}{\lambda - 2\eta - 2}\right) \quad \text{(Using that } \lambda > 2\eta + 2 \quad \text{and } \lambda, \eta, \quad > 0) \\
&= \tau \cdot \left(1 - \frac{8(\eta + \quad)}{\lambda - 2\eta} \cdot \frac{1}{1 - \frac{2}{(\lambda \quad 2\eta)} \cdot}\right) \\
&\geq \tau \cdot \left(1 - \frac{8(\eta + \quad)}{\lambda - 2\eta} \cdot \left(1 + \frac{4}{(\lambda - 2\eta)} \cdot \quad\right)\right) \quad \text{(Using that } \frac{2}{(\lambda \quad 2\eta)} \cdot \quad \in [0, \tfrac{1}{2}]) \\
&= \tau \cdot \left(1 - \frac{8\eta}{\lambda - 2\eta} - \frac{8}{\lambda - 2\eta} - \frac{32\eta}{(\lambda - 2\eta)^2} - \frac{32 \quad ^2}{(\lambda - 2\eta)^2}\right) \\
&\overset{(9)}{=} \tau \cdot \left(1 - \frac{8\eta}{\lambda - 2\eta} - \frac{0}{3} - \frac{0}{3} - \frac{0}{3}\right).
\end{aligned}
$$

Finally, from the discussion in the proof of Lemma A.6, it follows that for our choice of N, the above equation holds with probability at least $1 - 2p\delta_0$. □

A.1.4 Step 3: $f^\star$ is feasible for Program (ErrTolerant) with high probability

In this step, we conclude the proof of Theorem 4.3. It remains show that $f^\star$ is feasible for Program (ErrTolerant): The fairness guarantee follows from Lemma A.8, and if $f^\star$ is feasible for Program (ErrTolerant), then the accuracy guarantee follows from Lemma A.4.

Lemma A.10 (Structure of the optimal fair classifier). *If Assumption 1 holds, then For all* $\quad, \delta_0 \in (0,1)$*, given* $N \geq \quad \left(\quad ^2 \cdot \left(\mathsf{VC}(\mathcal{F}) \cdot \log\left(\mathsf{VC}(F)/ \quad\right) + \log\left(1/\delta_0\right)\right)\right.$ *iid samples* S *from* $\mathcal{D}$*, and corresponding perturbed samples* $A(S) := \{(x_i, y_i, \widehat{z}_i)\}_{i2[N]}$*, with probability at least* $1 - \delta_0$*,* $f^\star$ *is feasible for Program (ErrTolerant).*

Proof. Let G be the event that for all $\ell \in [p]$ and $f \in \mathcal{F}$

$$
\left|\mathsf{Err}_{\widehat{D}}(f) - \mathsf{Err}_D(f)\right| < \eta + \quad . \tag{26}
$$

Using Lemma A.3 with $g(\widetilde{y}, y, z) := \mathtt{I}[\widetilde{y} \neq y]$, shows that with probability at least $1 - \delta_0$ Inequality (26) holds for all $f \in \mathcal{F}$. Thus,

$$
\Pr[G] \geq 1 - \delta_0. \tag{27}
$$

Suppose events G and J hold. To show that $f^\star$ is feasible for Program (ErrTolerant), we have to show that

$$
(f^\star, \widehat{S}) \geq \tau \cdot \left(1 - \frac{4\eta}{\lambda} - \quad _0\right), \tag{28}
$$

$$
\forall\, \ell \in [p], \quad \Pr_{\widehat{D}}[\mathcal{E}(f^\star), \mathcal{E}^0(f^\star), \widehat{Z} = \ell] \geq \lambda - \eta - \quad _0. \tag{29}
$$

Since event J and Assumption 1 hold, we can apply Lemma A.6 to get that $\quad(f^\star, \widehat{S}) \geq \quad _D(f^\star) \cdot (1 - \frac{4\eta}{\lambda} - \quad _0)$. Then, Equation (28) holds since $\quad _D(f^\star) \geq \tau$. Finally, Equation (29) follows by using Assumption 1 and Equation (15) in the definition of event J. □

Proof of Theorem 4.3. Let J be the event that the fairness guarantee in Lemma A.8 holds and G be the event that $f^\star$ is feasible for Program (ErrTolerant). Using the union bound J and G, we get that

$$
\begin{aligned}
\Pr_D[J \ \wedge G] &\geq 1 - (2p + 1) \cdot \delta_0 \qquad \text{(Using Lemma A.8 and Lemma A.10)} \\
&\overset{(9)}{\geq} 1 - \delta.
\end{aligned}
$$

Suppose events J and G occur. By Lemma A.8, we know that the following holds $_D(f) \geq \tau - {}^{8\eta\tau}/_{(\lambda \ 2\eta)} - \ $. Since $\leq \ _0 \leq {}^{8\eta}/_{(\lambda \ 2\eta)} - {}^{\nu}/_{\tau}$, it follows that

$$_D(f) \geq \tau - \nu.$$

By Lemma A.10, we know that $f^\star$ is feasible for Program (ErrTolerant). Then, from Lemma A.4 it follows that $\mathsf{Err}_D(f_{\mathsf{ET}}) - \mathsf{Err}_D(f^\star) \leq 2\eta + \ $. Since $\leq \varepsilon - 2\eta$, it follows that

$$\mathsf{Err}_D(f_{\mathsf{ET}}) - \mathsf{Err}_D(f^\star) \leq \varepsilon.$$

□

A.1.5 Generalization of Theorem 4.3 to the Nasty Sample Noise model

In this section, we generalize Theorem 4.3 to the Nasty sample noise model. Precisely, we show that if Assumption 1 holds with constant $\lambda > 0$ and $\mathcal{F}$ has VC dimension $d \in \mathsf{N}$, 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 $\delta \in (0, 1)$, given sufficiently many perturbed samples from the η-Hamming model, with probability at least $1 - \delta$, it holds that

$$\mathsf{Err}_D(f_{\mathsf{ET}}) - \mathsf{Err}_D(f^\star) \leq \varepsilon \text{ and } \ _D(f_{\mathsf{ET}}) \geq \tau - \nu,$$

where f_{ET} is an optimal solution of Program (ErrTolerant).

The proof of the above generalization is almost identical to the proof of Theorem 4.3. Instead of repeating the entire proof, we highlight the changes required in the proof of Theorem 4.3.

The generalization requires two changes: (1) Proving an analogue of Lemma A.3 for the Nasty Sample Noise model, (2) generalizing the definition of s-stability (Definition A.5) to the Nasty Sample Noise model. These changes are sufficient because all other arguments use the guarantees of Lemma A.3 or s-stability, without using any properties of the η-Hamming model. Updating Definition A.5 only requires changing the perturbation model from η-Hamming to η-Nasty Sample Noise; we omit the formal statement. Next, we prove Lemma A.11 which is the required analogue of Lemma A.3.

Lemma A.11 (Bound on difference in means of bounded functions on $\mathcal{D}$ and on $\widehat{S}$). *For any bounded function $\ell\colon \{0,1\} \times \{0,1\} \times [p] \to [0,1]$, constants $, \delta_0 \in (0,1)$, and adversaries A admissible under the η-Nasty Sample Noise model, given $N \geq \big(\ ^2 \cdot \big(\mathsf{VC}(\mathcal{F}) \cdot \log\big({}^{\mathsf{VC}(F)}/ \ \big) + \log\big({}^1/_{\delta_0}\big)\big)$ samples S iid from $\mathcal{D}$, and corresponding perturbed samples $A(S) := \{(\widehat{x}_i, \widehat{y}_i, \widehat{z}_i)\}_{i2[N]}$, with probability at least $1 - \delta_0$, it holds that*

$$\forall\, f \in \mathcal{F}, \quad \left| \underset{(\widehat{X},\widehat{Y},\widehat{Z}) \ \ \widehat{D}}{\mathsf{E}} g(f(\widehat{X},\widehat{Z}),\widehat{Y},\widehat{Z}) - \mathsf{E}_{(X,Y,Z) \ \ D}\left[g(f(X,Z),Y,Z)\right] \right| \leq \ + \eta,$$

where $\widehat{D}$ is the empirical distribution of $A(S)$.

Proof. Let $S := \{(x_i, z_i, y_i)\}_{i2[N]}$. Using the triangle inequality for absolute value, we have

$$\begin{aligned}
&\left|\mathsf{E}_{(\widehat{X},\widehat{Y},\widehat{Z}) \ \ \widehat{D}}[g(f(\widehat{X},\widehat{Z}),\widehat{Y},\widehat{Z})] - \mathsf{E}_{(X,Y,Z) \ \ D}[g(f(X,Z),Y,Z)]\right| \\
&\qquad \leq \left|\mathsf{E}_{(X,Y,Z) \ \ D}[g(f(X,Z),Y,Z)] - \mathsf{E}_{(X,Y,Z) \ \ D}[g(f(X,Z),Y,Z)]\right| \\
&\qquad + \left|\mathsf{E}_{(\widehat{X},\widehat{Y},\widehat{Z}) \ \ \widehat{D}}[g(f(\widehat{X},\widehat{Z}),\widehat{Y},\widehat{Z})] - \mathsf{E}_{(X,Y,Z) \ \ D}[g(f(X,Z),Y,Z)]\right|. \qquad (30)
\end{aligned}$$

We can upper bound the first term in the RHS using Lemma A.2; this is identical to the argument under the η-Hamming model. In particular, we have that with probability at least $1 - \delta$, for all $f \in \mathcal{F}$, it holds that

$$\left|\mathsf{E}_{(X,Y,Z) \ \ D}[g(f(X,Z),Y,Z)] - \mathsf{E}_{(X,Y,Z) \ \ D}\left[g(f(X,Z),Y,Z)\right]\right| \leq \ . \qquad (31)$$

(The proof of the upper bound on the second term in the RHS of Equation (30) is slightly different from the proof under the η-Hamming model.) For all $f \in \mathcal{F}$, it holds that

$$\begin{aligned}
&\left|\mathsf{E}_{(\widehat{X},\widehat{Y},\widehat{Z})\ \ \widehat{D}}[g(f(\widehat{X},\widehat{Z}),\widehat{Y},\widehat{Z})] - \mathsf{E}_{(X,Y,Z)\ \ D}[g(f(X,Z),Y,Z)]\right| \\
&= \frac{1}{N}\left|\sum_{i \in [N]} g(f(\widehat{x}_i,\widehat{z}_i),\widehat{y}_i,\widehat{z}_i) - g(f(x_i,z_i),y_i,z_i)\right| \\
&= \frac{1}{N}\left|\sum_{i \in [N]:\ (x_i,y_i,z_i) \neq (\widehat{x}_i,\widehat{y}_i,\widehat{z}_i)} g(f(\widehat{x}_i,\widehat{z}_i),\widehat{y}_i,\widehat{z}_i) - g(f(x_i,z_i),y_i,z_i)\right| \\
&\text{(For all } i \in [N], \text{ where } (x_i,y_i,z_i) = (\widehat{x}_i,\widehat{y}_i,\widehat{z}_i),\ g(f(\widehat{x}_i,\widehat{z}_i),\widehat{y}_i,\widehat{z}_i) = g(f(x_i,z_i),y_i,z_i).) \\
&\leq \frac{1}{N}\left|\sum_{i \in [N]:\ (x_i,y_i,z_i) \neq (\widehat{x}_i,\widehat{y}_i,\widehat{z}_i)} 1\right| \qquad \text{(Using that } g \text{ is bounded by 0 and 1)} \\
&\leq \eta. \qquad (32)
\end{aligned}$$

Since with probability at least $1-\delta$, both Equations (31) and (32) hold for all $f \in \mathcal{F}$, substituting them in Equation (30) gives us the required bound. □

One can substitute Lemma A.3 by Lemma A.11 and repeat the proof of Theorem 4.3 for the η-Nasty Sample Noise model.

A.2 Proof of Theorem 4.4

Theorem 4.4 assumes that $\mathcal{F}$ shatters the set $\{x_A, x_B, x_C\} \times [2] \subseteq \mathcal{X} \times [p]$, the fairness threshold $\tau \in (1/2, 1)$, and statistical rate is the fairness metric; recall that the statistical rate of a classifier $f \in \mathcal{F}$ is

$$_D(f) := \frac{\min_{\ell \in [p]} \Pr_D[f = 1 \mid Z = \ell]}{\max_{\ell \in [p]} \Pr_D[f = 1 \mid Z = \ell]}.$$

Then, given parameters $\tau \in (1/2, 1)$ and $\delta \in [0, 1/2)$, our goal is to show that for any

$$\varepsilon \in \left[0, \frac{1}{2}\right) \text{ and } \nu \in \left[0, \tau - \frac{1}{2}\right),$$

$\mathcal{F}$ is not (ε, ν)-learnable with perturbation rate η and confidence δ. We prove a more general result: Given parameters $\tau \in [0,1)$, $c \in (0, \min\{\tau, 1/2\})$, and $\delta \in [0, 1/2)$, we show that for any

$$\varepsilon \in [0, c) \text{ and } \nu \in [0, \tau - c),$$

$\mathcal{F}$ is not (ε, ν)-learnable with perturbation rate η and confidence δ. When $\tau > 1/2$, the original result follows by taking the limit as c approaches $1/2$.

A.2.1 Proof of Theorem 4.4

Proof of Theorem 4.4 (assuming Lemmas A.12 to A.14). Let $\mathcal{L}$ be any learner. Set

$$\alpha := \min\left\{\frac{\eta}{2}, 1-\tau, \tau - c - \nu, c - \varepsilon\right\}, \qquad (33)$$

and confidence parameter $\delta \in (0, 1/2)$ (in Definition 3.2). We construct three distributions $\mathcal{D}_1, \mathcal{D}_2$, and $\mathcal{D}_3$, parameterized by α, that satisfy Lemma A.12. (See Table 2 for details of the distributions).

Lemma A.12 (Adversary can hide the true distribution). *For all $\delta_0 \in (0,1)$, $\alpha \in (0, \eta)$, and $\ell, k \in [3]$, given $N \geq 3 \cdot \ln(1/\delta_0) \cdot (\eta - \alpha)^{-2}$ iid samples, S, from $\mathcal{D}_\ell$ there is an adversary $A \in \mathcal{A}(\eta)$ such that with probability $1 - \delta_0$ (over the draw of S) the perturbed samples $\widehat{S} := A(S)$ are distributed as iid draws from $\mathcal{D}_k$.*

Suppose $\mathcal{L}$ is given N samples, where

$$N \geq 3 \cdot \ln 2 \cdot (\eta - \alpha)^{-2}. \qquad (34)$$

Consider three cases, where in the k-th case ($k \in [3]$) the samples in S are iid from $\mathcal{D}_k$. By Lemma A.12, in each case there is an $A \in \mathcal{A}(\eta)$, such that with probability at least $1/2$, $\widehat{S} := A(S)$ have the distribution $\mathcal{D}_1$.

Thus, given $\widehat{S}$, with probability at least $1/2$, $\mathcal{L}$ cannot identify the distribution from which S was drawn. As a result, with probability at least $1/2$, $\mathcal{L}$ outputs the same classifier, say $f_{\mathsf{Com}} \in \mathcal{F}$, in all three cases. We show that no $f_{\mathsf{Com}} \in \mathcal{F}$ satisfies the accuracy and fairness guarantee in all three cases.

Lemma A.13 (**No good classifier for all cases**). *There is no classifier $f \in \mathcal{F}$ such that for all $k \in [3]$, $\mathsf{Err}_{D_k}(f) < c \cdot (1-\alpha)$ and $\quad_{D_k}(f) \geq c + \alpha$.*

Lemma A.14 (**A good classifier for each case**). *For each $k \in [3]$, there is an $h_k \in \mathcal{F}$ such that $\mathsf{Err}_{D_k}(h_k) < c\alpha/2$ and $\quad_{D_k}(h_k) > 1-\alpha$.*

Note that for each $k \in [3]$, $f_k^\star$ satisfies the fairness constraint because $\tau < 1-\alpha$ (Equation (33)). Thus, the optimal fair classifier $f_k^\star$ for $\mathcal{D}_k$ subject to having a statistical rate τ must satisfy

$$\mathsf{Err}_{D_k}(f_k^\star) < \frac{c\alpha}{2}. \tag{35}$$

(Otherwise, we have a contradiction as h_k satisfies the fairness constraints and has a smaller error than $f_k^\star$.) If $\mathcal{L}$ is a (ε,ν)-learner, then in the k-th case, with probability at least $1-\delta > 1/2$, $\mathcal{L}$ must output a classifier f_k which satisfies

$$\mathsf{Err}_{D_k}(f_k) - \mathsf{Err}_{D_k}(f_k^\star) \leq \varepsilon \text{ and } \tau - \quad_{D_k}(f_k) \leq \nu.$$

But in all cases with probability at least $1/2$, $\mathcal{L}$ outputs f_{Com}. Because $1/2 > \delta$, f_{Com} must satisfy:

$$\text{For all } k \in [3], \quad \mathsf{Err}_{D_k}(f_{\mathsf{Com}}) - \mathsf{Err}_{D_k}(f_k^\star) \leq \varepsilon \text{ and } \tau - \quad_{D_k}(f_{\mathsf{Com}}) \leq \nu. \tag{36}$$

But from Lemma A.13 we know that for each $k \in [3]$ either

$$\mathsf{Err}_{D_k}(f_{\mathsf{Com}}) \geq c \cdot (1-\alpha) \quad \text{or} \quad \quad_{D_k}(f_{\mathsf{Com}}) < c + \alpha. \tag{37}$$

(Case A) $\mathsf{Err}_{D_k}(f_{\mathsf{Com}}) \geq c \cdot (1-\alpha)$**:** In this case, from Equation (36), we have

$$\begin{aligned} \varepsilon &\geq c \cdot (1-\alpha) - \mathsf{Err}_{D_k}(f_k^\star) \\ &\overset{(35)}{>} c \cdot (1-\alpha) - \frac{c\alpha}{2} \\ &> c - \alpha \qquad \text{(Using that } c < 1/2 \text{ and } \alpha > 0\text{)} \end{aligned} \tag{38}$$

But because $\alpha \leq c - \varepsilon$, Equation (38) cannot hold.

(Case B) $\quad_{D_k}(f_{\mathsf{Com}}) < c + \alpha$**:** In this case, from Equation (36), we have

$$\nu > \tau - c - \alpha. \tag{39}$$

But because $\alpha \leq \tau - c - \nu$, Equation (39) does not hold.

Therefore, we have a contradiction. Hence, $\mathcal{L}$ is not an (ε,ν)-learner for $\mathcal{F}$. Since the choice of $\mathcal{L}$ was arbitrary, we have shown that there is no learner which (ε,ν)-learns $\mathcal{F}$. It remains to prove Lemmas A.12 to A.14. □

A.2.2 Proof of Lemma A.12

Set $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ to be unique distributions with marginal distribution specified in Table 2, such that, for any draw $(X,Y,Z) \sim \mathcal{D}_k$ ($k \in [3]$) Y takes the value $\mathtt{I}[X = x_A]$, i.e.,

$$Y = \begin{cases} 1 & \text{if } X = x_A, \\ 0 & \text{otherwise.} \end{cases} \tag{40}$$

In particular, the construction of $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ ensures that the total variation distance between any pair of distributions is less than α. In this section, we prove the following generalization of Lemma A.12.

Proposition A.15 (**Adversary can hide the true distribution**). *For all $\delta,\eta \in (0,1)$, $\alpha \in (0,\eta)$, and two distributions $\mathcal{P}$ and $\mathcal{Q}$ over $\mathcal{X} \times \{0,1\} \times [p]$ that satisfy the following conditions:*

(C1) $\mathsf{TV}(\mathcal{P},\mathcal{Q}) = \alpha$,

(a) D_1: $\Pr_{D_1}[(X,Z) = (r,s)]$ for $(r,s) \in \{x_A, x_B, x_C\}$ [2].

	x_A	x_B	x_C
1	$c(1-\alpha)$	$(1-c)(1-\alpha)$	$\alpha/2$
2	$c\alpha/2$	$\alpha(1-c)/2$	0

(b) D_2: $\Pr_{D_2}[(X,Z) = (r,s)]$ for $(r,s) \in \{x_A, x_B, x_C\}$ [2].

	x_A	x_B	x_C
1	$c(1-\alpha)$	$(1-c)(1-\alpha/2)$	$c\alpha/2$
2	$c\alpha/2$	0	$\alpha(1-c)/2$

(c) D_3: $\Pr_{D_3}[(X,Z) = (r,s)]$ for $(r,s) \in \{x_A, x_B, x_C\}$ [2].

	x_A	x_B	x_C
1	$c(1-\alpha/2)$	$(1-c)(1-\alpha)$	$\alpha(1-c)/2$
2	0	$\alpha(1-c)/2$	$c\alpha/2$

Table 2: Marginal distributions of D_1, D_2, D_3 over X [2]. Recall that for a sample (X,Y,Z) D_k ($k \in [3]$), Y takes the value $\mathbb{I}[X = x_A]$.

(C2) *$\mathcal{P}$ and $\mathcal{Q}$ have the same marginal on $\mathcal{X}$, i.e., for all $T \subseteq \mathcal{X}$,*

$$\Pr_{(X,Y,Z) \sim P}[X \in T] = \Pr_{(X,Y,Z) \sim Q}[X \in T],$$

(C3) *for a random sample (X,Y,Z) drawn from $\mathcal{P}$, the label Y is independent of Z conditioned on X, i.e., $Y \perp Z \mid X$, similarly for a random sample (X,Y,Z) drawn from $\mathcal{Q}$, the label Y is independent of Z conditioned on X.*

Then, there is an adversary $A \in \mathcal{A}(\eta)$, that given N iid samples S from $\mathcal{P}$, where

$$N \geq 3 \cdot \ln(1/\delta) \cdot (\eta - \alpha)^{-2},$$

outputs a perturbed samples $\widehat{S} := A(S)$ such that with probability $1-\delta$ (over the draw of S) samples in $\widehat{S}$ are distributed as iid draws from $\mathcal{Q}$.

Note that Lemma A.12 follows by substituting $\mathcal{P}$ and $\mathcal{Q}$ by $\mathcal{D}_\ell$ and $\mathcal{D}_k$ respectively.

Proof of Proposition A.15. Let $A \in \mathcal{A}(\eta)$ use the following algorithm:

1. **For** $r,s \in \mathcal{X}$ **do: Set** $p(r,s) := \min\left\{\frac{\Pr_{(X,Y,Z) \sim Q}[(X,Z)=(r,s)]}{\Pr_{(X,Y,Z) \sim P}[(X,Z)=(r,s)]}, 1\right\}$
   //Since A knows the distributions P and Q, it can compute $p(r,s)$
2. **For** $i \in [N]$ **do:**
   (a) **Sample** a point t_i uniformly at random from $[0,1]$
   (b) **If** $t_i \leq p(X_i, Z_i)$ **then: Set** $\widetilde{Z}_i := Z_i$,
   (c) **Otherwise: Set** $\widetilde{Z}_i := 3 - Z_i$ *//If $Z_i = 1$ set $\widetilde{Z}_i = 2$, if $Z_i = 2$ set $\widetilde{Z}_i = 1$*
3. **If** $\sum_{i \in [N]} \mathbb{I}[\widetilde{Z}_i \neq Z_i] < \delta \cdot N$ **then: return** $\{(X_i, \widetilde{Z}_i, Y_i)\}_{i \in [N]}$,
4. **Otherwise: return** $\{(X_i, Z_i, Y_i)\}_{i \in [N]}$

(Since A knows the distributions $\mathcal{P}$ and $\mathcal{Q}$, it can compute $p(r,s)$.)

Lemma A.16. *If $N \geq 3 \cdot \ln(1/\delta) \cdot (\eta - \alpha)^{-2}$, then with probability at least $1-\delta$, $\sum_{i \in [N]} \mathbb{I}[\widetilde{Z}_i \neq Z_i] < \eta \cdot N$.*

Proof. For all $i \in [N]$, let $C_i \in \{0,1\}$ be a random variable indicating if $\widetilde{Z}_i \neq Z_i$. Since for each $i \in [N]$ the sample (X_i, Y_i, Z_i) and the point t_i is drawn independently of others, it follows that the random variables C_i are independent of each other. Suppose we can show that $\Pr[C_i] \leq \alpha$. Then, by linearity of expectation, it follows that $\mathsf{E}[\sum_{i \in [N]} C_i] \leq \alpha \cdot N$. Thus, using the Chernoff bound, we get that $\Pr[\sum_{i \in [N]} C_i \leq \eta \cdot N] \geq 1 - \delta$. This completes the proof of Lemma A.16, up to proving $\Pr[C_i = 1] \leq \alpha$. Towards this, observe that

$$
\begin{aligned}
&\Pr[C_i = 1] \\
&= \sum_{r \in \mathcal{X}, s \in [p]} \Pr[(X_i, Z_i) = (r,s)] \cdot \Pr[C_i = 1 \mid (X_i, Z_i) = (r,s)] \\
&= \sum_{r \in \mathcal{X}, s \in [p]} \Pr[(X_i, Z_i) = (r,s)] \cdot \Pr[Z_i \neq \widetilde{Z}_i \mid (X_i, Z_i) = (r,s)] && \text{(Definition of } C_i\text{)} \\
&= \sum_{r \in \mathcal{X}, s \in [p]} \Pr[(X_i, Z_i) = (r,s)] \cdot (1 - p(r,s)) && \text{(Definition of } p(r,s)\text{)} \\
&= \sum_{r \in \mathcal{X}, s \in [p]} \Pr[(X_i, Z_i) = (r,s)] \cdot \max\left\{1 - \frac{\Pr_Q[(X,Z) = (r,s)]}{\Pr_P[(X,Z) = (r,s)]}, 0\right\} \\
& && \text{(Definition of } p(r,s)\text{)} \\
&= \sum_{r \in \mathcal{X}, s \in [p]} \Pr_P[(X,Z) = (r,s)] \cdot \max\left\{1 - \frac{\Pr_Q[(X,Z) = (r,s)]}{\Pr_P[(X,Z) = (r,s)]}, 0\right\} \\
& && \text{(Using that for each } i \in [N],\ (X_i, Y_i, Z_i) \sim \mathcal{P}\text{)} \\
&= \sum_{r \in \mathcal{X}, s \in [p]} \max\{\Pr_P[(X,Z) = (r,s)] - \Pr_Q[(X,Z) = (r,s)], 0\} \\
&= \mathsf{TV}(\mathcal{P}, \mathcal{Q}) \\
&= \alpha. && \text{(Using that } \mathsf{TV}(\mathcal{P}, \mathcal{Q}) = \alpha\text{)}
\end{aligned}
$$

□

Lemma A.17. *Each sample in $\widetilde{S} := \{(X_i, \widetilde{Z}_i, Y_i)\}_{i \in [N]}$ is independent of each other and is distributed according to $\mathcal{Q}$.*

Proof. Since for each $i \in [N]$ the sample (X_i, Y_i, Z_i) and the point t_i is drawn independently of others, it follows that the samples $(X_i, \widetilde{Z}_i, Y_i)$ are independent of each other.

To see that $(X_i, \widetilde{Z}_i, Y_i) \sim \mathcal{Q}$, fix any $i \in [N]$, $r \in \mathcal{X}$, and $s \in [2]$. It holds that

$$\Pr[X_i = r, \widetilde{Z}_i = s, Y_i = 1] \overset{(40)}{=} \Pr[Y_i = 1 \mid X_i = r] \cdot \Pr[X_i = r, \widetilde{Z}_i = s], \tag{41}$$

$$\Pr[X_i = r, \widetilde{Z}_i = s, Y_i = 0] \overset{(40)}{=} \Pr[Y_i = 0 \mid X_i = r] \cdot \Pr[X_i = r, \widetilde{Z}_i = s], \tag{42}$$

here we used the fact that Y_i is independent of Z_i (see Equation (40)). Suppose that

$$\Pr[X_i = r, \widetilde{Z}_i = s] = \Pr_Q[(X,Z) = (r,s)]. \tag{43}$$

Then, from Equation (40) we have that

$$\Pr_Q[X = r, \widetilde{Z} = s, Y = 1] \overset{(40)}{=} \Pr[Y_i = 1 \mid X_i = r] \cdot \Pr_Q[(X,Z) = (r,s)], \tag{44}$$

$$\Pr_Q[X = r, \widetilde{Z} = s, Y = 0] \overset{(40)}{=} \Pr[Y_i = 0 \mid X_i = r] \cdot \Pr_Q[(X,Z) = (r,s)]. \tag{45}$$

Further, combining Equations (41) and (42) and Equations (44) and (45), we get for all $y \in \{0,1\}$

$$\Pr[X_i = r, \widetilde{Z}_i = s, Y_i = y] = \Pr_Q[X = r, \widetilde{Z} = s, Y = y].$$

It remains to prove Equation (43). Before proving it, we recall the following invariant from the statement of this proposition: For all $r \in \mathcal{X}$, it holds that

$$\Pr_P[X = r] = \Pr_Q[X = r]. \tag{46}$$

Consider $\mathsf{Pr}[X_i = r, \widetilde{Z}_i = 1]$ for some $r \in \mathcal{X}$. From the algorithm used by the adversary, we have

$$\begin{aligned}\mathsf{Pr}[X_i = r, \widetilde{Z}_i = 1] &= (1 - p(a,2)) \cdot \mathsf{Pr}[(X_i, Z_i) = (r,2)] + p(a,1) \cdot \mathsf{Pr}[(X_i, Z) = (r,1)] \\ &= (1 - p(a,2)) \cdot \mathsf{Pr}_P[(X,Z) = (r,2)] + p(a,1) \cdot \mathsf{Pr}_P[(X,Z) = (r,1)]. \\ &\qquad \text{(For all } i \in [N], (X_i, Y_i, Z_i) \sim \mathcal{P}) \quad (47)\end{aligned}$$

We consider two cases.

(Case A) $\mathsf{Pr}_Q[(X,Z) = (a,1)] \geq \mathsf{Pr}_P[(X,Z) = (a,1)]$**:** In this case, we have $p(a,1) = 1$.

$$\begin{aligned}\mathsf{Pr}[X_i = r, \widetilde{Z}_i = 1] &\overset{(47)}{=} (1 - p(a,2)) \cdot \mathsf{Pr}_P[(X,Z) = (r,2)] + p(a,1) \cdot \mathsf{Pr}_P[(X,Z) = (r,1)] \\ &= \mathsf{Pr}_P[(X,Z) = (r,2)] + \mathsf{Pr}_P[(X,Z) = (r,1)] \cdot \left(1 - \tfrac{\mathsf{Pr}_Q[(X,Z)=(r,1)]}{\mathsf{Pr}_P[(X,Z)=(r,1)]}\right) \\ &\qquad \text{(Definition of } p(r,s)) \\ &= \mathsf{Pr}_P[X = r] - \mathsf{Pr}_Q[(X,Z) = (r,2)] \\ &\overset{(46)}{=} \mathsf{Pr}_Q[X = r] - \mathsf{Pr}_Q[(X,Z) = (r,2)] \\ &= \mathsf{Pr}_Q[(X,Z) = (r,1)]\end{aligned}$$

(Case B) $\mathsf{Pr}_Q[(X,Z) = (a,1)] < \mathsf{Pr}_P[(X,Z) = (a,1)]$**:** In this case, we have $p(a,1) < 1$.

$$\begin{aligned}\mathsf{Pr}[X_i = r, \widetilde{Z}_i = 1] &\overset{(47)}{=} (1 - p(a,2)) \cdot \mathsf{Pr}_P[(X,Z) = (r,2)] + p(a,1) \cdot \mathsf{Pr}_P[(X,Z) = (r,1)] \\ &= \mathsf{Pr}_P[(X,Z) = (r,1)] \cdot \frac{\mathsf{Pr}_Q[(X,Z) = (r,1)]}{\mathsf{Pr}_P[(X,Z) = (r,1)]} \qquad \text{(Definition of } p(r,s)) \\ &= \mathsf{Pr}_Q[(X,Z) = (r,1)].\end{aligned}$$

In both, cases, we have $\mathsf{Pr}[X_i = r, \widetilde{Z}_i = 1] = \mathsf{Pr}_Q[(X,Z) = (r,1)]$. By swapping the protected labels, we can show that $\mathsf{Pr}[X_i = r, \widetilde{Z}_i = 2] = \mathsf{Pr}_Q[(X,Z) = (r,2)]$. This proves Equation (43). □

From Lemma A.16, with probability at least $1 - \delta$, $\widehat{S} := \{(X_i, \widetilde{Z}_i, Y_i)\}_{i \in [N]}$. By Lemma A.17, the samples $\{(X_i, \widetilde{Z}_i, Y_i)\}_{i \in [N]}$ are iid from $\mathcal{Q}$. Thus, Proposition A.15 follows. □

A.2.3 Proof of Lemma A.13

Proof of Lemma A.13. Our goal is to show that for every $f \in \mathcal{F}$, there exists a choice $k \in [3]$, such that, f has error at least $c(1-\alpha)$ or statistical rate at most $c + \alpha$ with respect to $\mathcal{D}_k$. Since $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ are supported on subsets of $\{x_A, x_B, x_C\} \times [2]$, it suffices to consider the restriction of $\mathcal{F}$ on this domain. There at most 2^6 classifiers in this restriction. We partition them into three cases.

(Case A) $f(x_B, 1) = 1$**:** For any $k \in [3]$, we have

$$\begin{aligned}\mathsf{Err}_{D_k}(f) &= \sum_{r \in \mathcal{X}, s \in [p]} \mathsf{Pr}_{D_k}[f(X,Z) \neq Y \mid (X,Z) = (r,s)] \cdot \mathsf{Pr}[(X,Z) = (r,s)] \\ &\geq \mathsf{Pr}_{D_k}[f(X,Z) \neq Y \mid X = x_B, Z = 1] \cdot \mathsf{Pr}[X = x_B, Z = 1] \\ &\overset{(40)}{\geq} \mathsf{Pr}[X = x_B, Z = 1] \qquad \text{(Using that, in this case, } f(x_B,1) = 1) \\ &\overset{\text{Table 2}}{\geq} (1-c) \cdot (1-\alpha) \\ &> c(1-\alpha). \qquad \text{(Using that } c < 1/2)\end{aligned}$$

Thus, in Case A, f has an error larger than $c \cdot (1-\alpha)$ on each of $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$.

(Case B) $f(x_A, 1) = 0$ **and** $f(x_B, 1) = 0$**:** For any $k \in [3]$, we have

$$\begin{aligned}
\mathsf{Err}_{D_k}(f) &= \sum_{r \in X, s \in [p]} \Pr_{D_k}[f(X,Z) \neq Y \mid (X,Z) = (r,s)] \cdot \Pr[(X,Z) = (r,s)] \\
&\geq \Pr_{D_k}[f(X,Z) \neq Y \mid X = x_A, Z = 1] \cdot \Pr[X = x_A, Z = 1] \\
&\overset{(40)}{\geq} \Pr[X = x_A, Z = 1] \qquad \text{(Using that, in this case, } f(x_B, 1) = 1\text{)} \\
&\overset{\text{Table 2}}{\geq} c(1-\alpha).
\end{aligned}$$

Thus, in Case B, f has an error larger than $c \cdot (1-\alpha)$ on each of $\mathcal{D}_1, \mathcal{D}_2$, and $\mathcal{D}_3$.

(Case C) $f(x_A, 1) = 1$ **and** $f(x_B, 1) = 0$**:**

(Case C.1) $\sum_{r \in X} f(r, 2) \geq 2$**:** In this case, f takes a value of 1 on at least two points in the tuple

$$L = ((x_A, 2), (x_B, 2), (x_C, 2)).$$

If f takes a value of 0 on a point in L, then fix $j \in [3]$ such that $h(L_j) = 0$. Let $k := 4 - j$. Consider the distribution $\mathcal{D}_k$. Notice that by our construction $\Pr_{D_k}[L_i] = 0$ (see Table 2). Since L_i has measure 0 for $\mathcal{D}_k$, the value of f at this point does not affect its accuracy or statistical rate on $\mathcal{D}_k$. Thus, we can assume that $f(L_i) = 1$. Or in other words, we can assume that

$$\text{For all } r \in \mathcal{X}, \quad f(r, 2) = 1. \tag{48}$$

We compute the performance of f on both protected groups $\ell \in [p]$. For $Z = 2$, we have

$$\begin{aligned}
\mathsf{E}_{D_k}[f(X,Z) \mid Z = 2] &= \sum_{r \in X} f(r,2) \cdot \Pr_{D_k}[X = r \mid Z = 2] \\
&\overset{(48)}{=} \sum_{r \in X} 1 \cdot \Pr_{D_k}[X = r \mid Z = 2] \\
&\overset{\text{Table 2}}{=} 1.
\end{aligned} \tag{49}$$

For $Z = 1$, we have

$$\begin{aligned}
\mathsf{E}_{D_k}[f(X,Z) \mid Z = 1] &= \sum_{r \in X} f(r,1) \cdot \Pr_{D_k}[X = r \mid Z = 1] \\
&= 1 \cdot \Pr_{D_k}[X = x_A \mid Z = 1] + f(x_C, 1) \cdot \Pr_{D_k}[X = x_C \mid Z = 1] \\
&\qquad \text{(In this case, } f(x_A, 1) = 1 \text{ and } f(x_B, 1) = 0\text{)} \\
&\leq 1 \cdot \Pr_{D_k}[X = x_A \mid Z = 1] + \Pr_{D_k}[X = x_C \mid Z = 1] \\
&\qquad \text{(Using that } f(x_C, 1) \leq 1\text{)} \\
&\overset{\text{Table 2}}{\leq} \max\left\{ \tfrac{c\,(1\ \ \alpha) + \alpha/2}{1\ \ \alpha/2}, \tfrac{c\,(1\ \ \alpha) + c\alpha/2}{1\ \ \alpha/2}, \tfrac{c\,(1\ \ \alpha/2) + \alpha\,(1\ \ c)/2}{1\ \ \alpha/2} \right\} \\
&= \frac{c \cdot (1-\alpha) + \alpha/2}{1 - \alpha/2} \qquad \text{(Using } c, \alpha > 0\text{)} \\
&< c + \alpha. \qquad \text{(Using } c, \alpha > 0 \text{ and } \alpha \leq 1\text{)}
\end{aligned} \tag{50}$$

Since $c < 1/2$ and $\alpha \leq \min\{\eta, 1/2\}$, we have

$$c + \alpha < 1. \tag{51}$$

Now, we can compute the statistical rate of f using Equations (49), (50), and (51).

$$_{D_k}(f) = \frac{\min_{\ell \in [p]} \mathsf{E}_{D_k}[f(X,Z) \mid Z = \ell]}{\max_{\ell \in [p]} \mathsf{E}_{D_k}[f(X,Z) \mid Z = \ell]} \overset{(49),(50),(51)}{<} \frac{c + \alpha}{1}.$$

Thus, in Case C.1, f has a statistical rate smaller than $c + \alpha$ on distribution $\mathcal{D}_k$.

(Case C.2) $\sum_{r \in X} f(r, 2) \leq 1$**:** Thus, f takes a value of 0 on at least two points in the list

$$L = ((x_A, 2), (x_B, 2), (x_C, 2)).$$

If f takes a value of 1 on one of the points in L, then fix $j \in [3]$ such that $f(L_j) = 0$. Let $k := 4 - k$ Consider the distribution $\mathcal{D}_k$. Notice that by our construction $\Pr_{D_k}[L_j] = 0$ (see Table 2). Since L_i has measure 0 on $\mathcal{D}_k$, the value of f at this point does not affect its accuracy or statistical rate on $\mathcal{D}_k$. Thus, we can assume that $f(L_j) = 0$. Or in other words, we can assume that

$$\text{For all } r \in \mathcal{X}, \quad f(r, 2) = 0. \tag{52}$$

We would like to compute the statistical rate of f. Toward this, we first compute the performance of f on both protected groups. For $Z = 2$, we have

$$\begin{aligned} \mathrm{E}_{D_i}[f(X,Z) \mid Z = 2] &= \sum_{r \in X} f(r,2) \cdot \Pr[X = r \mid Z = 2] \\ &\overset{(52)}{=} 0. \end{aligned} \tag{53}$$

For $Z = 1$, we have

$$\begin{aligned} \mathrm{E}_{D_k}[f(X,Z) \mid Z = 1] &= \sum_{r \in X} f(r,1) \cdot \Pr_{D_k}[X = r \mid Z = 1] \\ &= 1 \cdot \Pr_{D_k}[X = x_A \mid Z = 1] + f(x_C, 1) \cdot \Pr_{D_k}[X = x_C \mid Z = 1] \\ &\qquad \text{(Using that, in this case, } f(x_A,1) = 1 \text{ and } f(x_B,1) = 0) \\ &\geq \Pr_{D_k}[X = x_A \mid Z = 1] \\ &\qquad \text{(Using } f(x_C,1) \cdot \Pr[X = x_C \mid Z = 1] \geq 0) \\ &\overset{\text{Table 2}}{\geq} \max\left\{ \frac{c \cdot (1-\alpha)}{1 - \alpha/2}, \frac{c(1-\alpha)}{1-\alpha/2}, \frac{c \cdot (1 - \alpha/2)}{1 - \alpha/2} \right\} \\ &= \frac{c \cdot (1-\alpha)}{1 - \alpha/2} \qquad \text{(Using that } \alpha, c > 0) \\ &> c. \qquad \text{(Using that } \alpha, c > 0) \end{aligned} \tag{54}$$

Now, we can compute the statistical rate of f using Equations (54) and (53).

$$_{D_i}(f) = \frac{\min_{\ell \in [p]} \mathrm{E}_{D_k}[f(X,Z) \mid Z = \ell]}{\max_{\ell \in [p]} \mathrm{E}_{D_k}[f(X,Z) \mid Z = \ell]} \overset{(54),(53),(c>0)}{\leq} 0$$

Thus, in Case C.2, f has a statistical rate 0 on the distribution $\mathcal{D}_k$.

Across all cases, we proved that all 2^6 classifiers in the restriction of $\mathcal{F}$ to $\{x_A, x_B, x_C\} \times [2]$, either have a error larger than $c \cdot (1 - \alpha)$ or statistical rate smaller than $c + \alpha$ on one of $\mathcal{D}_1$, $\mathcal{D}_2$, or $\mathcal{D}_3$. □

A.2.4 Proof of Lemma A.14

Proof of Lemma A.14. For each distribution $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$, we will give an classifier $f \in \mathcal{F}$ which satisfies the condition in Lemma A.14.

(Case A) $\mathcal{D}_1$**:** Define f as $f(x,z) := \mathbb{I}[x = x_A]$. Comparing this to Equation (40), we get that

$$\mathrm{Err}_{D_1}(f) = 0.$$

For $Z = 2$, we have

$$\begin{aligned} \mathrm{E}_{D_1}[f(X,Z) \mid Z = 2] &= \sum_{r \in X} f(r,2) \cdot \Pr_{D_1}[X = r \mid Z = 2] \\ &= \Pr_{D_1}[X = x_A \mid Z = 2] \\ &\overset{\text{Table 2}}{=} c. \end{aligned} \tag{55}$$

Similarly, for $Z = 1$, we have

$$\begin{aligned} \mathrm{E}_{D_1}[f(X,Z) \mid Z = 1] &= \sum_{r \in X} f(r,1) \cdot \Pr_{D_1}[X = r \mid Z = 1] \\ &= \Pr_{D_1}[X = x_A \mid Z = 1] \\ &\overset{\text{Table 2}}{=} \frac{c(1-\alpha)}{(1-\alpha/2)}. \end{aligned} \tag{56}$$

Thus, we have

$$\begin{aligned}
_{D_1}(h) &= \frac{\min_{\ell \in [p]} \mathsf{E}_{D_1}[f(X,Z) \mid Z = \ell]}{\max_{\ell \in [p]} \mathsf{E}_{D_1}[f(X,Z) \mid Z = \ell]} \\
&\overset{(55),(56)}{=} \frac{c(1-\alpha)}{(1-\alpha/2)} \cdot \frac{1}{c} && \text{(Using that } c, \alpha > 0\text{)} \\
&= 1 - \frac{\alpha}{(2-\alpha)} \\
&\geq 1 - \alpha. && \text{(Using that } \alpha \leq 1\text{)}
\end{aligned}$$

(Case B) $\mathcal{D}_2$**:** Define f as $f(x,z) := \mathtt{I}[x = x_A]$. Comparing this to Equation (40), we get that

$$\mathsf{Err}_{D_1}(f) = 0.$$

For $Z = 2$, we have

$$\begin{aligned}
\mathsf{E}_{D_2}[f(X,Z) \mid Z = 2] &= \sum_{r \in X} f(r,2) \cdot \mathsf{Pr}_{D_2}[X = r \mid Z = 2] \\
&= \mathsf{Pr}_{D_2}[X = x_A \mid Z = 2] \\
&\overset{\text{Table 2}}{=} c. && (57)
\end{aligned}$$

Similarly, for $Z = 1$, we have

$$\begin{aligned}
\mathsf{E}_{D_2}[f(X,Z) \mid Z = 1] &= \sum_{r \in X} f(r,1) \cdot \mathsf{Pr}_{D_2}[X = r \mid Z = 1] \\
&= \mathsf{Pr}_{D_2}[X = x_A \mid Z = 1] \\
&\overset{\text{Table 2}}{=} \frac{c \cdot (1-\alpha)}{(1-\alpha/2)}. && (58)
\end{aligned}$$

Thus, we have

$$\begin{aligned}
_{D_1}(h) &= \frac{\min_{\ell \in [p]} \mathsf{E}_{D_2}[f(X,Z) \mid Z = \ell]}{\max_{\ell \in [p]} \mathsf{E}_{D_2}[f(X,Z) \mid Z = \ell]} \\
&\overset{(57),(58)}{=} \frac{c \cdot (1-\alpha)}{(1-\alpha/2)} \cdot \frac{1}{c} && \text{(Using } c, \alpha > 0\text{)} \\
&\geq 1 - \alpha. && \text{(Using that } \alpha \leq 1\text{)}
\end{aligned}$$

(Case C) $\mathcal{D}_3$**:** Define $f \in \mathcal{F}$ to be the following classifier

$$f(x,z) := \begin{cases} 1 & \text{if } (x,z) \in \{(x_A,1),(x_C,2)\}, \\ 0 & \text{otherwise.} \end{cases} \tag{59}$$

Such an $f \in \mathcal{F}$ exists because $\mathcal{F}$ shatters the set $\{x_A, x_B, x_C\} \times [2]$. We have that

$$\begin{aligned}
\mathsf{Err}_{D_3}(f) &= \sum_{r \in X, s \in [p]} \mathsf{Pr}_{D_3}[f(r,s) \neq Y \mid (X,Z) = (r,s)] \cdot \mathsf{Pr}_{D_3}[(X,Z) = (r,s)] \\
&\overset{(40),(59)}{=} \mathsf{Pr}_{D_3}[(X,Z) = (x_A,2)] + \mathsf{Pr}_{D_3}[(X,Z) = (x_C,2)] \\
&\overset{\text{Table 2}}{=} \frac{c\alpha}{2}.
\end{aligned}$$

Further, for $Z = 2$, we have

$$\begin{aligned}
\mathsf{E}_{D_3}[f(X,Z) \mid Z = 2] &= \sum_{r \in X} f(r,2) \cdot \mathsf{Pr}[X = r \mid Z = 2] \\
&\overset{(59)}{=} \mathsf{Pr}_{D_3}[X = x_C \mid Z = 2] \\
&\overset{\text{Table 2}}{=} c. && (60)
\end{aligned}$$

Similarly, for $Z = 1$, we have

$$\begin{aligned}\mathsf{E}_{D_3}[f(X,Z) \mid Z = 1] &= \sum_{r \in X} f(r,1) \cdot \Pr[X = r \mid Z = 1] \\ &\overset{(59)}{=} \Pr_{D_3}[X = x_A \mid Z = 1] \\ &\overset{\text{Table 2}}{=} \frac{c \cdot (1-\alpha/2)}{(1-\alpha/2)} \\ &= c. \end{aligned} \tag{61}$$

Thus, we can compute $\mathsf{DI}_{D_3}(f)$ as follows

$$\mathsf{DI}_{D_3}(f) = \frac{\min_{\ell \in [p]} \mathsf{E}_{D_3}[f(X,Z) \mid Z = \ell]}{\max_{\ell \in [p]} \mathsf{E}_{D_3}[f(X,Z) \mid Z = \ell]} \overset{(60),(61)}{=} \frac{c}{c} = 1.$$

Thus, for each $\mathcal{D} \in \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$, we give a classifier $f \in \mathcal{F}$ such that that has error at most $c\alpha/2$ and a statistical rate at least $1 - \alpha$. □

A.3 Proof of Theorem 4.5

In Theorem 4.5, the hypothesis class $\mathcal{F}$ that shatters the set $\{x_A, x_B, x_C, x_D, x_E\} \times [2] \subseteq \mathcal{X} \times [p]$, Assumption 1 holds with a constant $\lambda \in (0, 1/4]$, the fairness threshold $\tau = 1$, and statistical rate is the fairness metric; recall that the statistical rate of a classifier $f \in \mathcal{F}$ is

$$\mathsf{DI}_D(f) := \frac{\min_{\ell \in [p]} \Pr_D[f = 1 \mid Z = \ell]}{\max_{\ell \in [p]} \Pr_D[f = 1 \mid Z = \ell]}.$$

Then, given parameters $\eta \in (0, 1]$ and $\delta \in [0, 1/2)$, our goal is to show that for any

$$\varepsilon < \frac{1}{4} - \frac{2\eta}{5} \text{ and } \nu < \frac{\eta}{10\lambda} \cdot (1 - 4\lambda) - O\left(\frac{\eta^2}{\lambda^2}\right), \tag{62}$$

$\mathcal{F}$ is not (ε, ν)-learnable with perturbation rate η and confidence δ. We prove a more general result: Given parameters $\eta \in (0, 1]$, $\delta \in [0, 1/2)$, for any $c \in (0, \min\{\eta, 2\lambda/9\}]$, we show that for any

$$0 < \varepsilon < \frac{1}{2} - \lambda - 2c \text{ and } 0 < \nu < \frac{c(1-4\lambda)}{2\lambda} - \frac{3c^2}{4\lambda^2}, \tag{63}$$

$\mathcal{F}$ is not (ε, ν)-learnable with perturbation rate η and confidence δ. Setting $c = 2\eta/5$ recovers Theorem 4.5.

Remark A.18. *The proof of Theorem 4.5 has a similar structure to the proof of Theorem 4.4, but the specific distributions constructed are different from those in the proof of Theorem 4.4. The proof of Theorem 4.5 also borrows Proposition A.15 from the proof of Theorem 4.4; Proposition A.15 is proved in Section A.2.2.*

A.3.1 Proof of Theorem 4.5

Proof of Theorem 4.5. Let $\mathcal{L}$ be any learner. Fix any constant

$$c \in (0, \min\{\eta, 2\lambda/9\}], \tag{64}$$

and confidence parameter $\delta \in (0, 1/2)$ (in Definition 3.2). We construct three distributions $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ (parameterized by c) that satisfy the requirements of Proposition A.15:

1. The total variation distance between any two distributions is bounded by η,
2. the distributions have the same marginal on $\mathcal{X}$, and
3. the label Y is independent of the protected attribute Z conditioned on features X.

Suppose $\mathcal{L}$ is given N samples, where

$$N \geq 3 \cdot \ln 2 \cdot (\eta - c)^{-2}. \tag{65}$$

Consider three cases, depending on whether the samples in S are iid from $\mathcal{D}_1$, $\mathcal{D}_2$, or $\mathcal{D}_3$.

By Proposition A.15, in each case, there is an adversary $A \in \mathcal{A}(\eta)$, that can ensure that, with probability at least $1/2$, the perturbed samples $\widehat{S} := A(S)$ have the same distribution as iid samples from $\mathcal{D}_1$. Thus, given $\widehat{S}$, $\mathcal{L}$ cannot differentiate between the three cases with probability at least $1/2$. As a result, with probability at least $1/2$, $\mathcal{L}$ outputs the same classifier, say $f_{\mathsf{Com}} \in \mathcal{F}$, in each case. We show that no $f_{\mathsf{Com}} \in \mathcal{F}$ satisfies the accuracy and fairness guarantee in all three cases.

Lemma A.19 (No good classifier for all cases). *There is no $f \in \mathcal{F}$, such that, for all $k \in [3]$, $\Pr_{D_k}[h(X,Z) \neq Y] < 1/2 - \lambda - c/2$ and $\mathrm{DP}_{D_k}(f) < 1 - c(1-4\lambda)/(2\lambda) + 3c^2/(4\lambda^2)$.*

Lemma A.20 (A good classifier for each case). *For each $k \in [3]$, there is an $f \in \mathcal{F}$ such that $\Pr_{D_k}[h(X,Z) \neq Y] \leq 3\alpha/2$ and $\mathrm{DP}_{D}(h) = 1$.*

Note that for each $k \in [3]$, h_k satisfies the fairness constraint. Thus, an optimal solution $f_k^\star \in \mathcal{F}$ of Program (2) for $\mathcal{D}_k$ and $\tau = 1$ must satisfy

$$\mathsf{Err}_{D_k}(f_k^\star) < \frac{3c}{2}. \tag{66}$$

(Otherwise, we have a contradiction as h_k satisfies the fairness constraints and has a smaller error than $f_k^\star$.) If $\mathcal{L}$ is a (ε, ν)-learner, then ibn the k-th case, with probability at least $1 - \delta > 1/2$, $\mathcal{L}$ must output a classifier f_k which satisfies

$$\mathsf{Err}_{D_k}(f_k) - \mathsf{Err}_{D_k}(f_k^\star) \leq \varepsilon \text{ and } \tau - \mathrm{DP}_{D_k}(f_k) \leq \nu.$$

However, in all cases with probability at least $1/2$, $\mathcal{L}$ outputs f_{Com}. Because $1/2 > \delta$, f_{Com} must satisfy:

$$\text{For all } k \in [3], \quad \mathsf{Err}_{D_k}(f_{\mathsf{Com}}) - \mathsf{Err}_{D_k}(f_k^\star) \leq \varepsilon \text{ and } \tau - \mathrm{DP}_{D_k}(f_{\mathsf{Com}}) \leq \nu. \tag{67}$$

But from Lemma A.19 we know that for each $k \in [3]$ either

$$\mathsf{Err}_{D_k}(f_{\mathsf{Com}}) \geq \frac{1}{2} - \lambda - \frac{c}{2} \quad \text{or} \quad \mathrm{DP}_{D_k}(f_{\mathsf{Com}}) \leq 1 - \frac{c(1-4\lambda)}{2\lambda} + \frac{3c^2}{4\lambda^2}. \tag{68}$$

(Case A) $\mathsf{Err}_{D_k}(f_{\mathsf{Com}}) \geq 1/2 - \lambda - c/2$: In this case, from Equation (67), we have

$$\begin{aligned}
\varepsilon &\geq \frac{1}{2} - \lambda - \frac{c}{2} - \mathsf{Err}_{D_k}(f_k^\star) \\
&\overset{(66)}{>} \frac{1}{2} - \lambda - \frac{c}{2} - \frac{3c}{2} \\
&= \frac{1}{2} - \lambda - 2c \qquad \text{(Using that } c < 1/2 \text{ and } \alpha > 0\text{)}
\end{aligned} \tag{69}$$

But because $\varepsilon \leq \frac{1}{2} - \lambda - 2c$ (see Equation (63)), Equation (69) cannot hold.

(Case B) $\mathrm{DP}_{D_k}(f_{\mathsf{Com}}) \leq 1 - c(1-4\lambda)/(2\lambda) + 3c^2/(4\lambda^2)$: In this case, from Equation (67), we have

$$\nu \geq \frac{c(1-4\lambda)}{2\lambda} - \frac{3c^2}{4\lambda^2}. \tag{70}$$

But because $\nu < c(1-4\lambda)/(2\lambda) - 3c^2/(4\lambda^2)$ (see Equation (63)), Equation (70) does not hold. Therefore, we have a contradiction. Hence, $\mathcal{L}$ is not an (ε, ν)-learner for $\mathcal{F}$. Since the choice of $\mathcal{L}$ was arbitrary, we have shown that there is no learner which (ε, ν)-learns $\mathcal{F}$. It remains to prove the Lemmas A.19 and A.20. □

A.3.2 Proof of Lemma A.19

Fix any $c \in (0, \min\{\eta, 2\lambda/9\}]$. Set $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ to be unique distributions with marginal distribution specified in Table 3, such that, for any draw $(X, Y, Z) \sim \mathcal{D}_k$ ($k \in [3]$) Y takes the value $\mathtt{I}[X \neq x_E]$, i.e.,

$$Y = \begin{cases} 1 & \text{if } X \neq x_E, \\ 0 & \text{otherwise.} \end{cases} \tag{71}$$

In particular, the construction of $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ ensures that the total variation distance between any pair of distributions is less than c.

(a) D_1: $\Pr_{D_1}[(X,Z) = (r,s)]$ for $(r,s) \in \{x_A, x_B, x_C, x_D, x_E\} \times [2]$.

	x_A	x_B	x_C	x_D	x_E
1	0	$c/2$	$c/2$	$1/2-\lambda-c/2$	$1/2-\lambda-c/2$
2	c	$c/2$	$c/2$	$\lambda - c$	$\lambda - c$

(b) D_2: $\Pr_{D_2}[(X,Z) = (r,s)]$ for $(r,s) \in \{x_A, x_B, x_C, x_D, x_E\} \times [2]$.

	x_A	x_B	x_C	x_D	x_E
1	$c/2$	0	$c/2$	$1/2-\lambda-c/2$	$1/2-\lambda-c/2$
2	$c/2$	c	$c/2$	$\lambda - c$	$\lambda - c$

(c) D_3: $\Pr_{D_3}[(X,Z) = (r,s)]$ for $(r,s) \in \{x_A, x_B, x_C, x_D, x_E\} \times [2]$.

	x_A	x_B	x_C	x_D	x_E
1	$c/2$	$c/2$	0	$1/2-\lambda-c/2$	$1/2-\lambda-c/2$
2	$c/2$	$c/2$	c	$\lambda - c$	$\lambda - c$

Table 3: Marginal distributions of D_1, D_2, D_3 over X. Recall that for a sample $(X, Y, Z) \sim D_k$ ($k \in [3]$), Y takes the value $\mathbb{I}[X = x_E]$.

Proof of Lemma A.19. Our goal is to show that for every $f \in \mathcal{F}$, there is a $k \in [3]$, such that, f has error at least $1/2 - \lambda - c/2$ or statistical rate at most $1 + c(1 - 4\lambda)/(2\lambda) + \frac{3c^2}{(4\lambda^2)}$ with respect to $\mathcal{D}_k$. Since $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ are supported on subsets of $\{x_A, x_B, x_C, x_D, x_E\} \times [2]$, it suffices to consider the restriction of $\mathcal{F}$ to $\{x_A, x_B, x_C, x_D, x_E\} \times [2]$. There at most 2^{10} classifiers in this restriction. We partition them into four cases.

(Case A) $f(x_D, 1) = 0$ **or** $f(x_E, 1) = 1$**:**

(Case A.1) $f(x_D, 1) = 0$**:** For any $k \in [3]$ it holds that

$$\begin{aligned}
\mathrm{Err}_{D_k}(f) &= \sum_{r \in \mathcal{X}, s \in [p]} \Pr_{D_k}[f(X,Z) \neq Y \mid (X,Z) = (r,s)] \cdot \Pr[(X,Z) = (r,s)] \\
&\geq \Pr_{D_k}[f(X,Z) \neq Y \mid X = x_D, Z = 1] \cdot \Pr[X = x_D, Z = 1] && (\forall\, x \in \mathcal{X}, f(x,2) \geq 0) \\
&\overset{(71)}{\geq} \Pr[X = x_D, Z = 1] && \text{(Using that, in this case, } f(x_D,1) = 0) \\
&\overset{\text{Table 3}}{=} 1/2 - \lambda - c/2.
\end{aligned}$$

Thus, in Case A.1, f has error at least $1/2 - \lambda - c/2$ on each of $\mathcal{D}_1, \mathcal{D}_2$, and $\mathcal{D}_3$.

(Case A.2) $f(x_E, 1) = 1$**:** For any $k \in [3]$ it holds that

$$\begin{aligned}
\mathrm{Err}_{D_k}(f) &= \sum_{r \in \mathcal{X}, s \in [p]} \Pr_{D_k}[f(X,Z) \neq Y \mid (X,Z) = (r,s)] \cdot \Pr[(X,Z) = (r,s)] \\
&\geq \Pr_{D_k}[f(X,Z) \neq Y \mid X = x_E, Z = 1] \cdot \Pr[X = x_E, Z = 1] && (\forall\, x \in \mathcal{X}, f(x,2) \geq 0) \\
&\overset{(71)}{\geq} \Pr[X = x_E, Z = 1] && \text{(Using that, in this case, } f(x_E,1) = 1) \\
&\overset{\text{Table 3}}{=} 1/2 - \lambda - c/2.
\end{aligned}$$

Thus, in Case A.2, f has error at least $1/2 - \lambda - c/2$ on each of $\mathcal{D}_1, \mathcal{D}_2$, and $\mathcal{D}_3$.

In the rest of the cases, assume that $f(x_D, 1) = 1$ and $f(x_E, 1) = 0$.

(Case B) $f(x_D, 2) = f(x_E, 2)$**:**

(Case B.1) $f(x_D, 2) = f(x_E, 2) = 0$**:** For any $k \in [3]$ and $Z = 2$ it holds that

$$\begin{aligned}
\mathsf{E}_{D_k}[f(X,Z) \mid Z=2] &= \sum_{r \in \mathcal{X}} f(r,2) \cdot \Pr_{D_k}[X = r \mid Z = 2] \\
&= \sum_{r \in \mathcal{X} \setminus \{x_D, x_E\}} f(r,2) \cdot \Pr_{D_k}[X = r \mid Z = 2] \\
&\qquad \text{(Using that, in this case, } f(x_D,2) = f(x_E,2) = 0\text{)} \\
&\leq \sum_{r \in \mathcal{X} \setminus \{x_D, x_E\}} \Pr_{D_k}[X = r \mid Z = 2] \\
&\qquad \text{(Using that } \forall\, r \in \mathcal{X},\ f(r,2) \leq 1\text{)} \\
&\overset{\text{Table 3}}{\leq} \frac{2c}{2\lambda}. \qquad (72)
\end{aligned}$$

For $Z = 1$, we have

$$\begin{aligned}
\mathsf{E}_{D_k}[f(X,Z) \mid Z=1] &= \sum_{r \in \mathcal{X}} f(r,1) \cdot \Pr_{D_k}[X = r \mid Z = 1] \\
&\geq 1 \cdot \Pr_{D_k}[X = x_D \mid Z = 1] \qquad \text{(From Case A, } f(x_D,1) = 1\text{)} \\
&= \frac{1 - 2\lambda - c}{2(1-2\lambda)} \\
&\geq \frac{5/6 - 2\lambda}{2} \qquad \text{(Using that } c \leq 1/6 \text{ and } \lambda > 0\text{)} \\
&\geq 1/6. \qquad \text{(Using that } \lambda \leq 1/4\text{)} \quad (73)
\end{aligned}$$

Since $c < \lambda/2$, the statistical rate of f is as follows

$$\mathrm{SR}_{D_k}(f) = \frac{\min_{\ell \in [p]} \mathsf{E}_{D_k}[f(X,Z) \mid Z = \ell]}{\max_{\ell \in [p]} \mathsf{E}_{D_k}[f(X,Z) \mid Z = \ell]} \overset{(72),(73)}{\leq} \frac{6c}{\lambda}.$$

Thus, in Case B.1, f has a statistical rate at most $\frac{6c}{\lambda}$ on each of $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$.

(Case B.2) $f(x_D, 2) = f(x_E, 2) = 1$**:** For any $k \in [3]$ and $Z = 2$ it holds that

$$\begin{aligned}
\mathsf{E}_{D_k}[f(X,Z) \mid Z=2] &= \sum_{r \in \mathcal{X}} f(r,2) \cdot \Pr_{D_k}[X = r \mid Z = 2] \\
&\geq \sum_{r \in \{x_D, x_E\}} f(r,2) \cdot \Pr_{D_k}[X = r \mid Z = 2] \\
&\qquad \text{(Using that } f(x_D,2) = f(x_E,2) = 1 \text{ and for all } r \in \mathcal{X},\ f(r,2) \geq 0\text{)} \\
&\overset{\text{Table 3}}{=} 1 - \frac{c}{\lambda}. \qquad (74)
\end{aligned}$$

For $Z = 1$, we have

$$\begin{aligned}
\mathsf{E}_{D_k}[f(X,Z) \mid Z=1] &= \sum_{r \in \mathcal{X}} f(r,1) \cdot \Pr_{D_k}[X = r \mid Z = 1] \\
&= \sum_{r \in \mathcal{X} \setminus \{x_E\}} f(r,1) \cdot \Pr_{D_k}[X = r \mid Z = 1] \\
&\qquad \text{(From Case B, } f(x_E,1) = 0\text{)} \\
&\leq \sum_{r \in \mathcal{X} \setminus \{x_E\}} 1 \cdot \Pr_{D_k}[X = r \mid Z = 1] \\
&\qquad \text{(Using that } \forall\, r \in \mathcal{X},\ f(r,2) \leq 1\text{)} \\
&\overset{\text{Table 3}}{=} \frac{1}{2} - \frac{c}{2(1-2\lambda)}. \qquad (75)
\end{aligned}$$

Using this, we can compute an upper bound on the statistical rate of f as follows

$$\begin{aligned}
\mathcal{D}_k(f) &= \frac{\min_{\ell \in [p]} \mathsf{E}_{D_k}[f(X,Z) \mid Z = \ell]}{\max_{\ell \in [p]} \mathsf{E}_{D_k}[f(X,Z) \mid Z = \ell]} \\
&\overset{(75),(74)}{\leq} \frac{\frac{1}{2} - \frac{c}{2(1-2\lambda)}}{1 - \frac{c}{\lambda}} \\
&= \frac{1}{2} \cdot \frac{1 - \frac{c}{(1-2\lambda)}}{1 - \frac{c}{\lambda}} \cdot \frac{1 + \frac{3c}{\lambda}}{1 + \frac{3c}{\lambda}} \\
&= \frac{1}{2} \cdot \frac{1 - \frac{c}{(1-2\lambda)}}{1 + \frac{2c}{\lambda} - \frac{3c^2}{\lambda^2}} \cdot \left(1 + \frac{3c}{\lambda}\right) \\
&\leq \frac{1}{2} + \frac{3c}{2\lambda}. \qquad \text{(Using that for all } 0 < c, \lambda \leq \tfrac{1}{4}\text{, if } \tfrac{c}{\lambda} \leq \tfrac{1}{3}\text{, then } \tfrac{c}{1-2\lambda} \geq \tfrac{3c^2}{\lambda^2} - \tfrac{2c}{\lambda}\text{)}
\end{aligned}$$

Thus, in Case B.2, f has a statistical rate at most $\frac{1}{2} + \frac{3c}{2\lambda}$ on each of $\mathcal{D}_1, \mathcal{D}_2$, and $\mathcal{D}_3$.

In the rest of the cases, we assume that $f(x_D, 2) \neq f(x_E, 2)$.

(Case C) $\sum_{r \in \{x_A, x_B, x_C\}} f(r,2) \geq 2$**:** In Case C, f outputs 0 on at most one point in the tuple

$$L = ((x_A, 2), (x_B, 2), (x_C, 2)).$$

If f takes a value of 0 on a point, say L_j, in L, then fix $k \in [3]$, such that $\mathcal{D}_k(L_j) = c/2$. (Such a distribution always exists by construction). Otherwise, fix any $k \in [3]$. For $Z = 2$, we have

$$\begin{aligned}
\mathsf{E}_{D_k}[f(X,Z) \mid Z = 2] &= \sum_{r \in X} f(r,2) \cdot \mathsf{Pr}_{D_k}[X = r \mid Z = 2] \\
&\geq \frac{\lambda + c}{2\lambda}. \qquad \text{(By our choice of } k\text{)} \quad (76)
\end{aligned}$$

For $Z = 1$, we have

$$\begin{aligned}
&\mathsf{E}_{D_k}[f(X,Z) \mid Z = 1] \\
&= \sum_{r \in X} f(r,1) \cdot \mathsf{Pr}_{D_k}[X = r \mid Z = 1] \\
&= \mathsf{Pr}_{D_k}[X = x_D \mid Z = 1] + \sum_{r \in \{x_A, x_B, x_C\}} f(r,1) \cdot \mathsf{Pr}_{D_k}[X = r \mid Z = 1] \\
&\qquad \text{(From previous cases, } f(x_D, 1) = 1 \text{ and } f(x_E, 1) = 0\text{)} \\
&\overset{\text{Table 3}}{\leq} \frac{(1 - 2\lambda + c)}{2(1 - 2\lambda)}. \qquad (77)
\end{aligned}$$

We can compute the statistical rate of f using Equations (76) and (77)

$$\begin{aligned}
\mathcal{D}_k(f) &= \frac{\min_{\ell \in [p]} \mathsf{E}_{D_k}[f(X,Z) \mid Z = \ell]}{\max_{\ell \in [p]} \mathsf{E}_{D_k}[f(X,Z) \mid Z = \ell]} \\
&\overset{(77),(76)}{\leq} \frac{1 + \frac{c}{1-2\lambda}}{1 + \frac{c}{2\lambda}} \\
&\leq \left(1 + \frac{c}{1-2\lambda}\right) \cdot \left(1 - \frac{c}{2\lambda} + \frac{c^2}{4\lambda^2}\right) \\
&\qquad \text{(Using that for all } x \geq 0\text{, } \tfrac{1}{1+x} \leq 1 - x + x^2\text{)} \\
&= 1 + \frac{c}{1-2\lambda} - \frac{c}{2\lambda} + \frac{c^2}{4\lambda^2} \cdot \left(1 + \frac{c}{1-2\lambda}\right) - \frac{c^2}{2\lambda(1-2\lambda)} \\
&\leq 1 - \frac{c(1-4\lambda)}{2\lambda(1-2\lambda)} + \frac{3c^2}{4\lambda^2} \qquad \text{(Using that } 0 < \lambda \leq \tfrac{1}{4} \text{ and } c > 0\text{)} \\
&\leq 1 - \frac{c}{2\lambda} \cdot (1 - 4\lambda) + \frac{3c^2}{4\lambda^2}. \qquad \text{(Using that } \lambda \geq 0\text{)}
\end{aligned}$$

Thus, in Case C, f has a statistical rate at most $1 - \frac{c}{2\lambda} \cdot (1 - 4\lambda) + \frac{3c^2}{4\lambda^2}$ on the chosen distribution $\mathcal{D}_k$.

(Case D) $\sum_{r \in \{x_A, x_B, x_C\}} f(r, 2) \leq 1$**:** In Case D, f outputs of 1 on at most one point in the tuple

$$L = ((x_A, 2), (x_B, 2), (x_C, 2)).$$

If f takes a value of 1 on a point, say L_j, in L, then fix $k \in [3]$, such that $\mathcal{D}_k(L_j) = c/2$. (Such a distribution always exists by construction). Otherwise, fix any $k \in [3]$. For $Z = 2$, we have

$$\mathbb{E}_{D_k}[f(X, Z) \mid Z = 2] = \sum_{r \in X} f(r, 2) \cdot \Pr_{D_k}[X = r \mid Z = 2] \leq \frac{\lambda - c}{2\lambda}.$$

(By our choice of k) (78)

For $Z = 1$, we have

$$\begin{aligned}
&\mathbb{E}_{D_k}[f(X, Z) \mid Z = 1] \\
&= \sum_{r \in X} f(r, 1) \cdot \Pr_{D_k}[X = r \mid Z = 1] \\
&= \Pr_{D_k}[X = x_D \mid Z = 1] + \sum_{r \in \{x_A, x_B, x_C\}} f(r, 1) \cdot \Pr_{D_k}[X = r \mid Z = 1] \\
&\qquad \text{(From previous cases, } f(x_D, 1) = 1 \text{ and } f(x_E, 1) = 0) \\
&\overset{\text{Table 3}}{\geq} \frac{(1 - 2\lambda - c)}{2(1 - 2\lambda)}. \qquad (79)
\end{aligned}$$

We can compute the statistical rate of f using Equations (78) and (79)

$$\begin{aligned}
_{D_k}(f) &= \frac{\min_{\ell \in [p]} \mathbb{E}_{D_k}[f(X, Z) \mid Z = \ell]}{\max_{\ell \in [p]} \mathbb{E}_{D_k}[f(X, Z) \mid Z = \ell]} \\
&\overset{(79),(78)}{\leq} \frac{1 - \frac{c}{2\lambda}}{1 - \frac{c}{1 \; 2\lambda}} \\
&\leq \left(1 - \frac{c}{2\lambda}\right) \cdot \left(1 + \frac{c}{1 - 2\lambda} + \frac{2c^2}{(1 - 2\lambda)^2}\right) \\
&\qquad \text{(Using that for all } x \leq \tfrac{1}{2}, \tfrac{1}{1 \; x} \leq 1 + x + 2x^2) \\
&= 1 + \frac{c}{1 - 2\lambda} - \frac{c}{2\lambda} + \frac{2c^2}{(1 - 2\lambda)^2} \cdot \left(1 - \frac{c}{2\lambda}\right) - \frac{c^3}{2\lambda(1 - 2\lambda)^2} \\
&\leq 1 - \frac{c(1 - 4\lambda)}{2\lambda(1 - 2\lambda)} + 8c^2 \qquad \text{(Using that } 0 < \lambda \leq \tfrac{1}{4} \text{ and } c > 0) \\
&\leq 1 - \frac{c}{2\lambda} \cdot (1 - 4\lambda) + \frac{3c^2}{4\lambda^2}. \qquad \text{(Using that } 0 \leq \lambda \leq \tfrac{1}{4})
\end{aligned}$$

Thus, in this case, f has a statistical rate at most $1 - \frac{c}{2\lambda} \cdot (1 - 4\lambda) + \frac{3c^2}{4\lambda^2}$ on the chosen distribution $\mathcal{D}_k$. □

A.3.3 Proof of Lemma A.20

Proof of Lemma A.20. For each distribution $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$, we will give classifiers $f_1, f_2, f_3 \in \mathcal{F}$, such that, for each $k \in [3]$, f_k has error at most $3\alpha/2$ and a statistical rate 1 with respect to $\mathcal{D}_k$. The idea is to choose a classifier f_k such that, conditioned on a value of Z, they label exactly $1/2$-fraction of the samples as positive on $\mathcal{D}_k$. Thus, they have a statistical rate of 1. Then, by the construction it follows that these classifiers also have a small error. We give the classifier f_1 for distribution $\mathcal{D}_1$. Classifiers f_2 and f_3 follow by symmetry.

Define $f_1 \in \mathcal{F}$ to be the following classifier

$$f_1(x, z) := \begin{cases} 1 & \text{if } x \in \{x_B, x_D\}, \\ 1 & \text{if } x = x_C \text{ and } z = 2, \\ 0 & \text{otherwise.} \end{cases}$$

Using Equation (71) and Table 3, we get that

$$\mathsf{Err}_{D_1}(f_1) = \frac{3c}{2}.$$

For $Z = 2$, we have

$$\begin{aligned}\mathsf{E}_{D_1}[f_1(X,Z) \mid Z = 2] &= \sum_{r \in \mathcal{X}} f_1(r,2) \cdot \mathsf{Pr}_{D_1}[X = r \mid Z = 2] \\ &= \mathsf{Pr}_{D_1}[X \in \{x_B, x_C, x_D\} \mid Z = 2] \\ &\overset{\text{Table 3}}{=} \frac{1}{2}. \end{aligned} \tag{80}$$

For $Z = 1$, we have

$$\begin{aligned}\mathsf{E}_{D_1}[f(X,Z) \mid Z = 1] &= \sum_{r \in \mathcal{X}} f_1(r,1) \cdot \mathsf{Pr}_{D_1}[X = r \mid Z = 1] \\ &= \mathsf{Pr}_{D_1}[X \in \{x_B, x_D\} \mid Z = 1] \\ &\overset{\text{Table 3}}{=} \frac{1}{2}. \end{aligned} \tag{81}$$

Thus, we have

$$_{D_1}(h) = \frac{\min_{\ell \in [p]} \mathsf{E}_{D_1}[f(X,Z) \mid Z = \ell]}{\max_{\ell \in [p]} \mathsf{E}_{D_1}[f(X,Z) \mid Z = \ell]} \overset{(80),(81)}{=} 1.$$

□

A.4 Impossibility result omitted from Section 4

Theorem A.21 (Fairness guarantee of Theorem 4.3 is optimal up to constant factors). *For all perturbation rates $\eta \in (0, 1/2]$, constants $\lambda \in (0, 1/4]$, and confidence parameters $\delta \in [0, 1/2)$, if the fairness metric is statistical rate and $\tau = 1$, then even with the knowledge that Assumption 1 holds with a (known) constant λ, for any $\varepsilon < \eta$ and $\upsilon < \tau$ 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 $x_A, x_B, x_C \in \mathcal{X}$.*

Thus, Theorem A.21 shows that for any $\varepsilon < \eta$, no algorithm can (ε, ν)-learn any hypothesis classes $\mathcal{F}$ satisfying mild assumptions; this shows that the accuracy guarantee in Theorem 4.3 is optimal up to a constant factor.

A.4.1 Proof of Theorem A.21

Theorem A.21's assumes that $\mathcal{F}$ shatters a set $\{x_A, x_B, x_C\} \times [2] \subseteq \mathcal{X} \times [p]$, Assumption 1 holds with a constant $\lambda \in (0, 1/4]$, $\tau = 1$, and the statistical rate is the fairness metric; recall that the statistical rate of $f \in \mathcal{F}$ with respect to distribution $\mathcal{D}$ is

$$_{D}(f) := \frac{\min_{\ell \in [p]} \mathsf{Pr}_D[f = 1 \mid Z = \ell]}{\max_{\ell \in [p]} \mathsf{Pr}_D[f = 1 \mid Z = \ell]}.$$

Then, given parameters $\eta \in (0, 1/2]$, $\lambda \in (0, 1/4]$, and $\delta \in [0, 1/2)$ our goal is to prove that for any

$$0 < \varepsilon < \eta \text{ and } \nu \in [0,1],$$

$\mathcal{F}$ is not (ε, ν)-learnable with perturbation rate η and confidence δ.

Our proof is inspired by the approach of [39, Theorem 1] and [11, Theorem 1]. It has a similar structure to the proof of Theorem 4.4; but uses a different construction in which label Y depends on Z conditioned on X. (For a reader who has read the proof of Theorem 4.4, this means that condition (C3) in Proposition A.15 does not hold; and we have to give a different algorithm for the adversary to "hide" the true distribution.)

Proof of Theorem A.21. Let $\mathcal{L}$ be any learner. We construct two distributions $\mathcal{P}$ and $\mathcal{Q}$ (parameterized by η) such that Lemma A.22 holds (see Figures 2 and 3 for a description of $\mathcal{P}$ and $\mathcal{Q}$).

	x_A	x_B	x_C
1	$\eta/2$	$1/2 - \eta$	$\eta/2$
2	$\eta/2$	$1/2 - \eta$	$\eta/2$

Figure 2: Marginal distributions of P and Q over X [2]. For (r, s) 2 $f x_A, x_B, x_C g$ [2], the corresponding cell denotes $\Pr_P [(X, Z) = (r, s)] = \Pr_Q [(X, Z) = (r, s)]$.

Lemma A.22 (Adversary can hide the true distribution). *For both $\mathcal{D} \in \{\mathcal{P}, \mathcal{Q}\}$, given $N \in \mathbb{N}$ iid samples S from $\mathcal{D}$, if $\eta \cdot N$ is integral, then there is a distribution $\mathcal{D}_{\mathsf{Mix}}$ over $\mathcal{X} \times \{0, 1\} \times [p]$ and an adversary $A \in \mathcal{A}(\eta)$ such that with probability at least $1/2$ (over the draw of S) samples in $\widehat{S} := A(S)$ are distributed as iid draws from $\mathcal{D}_{\mathsf{Mix}}$.*

Suppose $\mathcal{L}$ is given $N \in \mathbb{N}$ samples S; where $N \cdot \eta$ is integral. Consider two cases: In the first case, S is iid from $\mathcal{P}$ and in the second case, S is iid from $\mathcal{Q}$. By Lemma A.22, in each case there is an $A \in \mathcal{A}(\eta)$ such that, with probability at least $1/2$, the perturbed samples $\widehat{S} := A(S)$ are have the distribution $\mathcal{D}_{\mathsf{Mix}}$.

Thus, given $\widehat{S}$, with probability at least $1/2$, the learner $\mathcal{L}$ cannot identify the distribution from which S was drawn. As a result, with probability at least $1/2$, $\mathcal{L}$ must to output a common classifier, say $f_{\mathsf{Com}} \in \mathcal{F}$, which satisfies the accuracy and fairness guarantee for both $\mathcal{P}$ and $\mathcal{Q}$. We show that no $f_{\mathsf{Com}} \in \mathcal{F}$ satisfies the accuracy and fairness guarantee for both $\mathcal{P}$ and $\mathcal{Q}$.

Lemma A.23 (No good classifier for both $\mathcal{P}$ and $\mathcal{Q}$). *There is no $f \in \mathcal{F}$, such that*

$$\mathsf{Err}_P[f] < \eta \text{ and } \mathsf{Err}_Q[f] < \eta.$$

Lemma A.24 (A good classifier for each $\mathcal{P}$ and $\mathcal{Q}$). *For each $\mathcal{D} \in \{\mathcal{P}, \mathcal{Q}\}$, there is $f^\star_{\mathcal{D}} \in \mathcal{F}$ such that*

$$\mathsf{Err}_{\mathcal{D}}[f^\star_{\mathcal{D}}] = 0 \text{ and } \quad _{\mathcal{D}}(f_{\mathcal{D}}) = 1.$$

Note that $f^\star_P$ (respectively $f^\star_Q$) has perfect accuracy and satisfies the fairness constraints with respect to $\mathcal{P}$ (respectively $\mathcal{Q}$). If $\mathcal{L}$ is a (ε, ν)-learner, then the first case, with probability at least $1 - \delta > 1/2$, $\mathcal{L}$ must output a classifier f_1 which satisfies

$$\mathsf{Err}_P(f_1) - \mathsf{Err}_P(f^\star_P) \leq \varepsilon \text{ and } \tau - \quad _P(f_1) \leq \nu.$$

And in the second case, with probability at least $1 - \delta$, $\mathcal{L}$ must output a classifier f_2 which satisfies

$$\mathsf{Err}_Q(f_2) - \mathsf{Err}_Q(f^\star_Q) \leq \varepsilon \text{ and } \tau - \quad _Q(f_2) \leq \nu.$$

However, in both cases, with probability at least $1/2$, $\mathcal{L}$ outputs f_{Com}. Because $1 - \delta > 1/2$, f_{Com} must satisfy

$$\mathsf{Err}_P(f_{\mathsf{Com}}) - \mathsf{Err}_P(f^\star_P) \leq \varepsilon \text{ and } \mathsf{Err}_Q(f_{\mathsf{Com}}) - \mathsf{Err}_Q(f^\star_Q) \leq \varepsilon \tag{82}$$

$$\tau - \quad _P(f_{\mathsf{Com}}) \leq \nu \text{ and } \tau - \quad _Q(f_{\mathsf{Com}}) \leq \nu \tag{83}$$

But since $\varepsilon < \eta$, Equation (82) contradicts Lemma A.22. Thus, $\mathcal{L}$ is not an (ε, ν)-learner for $\mathcal{F}$. Since the choice of $\mathcal{L}$ was arbitrary, we have shown that there is no learner which (ε, ν)-learns $\mathcal{F}$.

It remains to prove the Lemma A.22, Lemma A.23, and Lemma A.24. □

A.4.2 Proof of Lemma A.22

Set $\mathcal{P}$ and $\mathcal{Q}$ to be the unique distributions satisfying the following properties:

(P1) $\mathcal{P}$ and $\mathcal{Q}$ have the same marginal distribution on $\mathcal{X} \times [p]$; their marginal distribution on $\mathcal{X} \times [p]$ is given in Figure 2.

(P2) for a sample $(X, Y, Z) \sim \mathcal{P}$, Figure 3(a) denotes the distribution of Y conditioned on X and Z, and for a sample $(X, Y, Z) \sim \mathcal{Q}$, Figure 3(b) denotes the distribution of Y conditioned on X and Z.

	x_A	x_B	x_C
1	1	1	0
2	0	1	1

(a) *Distribution P:* For (r, s) 2 $f x_A, x_B, x_C g$ [2], the corresponding cell denotes the value of Y conditioned on X and Z, when sample (X, Y, Z) P.

	x_A	x_B	x_C
1	0	1	1
2	1	1	0

(b) *Distribution Q:* For (r, s) 2 $f x_A, x_B, x_C g$ [2], the corresponding cell denotes the value of Y conditioned on X and Z, when sample (X, Y, Z) Q.

Figure 3: Figure 3(a) denotes the conditional distribution of Y when (X, Y, Z) P. Figure 3(b) denotes the conditional distribution of Y when (X, Y, Z) Q.

From Figures 2 and 3, it follows that, conditional on $X = x_B$ the samples $(X, Y, Z) \sim \mathcal{P}$ and $(X, Y, Z) \sim \mathcal{Q}$ follow the same distribution. Then, the goal of the adversary A (in Lemma A.22) is to specifically perturb the samples with $X \neq x_B$ so that the perturbed sample $(X, Y, \widetilde{Z})$ follows the same distribution irrespective of whether the original sample (X, Y, Z) is drawn from $\mathcal{P}$ or $\mathcal{Q}$.

Proof of Lemma A.22. Let $S := \{(X_i, Z_i, Y_i)\}_{i2[N]}$. Let $A \in \mathcal{A}(\eta)$ execute the following algorithm:

1. **For** $i \in [N]$ **do:**
   (a) **Sample** t_i uniformly at random from $[0, 1]$
   (b) **If** $X_i = x_B$ **or** $t_i \leq 1/2$ **then: Set** $\widetilde{Z}_i := Z_i$,
   (c) **Otherwise: Set** $\widetilde{Z}_i := 3 - Z_i$ *//If $Z_i = 1$ set $\widetilde{Z}_i = 2$, if $Z_i = 2$ set $\widetilde{Z}_i = 1$*
2. **If** $\sum_{i2[N]} \mathbb{I}[\widetilde{Z}_i \neq Z_i] \leq \eta \cdot N$ **then: return** $\{(X_i, \widetilde{Z}_i, Y_i)\}_{i2[N]}$,
3. **Otherwise: return** $\{(X_i, Z_i, Y_i)\}_{i2[N]}$

Intuitively, A perturbs the samples with $X \neq x_B$, such that whether S is drawn from $\mathcal{P}$ or $\mathcal{Q}$, the following invariant holds: For any $s \in [2]$,

$$\Pr[Y = 0 \mid X \neq x_B, \widetilde{Z} = s] = \Pr[Y = 1 \mid X \neq x_B, \widetilde{Z} = s] = 1/2.$$

We will show that combining this with the facts that $\mathcal{P}$ and $\mathcal{Q}$ have the same marginal distribution on $\mathcal{X} \times [p]$, and that conditioned on $X = x_B$ samples from $\mathcal{P}$ and $\mathcal{Q}$ have the same distribution suffices to prove Lemma A.22.

The check in Step 2 ensures that A never perturbs more than $\eta \cdot N$ samples; thus, A is admissible in the η-Hamming model. Lemma A.25 shows that the check in Step 2 is true with probability at least $1/2$.

Lemma A.25. *If $\eta \cdot N$ is integral, then with probability at least $1/2$, $\sum_{i2[N]} \mathbb{I}[\widetilde{Z}_i \neq Z_i] < \eta \cdot N$.*

Proof. For all $i \in [N]$, let $C_i \in \{0, 1\}$ be a random variable indicating if $\widetilde{Z}_i \neq Z_i$. It suffices to show that with probability at least $1/2$, $\sum_{i2[N]} C_i \leq \eta \cdot N$. Towards this, we first compute $\Pr[C_i = 1]$ as follows

$$\begin{aligned}
\Pr[C_i = 1] &= \Pr[Z_i \neq \widetilde{Z}_i] && \text{(Definition of } C_i\text{)} \\
&= \Pr[X_i \neq x_B, t_i > 1/2] && \text{(Using Step 1(b) and Step 1(c))} \\
&= \Pr[X_i \neq x_B] \cdot \Pr[t_i > 1/2] && \text{(}(X_i, Y_i, Z_i)\text{ and } t_i \text{ are independent)} \\
&= 2\eta \cdot \frac{1}{2}. && \text{(Using } t_i \text{ is uniformly at random from } [0, 1] \text{ and Figure 2)}
\end{aligned}$$

Since for each i, the sample (X_i, Y_i, Z_i) and the point t_i is drawn independently of each other, other samples, and other points, it follows that the random variables $\{C_i\}_i$ are independent. Hence, $\sum_{i \in [N]} C_i$ follows a binomial distribution with mean $\eta \cdot N$. Since $\eta \cdot N$ is integral, we have that the median of the distribution of $\sum_{i \in [N]} C_i$ is $\eta \cdot N$. Thus, we get that $\Pr[\sum_{i \in [N]} C_i \leq \eta \cdot N] = 1/2$. □

Thus, Lemma A.25 shows that with probability at least $1/2$, $\widehat{S} := \{(X_i, \widetilde{Z}_i, Y_i)\}_{i \in [N]}$.

Lemma A.26. *There is a distribution $\mathcal{D}_{\mathsf{Mix}}$, such that, for all $\mathcal{D} \in \{\mathcal{P}, \mathcal{Q}\}$, given S iid from $\mathcal{D}$, the samples $\widetilde{S} := \{(X_i, \widetilde{Z}_i, Y_i)\}_{i \in [N]}$ (computed in the algorithm of A) are independent of each other and is distributed according $\mathcal{D}_{\mathsf{Mix}}$.*

Proof. Since for each $i \in [N]$ the sample (X_i, Y_i, Z_i) and the point t_i is drawn independently of others, it follows that the samples $(X_i, \widetilde{Z}_i, Y_i)$ are independent of each other.

Next, we show that $\widetilde{S}$ follows the same distribution whether S is iid from $\mathcal{P}$ or from $\mathcal{Q}$. The distribution $\mathcal{D}_{\mathsf{Mix}}$ will be implicit in the proof. Since samples in S are iid and the algorithm of A does not depend on i, it follows that all samples in $\widetilde{S}$ follow the same distribution. Thus, it suffices to consider any one sample in $\widetilde{S}$. Fix any $i \in [N]$. Consider the i-th sample $(X_i, \widetilde{Z}_i, Y_i)$. For the rest of the proof we drop the subscript i in $(X_i, \widetilde{Z}_i, Y_i)$, (X_i, Z_i, Y_i), and t_i.

Fix any $y \in \{0, 1\}$ and $s \in [2]$. Then, it holds that

$$\begin{aligned}
\Pr[X = x_A, Y = y, \widetilde{Z} = s] = \quad & \Pr[X = x_A, Y = y, \widetilde{Z} = s \mid t \leq 1/2] \cdot \Pr[t \leq 1/2] \\
& + \Pr[X = x_A, Y = y, \widetilde{Z} = s \mid t > 1/2] \cdot \Pr[t > 1/2] \\
= \quad & \Pr[X = x_A, Y = y, Z = s \mid t \leq 1/2] \cdot \Pr[t \leq 1/2] \\
& + \Pr[X = x_A, Y = y, Z = 3 - s \mid t > 1/2] \cdot \Pr[t > 1/2] \\
= \quad & \Pr[X = x_A, Y = y, Z = s] \cdot 1/2 \\
& + \Pr[X = x_A, Y = y, Z = 3 - s] \cdot 1/2 \\
= \quad & \Pr[X = x_A, Y = y] \cdot 1/2. \qquad \text{(Using that } Z \in \{0, 1\}\text{)}
\end{aligned}$$

Now, from Figure 2 it follows that $\Pr[X = x_A, Y = y] = \eta/2$ both when $(X, Y, Z) \sim \mathcal{P}$ or when $(X, Y, Z) \sim \mathcal{Q}$. Thus, we get that

$$\Pr[X = x_A, Y = y, \widetilde{Z} = s] = \frac{\eta}{4}.$$

Replacing x_A by x_C in the above argument, we get that

$$\Pr[X = x_C, Y = y, \widetilde{Z} = s] = \frac{\eta}{4}.$$

Finally, since A does not perturb samples with $X = x_B$ (see Step 1(b)), from Figures 2 and 3 it follows that

$$\begin{aligned}
\Pr[X = x_B, Y = y, \widetilde{Z} = s] &= \Pr[X = x_B, Y = y, Z = s] \\
&= \mathtt{I}[y = 1] \cdot \left(\frac{1}{2} - \eta\right).
\end{aligned}$$

Since the choice of $i \in [N]$, $y \in \{0, 1\}$, and $s \in [2]$ was arbitrary. We get that all samples in $\widetilde{S}$ follow the same distribution whether S is iid from $\mathcal{P}$ or from $\mathcal{Q}$. □

Lemma A.22 follows because with probability at least $1/2$, $\widehat{S}$ is $\{(X_i, \widetilde{Z}_i, Y_i)\}_{i \in [N]}$ (by Lemma A.25) and the samples $\{(X_i, \widetilde{Z}_i, Y_i)\}_{i \in [N]}$ are independent and distributed according to $\mathcal{D}_{\mathsf{Mix}}$ (by Lemma A.17). □

A.4.3 Proof of Lemma A.23

Proof of Lemma A.23. Fix any $f \in \mathcal{F}$.

$$
\begin{aligned}
&\mathsf{Err}_P(f) + \mathsf{Err}_Q(f) \\
&\quad = \mathsf{Pr}_{(X,Y,Z)\ \ P}[f(X,Z) \neq Y] + \mathsf{Pr}_{(X,Y,Z)\ \ Q}[f(X,Z) \neq Y] \quad \text{(Using the definition of Err)} \\
&\quad = \sum_{r \in \mathcal{X}, s \in [2]} \mathsf{Pr}_{(X,Y,Z)\ \ P}[X = r, Z = s] \cdot \mathsf{Pr}_{(X,Y,Z)\ \ P}[f(X,Z) \neq Y \mid X = r, Z = s] \\
&\qquad\qquad + \mathsf{Pr}_{(X,Y,Z)\ \ Q}[X = r, Z = s] \cdot \mathsf{Pr}_{(X,Y,Z)\ \ Q}[f(X,Z) \neq Y \mid X = r, Z = s] \\
&\quad \geq \sum_{r \in \{x_A, x_C\}, s \in [2]} \frac{\eta}{2} \cdot \mathsf{Pr}_{(X,Y,Z)\ \ P}[f(X,Z) \neq Y \mid X = r, Z = s] \\
&\qquad\qquad + \frac{\eta}{2} \cdot \mathsf{Pr}_{(X,Y,Z)\ \ Q}[f(X,Z) \neq Y \mid X = r, Z = s] \\
&\quad = \sum_{r \in \{x_A, x_C\}, s \in [2]} \frac{\eta}{2} \cdot \mathsf{Pr}_{(X,Y,Z)\ \ P}[f(X,Z) \neq 0 \mid X = r, Z = s] \\
&\qquad\qquad + \frac{\eta}{2} \cdot \mathsf{Pr}_{(X,Y,Z)\ \ Q}[f(X,Z) \neq 1 \mid X = r, Z = s] \\
&\qquad\qquad\qquad \text{(Using Property (P2) of } \mathcal{P} \text{ and } \mathcal{Q}\text{; see Figure 3)} \\
&\quad = \sum_{r \in \{x_A, x_C\}, s \in [2]} \frac{\eta}{2} \cdot \mathsf{Pr}_{(X,Y,Z)\ \ P}[f(X,Z) \neq 0 \mid X = r, Z = s] \\
&\qquad\qquad + \frac{\eta}{2} \cdot \mathsf{Pr}_{(X,Y,Z)\ \ P}[f(X,Z) \neq 1 \mid X = r, Z = s] \\
&\qquad\qquad\qquad \text{(Using Property (P1) of } \mathcal{P} \text{ and } \mathcal{Q}\text{; see Figure 2)} \\
&\quad = \sum_{r \in \{x_A, x_C\}, s \in [2]} \frac{\eta}{2} \\
&\quad = 2\eta.
\end{aligned}
$$

Since $\mathsf{Err}_P(f), \mathsf{Err}_Q(f) \geq 0$, by the Pigeonhole principle either $\mathsf{Err}_P(f) \geq \eta$ or $\mathsf{Err}_Q(f) \geq \eta$. □

A.4.4 Proof of Lemma A.24

Proof of Lemma A.24. Consider a sample $(X, Y, Z) \sim \mathcal{P}$ Notice that, for some $r \in \mathcal{X}$ and $s \in [2]$, conditioned on $X = r$ and $Z = s$, the label Y is uniquely identified by Figure 3(a). Let $f_P^\star$ be the classifier than given (r, s) predicts the value in the corresponding cell of Figure 3(a). Clearly, $\mathsf{Err}_P(f_P^\star) = 0$. Further, by our assumption that $\mathcal{F}$ shatters the set $\{x_A, x_B, x_C\} \times [2] \subseteq \mathcal{X} \times [p]$, it follows that $f_P^\star \in \mathcal{F}$.

The construction of $f_Q^\star$ follows symmetrically by using Figure 3(b). □

A.5 Additional remarks about the η-Hamming model and theoretical results

In Example A.27, we show that the ratio of the fairness of a classifier $f \in \mathcal{F}$ with respect to the perturbed samples $\widehat{S}$ and with respect to the unperturbed samples S can be 0.

Example A.27 (**Ratio of fairness of a classifier on perturbed and unperturbed samples).** *Let $S := \{(x_i, y_i, z_i)\}_{i \in [N]}$ denote the N unperturbed samples. Suppose that the fairness metric is statistical rate and S has an equal number of samples from each protected group (i.e., for all $\ell \in [p]$, $\sum_{i \in [N]} \mathbb{I}[z_i = \ell] = N/p$.) Consider a classifier $f \in \mathcal{F}$ that has exactly $\eta \cdot N$ positive predictions on each protected group $\ell \in [p]$, i.e., for all $\ell \in [p]$, $|\{i \in [N] \mid f(x_i, z_i) = 1 \text{ and } z_i = \ell\}| = \eta \cdot N$. This implies that $(f, S) = 1$. Fix any protected group $\ell \in [p]$. An adversary $A \in \mathcal{A}(\eta)$, can perturb the protected attributes of all $\eta \cdot N$ samples in the set $\{i \in [N] \mid f(x_i, z_i) = 1 \text{ and } z_i = \ell\}$. In this case, $\mathsf{Pr}_{\widehat{S}}[f = 1 \mid Z = \ell] = 0$. This implies that $(f, \widehat{S}) = 0$. Thus, in this example, $\frac{(f,\widehat{S})}{(f,S)} = 0$.*

In Remark A.28 we give two example hypothesis classes that satisfy the assumptions in Theorems 4.4 and 4.5. Theorem 4.5 assumes that there exist five distinct points $x_A, x_B, x_C, x_D, x_E \in \mathcal{X}$ such that $\mathcal{F}$ shatters the set of points $P := \{x_A, x_B, x_C, x_D, x_E\} \times [2] \subseteq \mathcal{X}$. Theorem 4.4 makes the weaker assumption that $\mathcal{F}$ shatters the set $\{x_A, x_B, x_C\} \times [2] \subseteq \mathcal{X}$.

Remark A.28.

1. **Decision trees.** *Suppose $\mathcal{X} := \mathsf{R}$. Then, the following hypothesis class of two-layer decision trees shatters P: On the first layer, the decision tree splits the root node into five nodes by thresholding $X \in \mathcal{X}$. On the second layer, it further splits each node in the first layer into two leaves depending on whether $Z = 1$ or $Z = 2$. The resulting tree has 10 leaves. This hypothesis class shatters P: One can choose the thresholds in the first layer so that each of x_A, x_B, x_C, x_D, and x_E belong to different nodes on the first layer. Then, the 2^{10} hypothesis generated by assigning different outcomes to the leaves shatter P.*

2. **SVM with kernels.** *Suppose $\mathcal{X} := \mathsf{R}^5$, and $x_A = e_1$, $x_B = e_2$, $x_C = e_3$, $x_D = e_4$, $x_E = e_5$, where e_i is the i-th standard basis in R^5. Consider the hypothesis class of SVM classifiers on the feature space $\mathcal{X}^2$; where given a sample $(x, z) \in \mathcal{X} \times [p]$, we map it to $\mathcal{X}^2$ using the map $\psi\colon \mathcal{X} \times [p] \to \mathcal{X}^2$ defined as follows: $\psi(x, z) := (x, z \cdot x) \in \mathcal{X}^2$. One can verify that the hypothesis class of SVM classifiers in this feature space shatter the set P (more precisely, it shatters the image of P under map ψ).*

Remark A.29 (Test errors). *In the paper, we focus on the setting where given perturbed samples as input, the learner's goal is to find a classifier that satisfies a given set of fairness constraints with the highest accuracy, where both accuracy and fairness are measured with respect to* true distribution *$\mathcal{D}$. In some applications, for example, when protected attributes are self-reported after deployment, the test samples can also have perturbations. Given a number > 0, a sufficient number of test samples T, and an η-Hamming adversary A, one can show that $\mathsf{Err}_{\widehat{D}}(f_{\mathsf{ET}}) - \mathsf{Err}_D(f^\star) \le 3\eta +$ and $(f_{\mathsf{ET}}, A(T)) \ge \tau - 12\eta\tau/(\lambda \ 2\eta) -$, where $A(T)$ denotes the perturbed test samples and $\widehat{D}$ is the empirical distribution of $A(T)$.*

B Extensions of theoretical results

B.1 Theoretical results with multiple protected attributes and fairness metrics

In this section, we extend Program (ErrTolerant) to the general case with $m \in \mathsf{N}$ fairness constraints $^{(1)}$, $^{(2)}, \ldots,$ $^{(k)}$ with respect to m (not necessarily distinct) protected attributes $Z^{(1)}, Z^{(2)}, \ldots, Z^{(m)}$.

Definition B.1 (General Error-tolerant program). *Given a perturbation rate $\eta \in [0, 1]$, constants $\lambda, \ \in (0, 1]$, and for each $r \in [m]$, given a fairness constraint $^{(r)}$, corresponding events $\mathcal{E}^{(r)}$ and $(\mathcal{E}^0)^{(r)}$ (as in Definition 3.1), and protected attribute $Z^{(r)} \in [p_r]$, we define the general error-tolerant program for perturbed samples $\widehat{S}$, with empirical distribution is $\widehat{D}$, as*

$$\min_{f \in \mathcal{F}} \ \mathsf{Err}_{\widehat{D}}(f), \qquad \text{(General-ErrTolerant)} \quad (84)$$

$$\text{s.t.}, \quad \forall\, r \in [m], \qquad {}^{(r)}(f, \widehat{S}) \ge \tau \cdot \left(\frac{1 - (\eta + \ )/\lambda}{1 + (\eta + \ )/\lambda}\right)^2, \quad (85)$$

$$\forall\, r \in [m] \text{ and } \ell \in [p_r], \quad \mathsf{Pr}_{\widehat{D}}\left[\mathcal{E}^{(r)}(f), (\mathcal{E}^0)^{(r)}(f), \widehat{Z} = \ell\right] \ge \lambda - \eta - \quad . \quad (86)$$

Let $f^\star \in \mathcal{F}$ have the lowest error subject to satisfying all fairness constraints with respect to $\mathcal{D}$:

$$f^\star := \operatorname{argmin}_{f \in \mathcal{F}} \mathsf{Err}_D(f) \quad \text{s.t.}, \quad \text{for all } r \in [m], \quad {}^{(r)}_D(f) \ge \tau.$$

We need the following generalization of Assumption 1:

Assumption 2. *There is a known constant $\lambda > 0$ such that*

$$\min_{r \in [m]} \min_{\ell \in [p_r]} \mathsf{Pr}_D[\mathcal{E}^{(r)}(f^\star), (\mathcal{E}^0)^{(r)}(f^\star), Z = \ell] \ge \lambda.$$

Theorem B.2 (Extending Theorem 4.3 to multiple protected attributes and fairness constraints). *Suppose Assumption 2 holds with constant $\lambda > 0$ and $\mathcal{F}$ has VC dimension $d \in \mathsf{N}$. 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 $f_{\mathsf{ET}} \in \mathcal{F}$ of Program (General-ErrTolerant) with parameters η, λ, and $:= O(\varepsilon - 2\eta, \nu - 8\eta\tau/(\lambda \ 2\eta), \lambda - 2\eta)$, and $N = \mathrm{poly}(d, 1/\ , \log(\delta^{\ 1} \cdot \sum_{i \in [m]} p_i))$ perturbed samples from the η-Hamming model satisfies*

$$\mathsf{Err}_D(f_{\mathsf{ET}}) - \mathsf{Err}_D(f^\star) \le \varepsilon, \quad (87)$$

$$\text{for all } r \in [m], \quad {}^{(r)}_D(f_{\mathsf{ET}}) \ge \tau - \nu. \quad (88)$$

The proof of Theorem B.2 is similar to the proof of Theorem 4.3. Instead of repeating the entire proof, we highlight the differences between the two proofs.

Proof. Set

$$N := \left(\frac{1}{2 \cdot (\lambda - 2\eta)^4} \cdot \left(d \log \left(\frac{d}{2 \cdot (\lambda - 2\eta)^4} \right) + \log \left(\delta \ ^1 \cdot \sum_{r 2 [m]} p_r \right) \right) \right).$$

Fix any fairness constraint $r \in [m]$. Applying Lemma 4.8 to the r-th fairness constraint, we get that any $f \in \mathcal{F}$ feasible for Program (General-ErrTolerant) is $(\frac{1 \ (\eta + \ )/(\lambda \ 2\eta)}{1+(\eta + \ )/(\lambda \ 2\eta)})^2$-stable with respect to the fairness constraint $^{(r)}$. This satisfies the first condition in Lemma A.8. The second condition holds because any $f \in \mathcal{F}$ feasible for Program (General-ErrTolerant) satisfies the fairness constraint in Equation (85). This allows us to use Lemma A.8; we get that with probability at least $1 - \delta \cdot \frac{p_r}{\sum_i p_i}$, any $f \in \mathcal{F}$ feasible for Program (General-ErrTolerant) satisfies that

$$^{(r)}_{D}(f_{\mathsf{ET}}) \geq \tau - \nu.$$

Using the union bound over $r \in [m]$, implies that Equation (88) holds with probability at least $1 - \delta$.

If we can show that $f^\star$ is feasible for Program (General-ErrTolerant), then Lemma A.4 implies Equation (87). Using Assumption 2 and Lemma 4.6, it follows that $f^\star$ satisfies the lower bounds in Equation (86) with probability at least $1 - \frac{\delta}{\sum_i p_i}$. This, along with Lemma 4.8, implies that $f^\star$ is $(\frac{1 \ (\eta + \ )/\lambda}{1+(\eta + \ )/\lambda})^2$-stable for all fairness metrics. Then, using Lemma A.10, it follows that $f^\star$ satisfies the constraints for a particular $r \in [m]$ with probability at least $1 - \delta \cdot \frac{p_r}{\sum_i p_i}$. Taking the union bound over all $r \in [m]$, it follows that $f^\star$ is feasible for Program (General-ErrTolerant) with probability at least $1 - \delta$. □

Remark B.3. *Like Program (ErrTolerant), in general, Program (General-ErrTolerant) is also a nonconvex optimization program. But, for any arbitrarily small $\alpha > 0$, the techniques from [13] can be used to find an $f \in \mathcal{F}$ that has the optimal objective value for Program (General-ErrTolerant) and that additively violates its fairness constraint (85) by at most α by solving a set of $O((\lambda\alpha) \ ^m)$ convex programs (see Section C for details on an analogous argument for Program (ErrTolerant)).*

B.2 Theoretical results for Program (ErrTolerant+)

In this section, we show that Program (ErrTolerant+) offers a better fairness guarantee than Program (ErrTolerant) (up to a constant) if the classifiers in $\mathcal{F}$ do not use the protected attributes for prediction. In particular, we prove Theorem B.4.

Theorem B.4 (Guarantees for Program (ErrTolerant+)). *Suppose for each $\ell \in [p]$*

$$\lambda_\ell := \mathsf{Pr}_D[\mathcal{E}(f^\star), \mathcal{E}^0(f^\star), Z = \ell] \quad \textit{and} \quad \gamma_\ell := \mathsf{Pr}_D[\mathcal{E}^0(f^\star), Z = \ell],$$

$\mathcal{F}$ has VC dimension $d \in \mathsf{N}$, and classifiers in $\mathcal{F}$ do not use the protected attributes for prediction. Let s be the optimal value of Program (7) and $\lambda := \min_{\ell 2 [p]} \lambda_\ell$. 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

$$\nu > \tau \cdot \left(1 - s + \frac{4\eta}{(\lambda - 2\eta)} \right),$$

and confidence parameters $\delta \in (0, 1)$ with probability at least $1 - \delta$, the optimal solution $f_{\mathsf{ET}} \in \mathcal{F}$ of Program (ErrTolerant) with parameters η, λ,

$$:= O\left(\min \left\{ \varepsilon - 2\eta, \nu - \tau \cdot \left(1 - s + \frac{4\eta}{(\lambda - 2\eta)} \right), \lambda - 2\eta \right\} \right),$$

and $N := \mathrm{poly}(d, 1/\ , \log(p/\delta))$ perturbed samples from an η-Hamming adversary satisfies

$$\mathsf{Err}_D(f_{\mathsf{ET}}) - \mathsf{Err}_D(f^\star) \leq \varepsilon \quad \textit{and} \quad _D(f_{\mathsf{ET}}) \geq \tau - \nu. \tag{89}$$

B.2.1 Proof of Theorem B.4

The proof of Theorem B.4 is similar to the proof of Theorem 4.3. Instead of repeating the entire proof, we highlight the differences from the proof of Theorem 4.3.

Proof of Theorem B.4. The proof of Theorem 4.3 has three main steps:

1. Step 1 proves that if $f^\star$ is feasible for Program (ErrTolerant), then the f_{ET} has accuracy 2η-close to the accuracy of $f^\star$ (Lemma A.4).
2. Step 2 proves that any $f \in \mathcal{F}$ feasible for Program (ErrTolerant) satisfies the fairness guarantee in Theorem 4.3 with high probability (Lemma A.8). The main substep in the proof of Lemma A.8 is to show that any $f \in \mathcal{F}$ feasible for Program (ErrTolerant) is $(\frac{1 \quad (\eta + \quad)/\lambda}{1+(\eta + \quad)/\lambda})^2$-stable.
3. Step 3 proves that, with high probability, $f^\star$ is feasible for Program (ErrTolerant) (Lemma A.10). Combining this with Step 1 shows that f_{ET} satisfies the accuracy guarantee in Theorem 4.3 with high probability.

The proof of Step 1 (Lemma A.4) only depends on the perturbation model and the objective of Program (ErrTolerant); thus, it also generalizes to Program (ErrTolerant+). The proof of Step 2 (Lemma A.8), follows because any $f \in \mathcal{F}$ feasible for Program (ErrTolerant) satisfies the following inequality:

$$\mathsf{Pr}_{\widehat{D}}\left[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell\right] \geq \lambda - \eta - \quad . \tag{90}$$

Since for all $\ell \in [p]$, $\lambda_\ell \geq \lambda$, any $f \in \mathcal{F}$ feasible for Program (ErrTolerant+) also satisfies Equation (90). Thus, Lemma A.8 also holds for Program (ErrTolerant+). This shows that any $f \in \mathcal{F}$ feasible for Program (ErrTolerant+) satisfies

$$_D(f_{\mathsf{ET}}) \geq \tau - \frac{8\eta\tau}{\lambda - 2\eta}.$$

However, Theorem B.4 has a tighter fairness guarantee.[5] The tighter guarantee follows by lower bounding $(f, \widehat{S})$ by $\tau \cdot s$ in Equation (25) in the proof of Lemma A.8.

The main difference in the proofs of Theorem 4.3 and Theorem B.4 is in Step 3. Using Assumption 1 and Lemma A.3, we can show that $f^\star$ satisfies: $\mathsf{Pr}_{\widehat{D}}\left[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell\right] \geq \lambda_\ell - \eta - \quad$. It remains to show that $f^\star$ satisfies the fairness constraint: $(f, \widehat{S}) \geq \tau \cdot s$.

In the proof of Lemma A.10 we show a weaker result: $f^\star$ satisfies the fairness constraint: $(f, \widehat{S}) \geq \tau \cdot (\frac{1 \quad (\eta + \quad)/\lambda}{1+(\eta + \quad)/\lambda})^2$. This follows because $f^\star$ is $(\frac{1 \quad (\eta + \quad)/\lambda}{1+(\eta + \quad)/\lambda})^2$-stable. Instead, here, we prove that with high probability $f^\star$ satisfies the following inequality

$$\frac{(f^\star, \widehat{S})}{_D(f^\star)} \geq s. \tag{91}$$

(Equation (91) does not imply to s-stability because s-stability also requires an upper bound of $1/s$.) This suffices to show that $f^\star$ is feasible for Program (ErrTolerant+) with high probability; by our discussion so far, it also proves Theorem B.4. □

It remains to prove Equation (91). Formally, we prove Lemma B.5.

Lemma B.5. *Let s be the optimal value of Program* (7)*, then with probability at least $1 - \delta_0$*

$$\frac{(f^\star, \widehat{S})}{_D(f^\star)} \geq s, \tag{92}$$

where $f^\star$ is an optimal solution of Program (2) *and δ_0 is as defined in Equation* (9)*.*

[5] Because for all $\ell \in [p]$, $\lambda_\ell \quad \lambda$ and $\gamma_\ell \quad \lambda$, it can be shown that $s \quad (\frac{1 \quad (\eta + \quad)/\lambda}{1+(\eta + \quad)/\lambda})^2$. Thus, the guarantee in Theorem B.4 is stronger than the guarantee in Theorem 4.3.

Proof. Let the unperturbed samples be $S := \{(x_i, y_i, z_i)\}_{i \in [N]}$ and the perturbs samples be $\widehat{S} := \{(x_i, y_i, \widehat{z}_i)\}_{i \in [N]}$. We will prove that $(f, \widehat{S}) / (f, S) \geq s$. Then the result follows as give $N = \text{poly}(d, 1/, \log(p/\delta_0))$ iid samples from $\mathcal{D}$, it holds that with probability at lest $1 - \delta_0$, $(f, S) / D(f) \geq 1 -$.

Let $\mathcal{E}$ and $\mathcal{E}^0$ be the events defining the fairness metric (Definition 3.1). For each $\ell \in [p]$, we define

$$\eta_\ell^1 := \Pr_{\widehat{D}}[\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X)), \widehat{Z} = \ell] - \Pr_D[\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X)), Z = \ell],$$
$$\eta_\ell^2 := \Pr_{\widehat{D}}[\neg\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X)), \widehat{Z} = \ell] - \Pr_D[\neg\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X)), Z = \ell].$$

Substituting the values of λ_ℓ and γ_ℓ, we get that[6]

$$\eta_\ell^1 := \Pr_{\widehat{D}}[\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X)), \widehat{Z} = \ell] - \lambda_\ell, \tag{93}$$
$$\eta_\ell^2 := \Pr_{\widehat{D}}[\neg\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X)), \widehat{Z} = \ell] - (\gamma_\ell - \lambda_\ell)\,. \tag{94}$$

Intuitively, η_ℓ^1 is the number of samples with $\mathtt{I}[\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X))] = 1$ added to protected group $Z = \ell$ and η_ℓ^2 is the number of samples with $\mathtt{I}[\neg\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X))] = 1$ added to protected group $Z = \ell$. (Note that the event $\mathcal{E}(f^\star(X))$ and $\neg\mathcal{E}(f^\star(X))$ are disjoint.)

The values $\{\eta_\ell^1, \eta_\ell^2\}_\ell$ satisfy several conditions: Because the total number of samples added or removed from all protected groups is 0, it holds that $\sum_{\ell \in [p]} \eta_\ell^1 + \eta_\ell^2 = 0$. Moreover, because f does not use Z for prediction, perturbing Z does not change the value of $\mathtt{I}[\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X))]$, and hence, it holds that

$$\sum\nolimits_{\ell \in [p]} \eta_\ell^1 = 0 \quad \text{and} \quad \sum\nolimits_{\ell \in [p]} \eta_\ell^2 = 0. \tag{95}$$

Next, because any η-Hamming adversary perturbs at most η-fraction of the samples, we have that

$$\sum\nolimits_{\ell \in [p]} |\eta_\ell^1| + |\eta_\ell^2| \leq 2\eta. \tag{96}$$

Finally, as the probability in the RHS of Equation (94) is nonnegative, it follows that for all $\ell \in [p]$

$$\eta_\ell^2 \geq -(\gamma_\ell - \lambda_\ell)\,. \tag{97}$$

Now we are ready to prove the result

$$(f^\star, \widehat{S}) := \min_{\ell, k \in [p]} \frac{\Pr_{\widehat{D}}[\mathcal{E}^0(f^\star(X)), \widehat{Z} = \ell]}{\Pr_{\widehat{D}}[\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X)), \widehat{Z} = \ell]} \cdot \frac{\Pr_{\widehat{D}}[\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X)), \widehat{Z} = k]}{\Pr_{\widehat{D}}[\mathcal{E}^0(f^\star(X)), \widehat{Z} = k]}. \tag{98}$$

Fix any $\ell, k \in [p]$. From Equations (93) and (94) we have that

$$\frac{\Pr_{\widehat{D}}[\mathcal{E}^0(f^\star(X)), \widehat{Z} = \ell]}{\Pr_{\widehat{D}}[\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X)), \widehat{Z} = \ell]} \frac{\Pr_{\widehat{D}}[\mathcal{E}(f^\star(X)), \mathcal{E}^0(f^\star(X)), \widehat{Z} = k]}{\Pr_{\widehat{D}}[\mathcal{E}^0(f^\star(X)), \widehat{Z} = k]} = \frac{\lambda_\ell + \eta_\ell^1}{\gamma_\ell + \eta_\ell^1 + \eta_\ell^2} \cdot \frac{\gamma_k + \eta_k^1 + \eta_k^2}{\lambda_k + \eta_k^1}.$$

Our goal is to lower bound the LHS of the above equation because it implies a lower bound on $(f^\star, \widehat{S})$ by Equation (98). Towards this, given vectors $\eta^1, \eta^2 \in \mathbb{R}^p$, define the following objective:

$$\text{Obj}(\eta^1, \eta^2) := \frac{\lambda_\ell + \eta_\ell^1}{\gamma_\ell + \eta_\ell^1 + \eta_\ell^2} \cdot \frac{\gamma_k + \eta_k^1 + \eta_k^2}{\lambda_k + \eta_k^1}.$$

We would like to lower bound $\text{Obj}(\eta^1, \eta^2)$ subject to Equations (95) to (97), i.e., we would like to lower bound the value of the following program

$$\begin{aligned} & \text{Obj}(\eta^1, \eta^2), \\ \text{s.t.}, \quad & \text{eqs. (95), (96) and (97) hold} \end{aligned} \tag{99}$$

We claim that for all $i \notin \{\ell, k\}$ it is optimal to set $\eta_i^1 = 0$ in Program (99). To see this, note that given any solution $\eta^1, \eta^2 \in \mathbb{R}^p$ to Program (99), with $\alpha := \sum_{i \neq \ell, k} \eta_i^1$, we can construct another solution $\delta^1, \delta^2 \in \mathbb{R}^p$ that has a better objective while satisfying $\delta_i^1 = 0$ for all $i \notin \{\ell, k\}$: For all

[6] Here, we implicitly use the fact that $f \in \mathcal{F}$ does not use the protected attribute Z for prediction.

$i \notin \{\ell, k\}$, set $\delta_i^1 := 0$. Next, if $\alpha > 0$, set $\delta_k^1 := \eta_k^1 + \alpha$, and otherwise, set $\delta_\ell^1 := \eta_\ell^1 + \alpha$. Finally, let all other variables remain unchanged. (Under Equation (97), (δ^1, δ^2) has a smaller objective than (η^1, η^2).) Thus, we get that

$$\begin{aligned} &\min_{\eta^1,\eta^2 \in \mathbb{R}^p} \mathsf{Obj}(\eta^1, \eta^2), && = && \min_{\eta^1,\eta^2 \in \mathbb{R}^p} \mathsf{Obj}(\eta^1, \eta^2), \\ &\text{s.t.,} \quad \text{eqs. (95), (96) and (97) hold} && && \text{s.t.,} \quad \text{eqs. (95), (96) and (97) hold} \\ & && && \quad \text{for all } i \notin \{\ell, k\},\ \eta_i^1 = 0 \end{aligned} \tag{100}$$

Next, dropping Equation (97) from the constraint, which only improves the objective, we get that:

$$\begin{aligned} &\min_{\eta^1,\eta^2 \in \mathbb{R}^p} \mathsf{Obj}(\eta^1, \eta^2), && \geq && \min_{\eta^1,\eta^2 \in \mathbb{R}^p} \mathsf{Obj}(\eta^1, \eta^2), \\ &\text{s.t.,} \quad \text{eqs. (95), (96) and (97) hold} && && \text{s.t.,} \quad \text{eqs. (95) and (96) hold} \\ &\quad \text{for all } i \notin \{\ell, k\},\ \eta_i^1 = 0 && && \quad \text{for all } i \notin \{\ell, k\},\ \eta_i^1 = 0. \end{aligned} \tag{101}$$

We claim that for all $i \notin \{\ell, k\}$ it is optimal to set $\eta_i^2 := 0$ in the program in the RHS of Equation (101). This follows by a similar construction used to prove Equation (100). To see this, note that given any solution $\eta^1, \eta^2 \in \mathbb{R}^p$ to the program in the RHS of Equation (101), with $\alpha := \sum_{i \notin \{\ell,k\}} \eta_i^2$, there is another solution $\delta^1, \delta^2 \in \mathbb{R}^p$ that has a better objective value while satisfying $\delta_i^2 = 0$ for all $i \notin \{\ell, k\}$: For all $i \notin \{\ell, k\}$, set $\delta_i^1 := 0$. Next, if $\alpha > 0$, set $\delta_\ell^2 := \eta_\ell^2 + \alpha$, otherwise set $\delta_k^2 := \eta_k^2 + \alpha$. Finally, let all other variables remain unchanged. ((δ^1, δ^2) always has a smaller objective than (η^1, η^2).) Thus, we have

$$\begin{aligned} &\min_{\eta^1,\eta^2 \in \mathbb{R}^p} \mathsf{Obj}(\eta^1, \eta^2), && = && \min_{\eta^1,\eta^2 \in \mathbb{R}^p} \mathsf{Obj}(\eta^1, \eta^2), \\ &\text{s.t.,} \quad \text{eqs. (95) and (96) hold} && && \text{s.t.,} \quad \text{eqs. (95) and (96) hold} \\ &\quad \text{for all } i \notin \{\ell, k\},\ \eta_i^1 = 0 && && \quad \text{for all } i \notin \{\ell, k\},\ \eta_i^1 = 0,\ \eta_i^2 = 0. \end{aligned} \tag{102}$$

Rewriting the program in the RHS of Equation (102), by dropping the always 0 variables, we get

$$\begin{aligned} &\min_{\eta^1,\eta^2 \in \mathbb{R}^p} \mathsf{Obj}(\eta^1, \eta^2), && = && \min_{\eta_\ell^1,\eta_k^1,\eta_\ell^2,\eta_k^2 \in \mathbb{R}} \frac{\lambda_\ell + \eta_\ell^1}{\gamma_\ell + \eta_\ell^1 + \eta_\ell^2} \cdot \frac{\gamma_k + \eta_k^1 + \eta_k^2}{\lambda_k + \eta_k^1}, \\ &\text{s.t.,} \quad \text{eqs. (95) and (96) hold} && && \text{s.t.,} \quad \eta_\ell^1 = -\eta_k^1 \text{ and } \eta_\ell^2 = -\eta_k^2 \\ &\quad \text{for all } i \notin \{\ell, k\},\ \eta_i^1 = 0,\ \eta_i^2 = 0 && && \quad \left|\eta_\ell^1\right| + \left|\eta_\ell^2\right| \leq \eta \end{aligned} \tag{103}$$

Rewriting the program in the RHS of Equation (103), we get

$$\begin{aligned} &\min_{\eta_\ell^1,\eta_k^1,\eta_\ell^2,\eta_k^2 \in \mathbb{R}} \frac{\lambda_\ell + \eta_\ell^1}{\gamma_\ell + \eta_\ell^1 + \eta_\ell^2} \cdot \frac{\gamma_k + \eta_k^1 + \eta_k^2}{\lambda_k + \eta_k^1}, && = && \min_{\eta_\ell,\eta_k \in \mathbb{R}} \frac{\lambda_\ell - \eta_\ell}{\gamma_\ell - \eta_\ell + \eta_k} \cdot \frac{\gamma_k + \eta_\ell - \eta_k}{\lambda_k + \eta_\ell}, \\ &\text{s.t.,} \quad \eta_\ell^1 = -\eta_k^1 \text{ and } \eta_\ell^2 = -\eta_k^2, && && \text{s.t.,} \quad \left|\eta_\ell^1\right| + \left|\eta_\ell^2\right| \leq \eta \\ &\quad \left|\eta_\ell^1\right| + \left|\eta_\ell^2\right| \leq \eta && && \end{aligned} \tag{104}$$

In the program in the RHS of Equation (104), it is optimal to set $\eta_\ell, \eta_k \geq 0$; this simplifies the constraint $|\eta_\ell^1| + |\eta_\ell^2| \leq \eta$ to $\eta_\ell^1 + \eta_\ell^2 \leq \eta$. The sequence of equations, Equations (100) to (104), implies that

$$\begin{aligned} (f^\star, \widehat{S}) &\geq \min_{\ell,k \in [p]} \min_{\eta_\ell,\eta_k \geq 0} \frac{\lambda_\ell - \eta_\ell}{\gamma_\ell - \eta_\ell + \eta_k} \cdot \frac{\gamma_k + \eta_\ell - \eta_k}{\lambda_k + \eta_\ell}, \quad \text{s.t.,} \quad \eta_\ell^1 + \eta_\ell^2 \leq \eta \\ &= \min_{\ell,k \in [p]} \frac{\lambda_\ell \cdot \gamma_k}{\gamma_\ell \cdot \lambda_k} \cdot \min_{\eta_\ell,\eta_k \geq 0} \frac{1 - \eta_\ell/\lambda_\ell}{1 - (\eta_\ell - \eta_k)/\gamma_\ell} \cdot \frac{1 + (\eta_\ell - \eta_k)/\gamma_k}{1 + \eta_\ell/\lambda_k}, \ \text{s.t.,} \ \eta_\ell^1 + \eta_\ell^2 \leq \eta \\ &= (f^\star, S) \cdot \min_{\ell,k \in [p]} \min_{\eta_\ell,\eta_k \geq 0} \frac{1 - \eta_\ell/\lambda_\ell}{1 - (\eta_\ell - \eta_k)/\gamma_\ell} \cdot \frac{1 + (\eta_\ell - \eta_k)/\gamma_k}{1 + \eta_\ell/\lambda_k}, \ \text{s.t.,} \ \eta_\ell^1 + \eta_\ell^2 \leq \eta \\ &\qquad \text{(By the definition of } \lambda_\ell \text{ and } \gamma_\ell) \\ &= (f^\star, S) \cdot s. \qquad \text{(By the definition of } s) \end{aligned}$$

□

C Reduction from Program (ErrTolerant) to a set of convex programs

In general, Program (ErrTolerant) is a nonconvex optimization program. But, we can reduce Program (ErrTolerant) to a set of convex programs. Formally, for any arbitrarily small $\alpha > 0$, we can find an $f \in \mathcal{F}$ that has the optimal objective value for Program (ErrTolerant) and that additively violates its fairness constraint (Equation (4)) by at most α, by solving a set of $O(1/(\lambda\alpha))$ convex programs. In this section, we present this reduction. It largely follows from [13], but is included for completeness.

Recall that given a fairness metric and corresponding events $\mathcal{E}$ and $\mathcal{E}^0$ (as in Definition 3.1), perturbed samples $\widehat{S}$, whose empirical distribution is $\widehat{D}$, a perturbation rate $\eta \in [0, 1]$, and constants $\lambda, \in (0, 1]$, Program (ErrTolerant) is the following program:

$$\min_{f \in \mathcal{F}} \quad \mathrm{Err}_{\widehat{D}}(f), \tag{105}$$

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

$$\forall\, \ell \in [p],\ \Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell] \geq \lambda - \eta - \quad .$$

Equivalently defining scalars $\widehat{\tau} := \left(\frac{1 \quad (\eta + \quad)/\lambda}{1 + (\eta + \quad)/\lambda}\right)^2$ and $\widehat{\lambda} := \lambda - \eta - \quad$, our goal is to solve

$$\min_{f \in \mathcal{F}} \quad \mathrm{Err}_{\widehat{D}}(f), \tag{106}$$

$$\text{s.t.,} \quad (f, \widehat{S}) \geq \widehat{\tau}, \tag{107}$$

$$\forall\, \ell \in [p],\ \Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell] \geq \widehat{\lambda}. \tag{108}$$

Remark C.1. *All references to the results in [13] are to its arXiv version.*

Remark C.2. *In this section, all probabilities and expectations are with respect to the draw of perturbed samples $(X, Y, \widehat{Z})$. Given $\widehat{D}$, the empirical distribution over $\widehat{S}$, we use $\Pr_{\widehat{D}}[\cdot]$ to denote $\Pr_{(X,Y,\widehat{Z}) \quad \widehat{D}}[\cdot]$ and $\mathrm{E}_{\widehat{D}}[\cdot]$ to denote $\mathrm{E}_{(X,Y,\widehat{Z}) \quad \widehat{D}}[\cdot]$.*

C.1 Performance metrics in Definition 3.1 are a special case of the metrics in [13]

To use the results in [13], we need to show that Definition 3.1 is a special case of [13, Definition 2.3].

Lemma C.3. *Suppose $\mathcal{F} := \{0, 1\}^{X \quad [p]}$. For all events $\mathcal{E}$ and $\mathcal{E}^0$, that can depend on f, the corresponding metric q is a "performance function" as defined in [13, Definition 2.3].*

Proof. Observe that

$$q_\ell(f) := \Pr_{\widehat{D}}[\mathcal{E}(f) \mid \mathcal{E}^0(f), \widehat{Z} = \ell] = \frac{\Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f) \mid \widehat{Z} = \ell]}{\Pr_{\widehat{D}}[\mathcal{E}^0(f) \mid \widehat{Z} = \ell]}.$$

To simplify the notation, below, we use f to denote $f(X, \widehat{Z})$. We can rewrite the denominator as:

$$\begin{aligned}
\Pr_{\widehat{D}}[\mathcal{E}^0(f) \mid \widehat{Z} = \ell] &= \Pr_{\widehat{D}}[\mathcal{E}^0(f) \mid \widehat{Z} = \ell] \\
&= \Pr_{\widehat{D}}[\mathcal{E}^0(f) \mid f = 0, \widehat{Z} = \ell] \cdot \Pr_{\widehat{D}}[f = 0 \mid \widehat{Z} = \ell] \\
&\quad + \Pr_{\widehat{D}}[\mathcal{E}^0(f) \mid f = 1, \widehat{Z} = \ell] \cdot \Pr_{\widehat{D}}[f = 1 \mid \widehat{Z} = \ell] \\
&= c_0 \cdot \Pr_{\widehat{D}}[f = 0 \mid \widehat{Z} = \ell] + c_1 \cdot \Pr_{\widehat{D}}[f = 1 \mid \widehat{Z} = \ell],
\end{aligned}$$

where we defined

$$\begin{aligned}
c_0 &:= \Pr_{\widehat{D}}[\mathcal{E}^0(f) \mid f = 0, \widehat{Z} = \ell], \\
c_1 &:= \Pr_{\widehat{D}}[\mathcal{E}^0(f) \mid f = 1, \widehat{Z} = \ell].
\end{aligned}$$

Let $\alpha_0 := c_0$ and $\alpha_1 := c_1 - c_0$. Then, we have

$$\begin{aligned}
\Pr_{\widehat{D}}[\mathcal{E}^0(f) \mid \widehat{Z} = \ell] &= c_0 \cdot \left(1 - \Pr_{\widehat{D}}[f = 1 \mid \widehat{Z} = \ell]\right) + c_1 \cdot \Pr_{\widehat{D}}[f = 1 \mid \widehat{Z} = \ell] \\
&= \alpha_0 + \alpha_1 \cdot \Pr_{\widehat{D}}[f = 1 \mid \widehat{Z} = \ell] \qquad \text{(Using } \alpha_0 := c_0 \text{ and } \alpha_1 := c_1 - c_0\text{)} \\
&= \alpha_0 + \alpha_1 \cdot \Pr_{\widehat{D}}[f(X, \widehat{Z}) = 1 \mid \widehat{Z} = \ell]. \qquad (109)
\end{aligned}$$

Where the last equality follows due to our notation that the event $f = 1$ denotes $f(X, \widehat{Z}) = 1$ for random draws $(X, Y, \widehat{Z}) \sim \widehat{D}$. Next, by replacing $\mathcal{E}^0(f)$ by $\mathcal{E}(f) \wedge \mathcal{E}^0(f)$, we get that

$$\Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f) \mid \widehat{Z} = \ell] = \beta_0 + \beta_1 \cdot \Pr_{\widehat{D}}[f(X, \widehat{Z}) = 1 \mid \widehat{Z} = \ell]. \tag{110}$$

for some $0 \leq \beta_0 \leq 1$ and $-1 \leq \beta_1 \leq 1$. Comparing Equations (109) and (110) with [13, Definition 2.3], it follows that $q_\ell(f)$ is a special case of the performance functions in [13, Definition 2.3]. □

C.2 Reduction from Program (ErrTolerant) to a set of convex programs

Before stating the result, we need some additional notation.

Definition C.4. *Given a fairness metric , the corresponding performance metric q_ℓ (from Definition 3.1), desired fairness threshold $\tau \in (0,1]$, approximation parameter $\alpha \in (0,1]$, and lower and upper bounds $L, U \in [0,1]$, define the sets $K(\tau,\alpha), P(L,U) \subseteq \mathcal{F}$ as*

$$K(\tau,\alpha) := \left\{ f \in \mathcal{F} \colon \min_{\ell 2[p]} q_\ell(f) \geq \tau \cdot \max_{\ell 2[p]} q_\ell(f) - \alpha \right\},$$
$$P(L,U) := \left\{ f \in \mathcal{F} \colon \text{ for all } \ell \in [p],\ L \leq q_\ell(f) \leq U \right\}.$$

Note that $K(\tau,0)$ (i.e., setting $\alpha = 0$) is the set of classifiers that satisfy the fairness constraint $(f) \geq \tau$ exactly. $K(\tau,\alpha)$ ($\alpha > 0$) is the set of classifiers that satisfy a relaxation of this constraint. Formally, for any $\alpha > 0$, the set of classifiers α-feasible for Program (106) are all f in $K(\tau,\alpha)$ that also satisfy Equation (108). Under Assumption 1, any α-feasible classifier $f \in \mathcal{F}$ additively violates the fairness constraint in Program (106) by at most α/λ. To see this, suppose $f \in \mathcal{F}$ is α-stable, then

$$\begin{aligned}
\min_{\ell 2[p]} q_\ell(f) &\geq \tau \cdot \max_{\ell 2[p]} q_\ell(f) - \alpha \\
&= \left(\tau - \frac{\alpha}{\max_{\ell 2[p]} q_\ell(f)} \right) \cdot \max_{\ell 2[p]} q_\ell(f) \\
&\geq \left(\tau - \frac{\alpha}{\lambda} \right) \cdot \max_{\ell 2[p]} q_\ell(f), \qquad \text{(Using that } \max_{\ell 2[p]} q_\ell(f) \geq \lambda)
\end{aligned}$$

and hence,

$$\widehat{D}(f) = \frac{\min_{\ell 2[p]} q_\ell(f)}{\min_{\ell 2[p]} q_\ell(f)} \geq \tau - \frac{\alpha}{\lambda}. \tag{111}$$

Using the above notation, we can write Program (106) as follows:

$$\begin{aligned}
\min_{f 2 F} \quad & \mathrm{Err}_{\widehat{D}}(f), \\
\text{s.t.}, \quad & f \in K(\widehat{\tau}, 0), \\
& \forall\, \ell \in [p],\ \Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell] \geq \widehat{\lambda}.
\end{aligned}$$

Here, $K(\widehat{\tau},0)$ can be a nonconvex set. But as [13] show, it can be approximated as a union of convex sets. In particular, they approximate $K(\widehat{\tau},0)$ as the union $\bigcup_{j=1}^{J} P(L_j,U_j)$ for some $J \in \mathbb{N}$ and vectors $L, U \in [0,1]^J$. (One can prove that for all $L, U \in [0,1]$, $P(L,U)$ is a convex set [13].)

Theorem C.5 (**Implicit in [13, Theorem 3.1]**). *Given constants $\tau, \alpha \in (0,1]$, let $J := \lceil \tau/\alpha \rceil$, and for all $j \in [J]$, let $L_j := (j-1)\alpha$ and $U_j := (j\alpha)/\tau$. For all fairness metrics and corresponding performance metric q_ℓ (as defined in Definition 3.1) it holds that*

$$K(\tau,0) \subseteq \bigcup_{j=1}^{J} P(L_j,U_j) \subseteq K(\tau,\alpha).$$

Theorem C.5 allows one to reduce the problem of finding an α-feasible classifier to solving a set of $\lceil \tau/\alpha \rceil$ convex programs of the following form: For some $L, U \in [0,1]$

$$\begin{aligned}
\min_{f 2 F} \quad & \mathrm{Err}_{\widehat{D}}(f), && (112) \\
\text{s.t.}, \quad & f \in P(L,U), && (113) \\
& \forall\, \ell \in [p],\ \Pr_{\widehat{D}}[\mathcal{E}(f), \mathcal{E}^0(f), \widehat{Z} = \ell] \geq \widehat{\lambda}. && (114)
\end{aligned}$$

Theorem C.6. *Given constants $\tau, \alpha \in (0,1]$, let $J := \lceil \tau/\alpha \rceil$, and for all $j \in [J]$, let $L_j := (j-1)\alpha$ and $U_j := (j\alpha)/\tau$. Further, let f_j be the optimal solution of Equation* (112) *with $L := L_j$ and $U := U_j$. Then, $f_\alpha := \mathrm{argmin}_{f_j} \mathrm{Err}(f_j, \widehat{S})$ has the optimal accuracy for Program* (106) *and is α-feasible for Program* (106)*.*

If Assumption 1 holds, then using Theorem C.6 and Equation (111), we can to find an $f \in \mathcal{F}$ that has the optimal objective value for Program (ErrTolerant) and that additively violates its fairness constraint (4) by at most α by solving a set of $\lceil \widehat{\tau}/(\lambda\alpha) \rceil$ convex programs.

Proof of Theorem C.6.

A. Fairness guarantee: Since f_α is an optimal solution of Program (112) with $L := L_j$ and $U := U_j$ for some $r \in [J]$, $f_\alpha \in P(L_r, U_r)$, and hence, $f \in \bigcup_{j=1}^{J} P(L_j, U_j)$. Therefore, by Theorem C.5, $f_\alpha \in K(\tau, \alpha)$. Since $f_\alpha \in K(\tau, \alpha)$ and f_α satisfies Equation (114), f_α is an α-feasible solution for Program (106).

B. Accuracy guarantee: Let $f_{\mathsf{ET}} \in \mathcal{F}$ be the optimal solution of Program (106). Since $f_{\mathsf{ET}} \in K(\tau, 0)$ and by Theorem C.5 the inclusion $\bigcup_{j=1}^{J} P(L_j, U_j) \supseteq K(\tau, 0)$ holds, $f_{\mathsf{ET}} \in \bigcup_{j=1}^{J} P(L_j, U_j)$. Further, since f_α minimizes $\mathsf{Err}(\cdot, \widehat{S})$ over $\bigcup_{j=1}^{J} P(L_j, U_j)$ and by Theorem C.5 the containment $\bigcup_{j=1}^{J} P(L_j, U_j) \supseteq K(\tau, 0)$ holds, it follows that $\mathsf{Err}(f_\alpha, \widehat{S}) \leq \mathsf{Err}(f_{\mathsf{ET}}, \widehat{S})$. □

D Further comparison to related work

D.1 Other related work

Fair classification without perturbations. A large body of work studies fair classification. Here, several works frame fair classification as a constrained optimization program and develop algorithms to solve these programs [65, 63, 62, 48, 27, 1, 13]. A different approach is to alter the decision boundary of a given classifier to improve its fairness [26, 32, 28, 51, 61, 24] (possibly with different alterations for different protected groups). Furthermore, some works preprocesses the training data to "correct" for its bias [37, 47, 38, 64, 25, 42]. However, these works require the protected attributes in the training samples to be known exactly, whereas in this paper we study the setting where fraction of the protected attributes are arbitrarily corrupted.

Missing protected attributes. Some works have studied fair classification in the absence of protected attributes–using auxiliary data. For example, [30] use other variables as proxies for protected attributes and [19] augment their dataset with "related data" (that includes protect attributes) to control fairness. In the absence of auxiliary data, [33] use distributionally robust optimization to minimize the maximum empirical risk across the protected groups, and [43] use a neural network to identify "potential" protected groups. However, these approaches do not offer provable guarantees on accuracy (with respect to $f^\star$). In contrast, our approach uses perturbed protected attributes and comes with provable guarantees on fairness and accuracy (with respect to $f^\star$).

Stochastic perturbations in labels. [10, 58] study fair classification with perturbations in the labels: [10] consider a model where perturbations arise due to bias in the training samples. They show that, under some models of bias, adding fairness constraints can improve the accuracy of the classifier on the unbiased data. [58] consider a model where the labels in each protected group are perturbed to a different value independently with a known (group-dependent) probability; they give a framework for a non-binary protected attribute that provably outputs a classifier with near-optimal accuracy that nearly satisfies the fair constraint with respect to equalized odds, true-positive rate, or false-positive rate fairness constraints. In contrast, we focus on adversarial perturbations in the features, and our framework can be extended to adversarial perturbations in both features and labels (see Section A.1.5). Finally, our framework works for a large class of linear-fractional fairness metrics (which include true-positive rate and false-positive rate fairness constraints, and can ensure equalized odds fairness).

D.2 Performance of prior frameworks under the η-Hamming model

In this section, we present examples showing that prior frameworks for fair classification can have low accuracy and fairness compared to our framework under the η-Hamming model.

D.2.1 [11]'s framework can output classifiers with low statistical rate

In this section, for any $\delta \in (0, 1/4)$, we give an example (Example D.1) where with high probability [11]'s framework outputs a classifier f_{OPT} that has perfect accuracy and 0 statistical rate. On the same example, an optimal solution f_{ET} of Program (ErrTolerant) has accuracy $1 - \delta$ and perfect statistical rate 1.

Example D.1. *Fix $\mathcal{X}$ to be any set with at least two distinct points, say x_A and x_B. Let $\mathcal{F}$ be any hypothesis class that shatters the set $\{x_A, x_B\} \times [2] \subseteq \mathcal{X} \times [p]$. Define the distribution $\mathcal{D}$ as follows*

$$\Pr_D[X = x, Y = y, Z = z] := \begin{cases} 1/3 - \delta/3 & \text{if } x = x_A, y = 1, z = 1, \\ 1/3 - \delta/3 & \text{if } x = x_B, y = 0, z = 1, \\ \delta & \text{if } x = x_A, y = 0, z = 2, \\ 1/3 - \delta/3 & \text{if } x = x_B, y = 0, z = 2, \\ 0 & \text{otherwise}, \end{cases}$$

where δ is some constant smaller than $1/4$. Note that for a sample $(X, Y, Z) \sim \mathcal{D}$, conditioned on $X = x$ and $Z = z$, Y takes the value $\mathbb{I}[x = x_A, z = 1]$. Thus, the classifier $f_{\mathsf{OPT}}(x, z) := \mathbb{I}[x = x_A, z = 1]$ has 0 predictive error. One can verify that f_{OPT} has a statistical rate of 0 with respect to $\mathcal{D}$. Since $\mathcal{F}$ shatters $\{x_A, x_B\} \times [2]$, $\mathcal{F}$ contains $f_{\mathsf{OPT}}(x, z)$. Further, any other classifier in $\mathcal{F}$ has an error at least δ with respect to $\mathcal{D}$.

[11]'s framework outputs the classifier with the minimum empirical risk on the given samples. Suppose the perturbation rate is $\eta := 0$. Then, given a sufficient number of samples from $\mathcal{D}$, with high probability, $f_{\mathsf{OPT}} \in \mathcal{F}$ has the minimum empirical error, and hence, is output by [11]'s framework; f_{OPT} satisfies

$$\mathsf{Err}_D(f_{\mathsf{OPT}}) = 0 \quad \text{and} \quad \_{}_D(f_{\mathsf{OPT}}) = 0,$$

where _ is the statistical rate fairness metric.

Next, we show that on this example, Program (ErrTolerant) outputs a classifier with a large statistical rate. Set the fairness threshold to be any value $\tau < 1$. Fix any $\lambda \leq \delta$ (this ensures that Assumption 1 is satisfied). Fix any _ > 0. Finally, as mentioned, $\eta := 0$.

One can verify $f_{\mathsf{ET}}(x, z) := \mathbb{I}[x = x_A]$ has error $\mathsf{Err}_D(f_{\mathsf{ET}}) = \delta$ and a statistical rate of 1 with respect to $\mathcal{D}$. In contrast, any other classifier with statistical rate at least $1 - 2\delta$ has error at least $1/3 - \delta/3 > \delta$ with respect to $\mathcal{D}$. (O1) Using this, one can show that, given a sufficient number of iid samples from $\mathcal{D}$, with high probability, any other classifier feasible for Equation (4) in Program (ErrTolerant) has an error larger than the error of f_{ET} (on the given samples). (O2) Further, because $\lambda \leq \delta$, one can verify that given a sufficient number of iid samples from $\mathcal{D}$, with high probability, f_{ET} satisfies Equation (5) in Program (ErrTolerant); thus, with high probability, f_{ET} is feasible for Program (ErrTolerant).

Combining observations (O1) and (O2), we get that: Given a sufficient number of iid samples from $\mathcal{D}$, with high probability, f_{ET} is the optimal solution Program (ErrTolerant) with parameters $\tau = 1 - \delta$, $\lambda = \delta$, $\eta = 0$, and _ > 0; f_{ET} satisfies

$$\mathsf{Err}_D(f_{\mathsf{ET}}) = \delta \quad \text{and} \quad \_{}_D(f_{\mathsf{ET}}) = 1.$$

D.2.2 [44]'s and [14]'s frameworks can output classifiers with low accuracy

In this section, for any $\eta \in (0, 1/2)$, we give an example where with high probability [44]'s and [14]'s frameworks output classifiers f_L and f_C (respectively) whose error is at least $1/4$ higher than the error of $f^\star$ (Theorem D.3); where $f^\star$ is an optimal solution to Program (2). On the same example, an optimal solution of Program (ErrTolerant) has error within 2η of the error of $f^\star$ and violates the fairness constraint by at most $O(\eta)$.

[44] and [14] 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. [14] also takes the constant λ from Assumption 1 as input. In addition, both [44] and [14] require group specific perturbation rates as input: for each pair $\ell, k \in [p]$, they require $P\ell k := \Pr_D[\widehat{Z} = k \mid Z = \ell]$.

Let $P \in [0, 1]^{p \ p}$ denote the resulting matrix. To give a meaningful estimate of P with adversarial noise, we define the following restriction of the Hamming adversary.

Definition D.2 (P**-restricted Hamming adversary).** *Given a matrix $P \in [0, 1]^{p \ p}$ and $N \in \mathbb{N}$ samples $\{(x_i, y_i, z_i)\}_{i2[N]}$, for each $\ell \in [p]$, let $G_\ell := \{i \in [N] \mid z_i = \ell\}$ be the set of samples with protected attribute ℓ. For each $\ell, k \in [p]$, the P-restricted Hamming adversary A_{RH} chooses $P\ell k \cdot |G_\ell|$ samples $i \in [N]$ from G_ℓ, and perturbs their protected attribute z_i from ℓ to $\widehat{z}_i = k$.*[7]

[7] We assume that $P\ell k \ \ N_\ell$ is integral for all ℓ, k 2 $[p]$. This can be ensured by slightly increasing $P\ell k$ or N.

The modifier "P-restricted" refers to the restriction placed by the matrix P on the adversary. Let $\mathcal{A}_{RH}(P)$ be the set of all P-restricted Hamming adversaries. Then one can show that $\mathcal{A}_{RH}(P) \subseteq \mathcal{A}(\eta)$ for any $\eta \geq \max_{\ell \in [p]} \sum_{k \in [p]} P\ell k$.

Theorem D.3. *Suppose that there are two protected groups ($p := 2$) and $\mathcal{X}$ contains at least two distinct points. Then, there is a family of hypothesis classes $\mathcal{F}$ such that for all fairness thresholds $\tau \in (0, 1]$ and perturbation rates $\eta \in (0, 1/2)$, there is*

1. *a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\} \times [2]$ that satisfies Assumption 1 with $\lambda := \tau/4$,*
2. *a matrix $P \in [0, 1]^{2 \times 2}$ such that $\mathcal{A}_{RH}(P) \subseteq \mathcal{A}(\eta)$, and*
3. *an adversary $A_{RH} \in \mathcal{A}_{RH}(P)$ that perturbs at most η-fraction of the samples,*

such that, if the fairness metric is statistical rate, then for a draw of $N \in \mathbb{N}$ iid samples S from $\mathcal{D}$, with probability at least $1 - e^{-(\eta^2 \tau^2 N)}$ (over the draw of S), it holds that the optimal classifiers

1. *$f_C \in \mathcal{F}$ of [14]'s program with parameters P, λ, and τ and samples $A_{RH}(S)$,*
2. *$f_L \in \mathcal{F}$ of [44]'s program with parameters P and $\delta_L := 1/2 - \tau/2$ and samples $A_{RH}(S)$, and*[8]
3. *$f_{\mathsf{ET}} \in \mathcal{F}$ of Program (ErrTolerant) with parameters η, λ, and τ and samples $A_{RH}(S)$*

have errors

$$\mathsf{Err}_D(f_C) - \mathsf{Err}_D(f^\star) \geq \frac{1}{4}, \tag{115}$$

$$\mathsf{Err}_D(f_L) - \mathsf{Err}_D(f^\star) \geq \frac{1}{4}, \tag{116}$$

$$\mathsf{Err}_D(f_{\mathsf{ET}}) - \mathsf{Err}_D(f^\star) \leq 2\eta. \tag{117}$$

Further, f_{ET} has statistical rates at least $\tau - O(\eta/\tau)$ with respect to $\mathcal{D}$, i.e., $\gamma_D(f_{\mathsf{ET}}) \geq \tau - O(\eta/\tau)$.

Proof for Theorem D.3. Let $S := \{(x_i, y_i, z_i)\}_{i \in [N]}$ denote N iid samples from $\mathcal{D}$.

Setting P, A, and $\mathcal{D}$. We let $P := \begin{bmatrix} 1 & \eta_1 \\ \eta_2 & 1 \end{bmatrix}^{\eta_1}_{\eta_2}$, where $\eta_1 := 0$ and $\eta_2 := \eta$. Since $\eta = \max_{\ell \in [p]} \sum_{k \in [p]} P\ell k$, we can verify that $\mathcal{A}_{\mathsf{RH}}(P) \subseteq \mathcal{A}(\eta)$. We fix $A \in \mathcal{A}_{\mathsf{RH}}(P)$ to be the following algorithm.

Input. A perturbation rate $\eta > 0$, matrix $P := \begin{bmatrix} 1 & \eta_1 \\ \eta_2 & 1 \end{bmatrix}^{\eta_1}_{\eta_2}$, where $\eta_1, \eta_2 \in [0, 1]$, and samples $S := \{(x_i, y_i, z_i)\}_{i \in [N]}$

Output. Samples $\widehat{S}$

---

1. **For** $\ell \in [2]$ **do:**
   (a) **Set** $N_\ell := \eta_\ell \cdot \sum_{i \in [N]} \mathbb{I}[z_i = \ell]$
   (b) **Set** $G_A := \{i \in [N] : z_i = \ell, x_i = x_A\}$ and $G_B := \{i \in [N] : z_i = \ell, x_i = x_B\}$
   (c) **Initialize** $C = \emptyset$ *// Corrupted samples*
   (d) Pick any $\min\{N_\ell, |G_B|\}$ items from G_B and add them to C
   (e) Pick any $N_\ell - \min\{N_\ell, |R_B|\}$ items from G_A and add them to C
   (f) **For** $i \in C$ **do: Set** $\widehat{z}_i = 3 - \ell$ *// If $z_i = 1$ the $\widehat{z}_i = 2$, and if $z_i = 2$ then $\widehat{z}_i = 1$*
   (g) **For** $i \in (G_A \cup G_B) \backslash C$ **do: Set** $\widehat{z}_i = \ell$
2. **return** $\widehat{S} := \{(x_i, y_i, \widehat{z}_i)\}_{i \in [N]}$

One can verify that A perturbs exactly $P\ell k \cdot |G_\ell|$ samples with protected attribute ℓ to protected attribute k. Hence, A is a P-restricted Hamming adversary. Further, as $\eta_1 + \eta_2 = \eta$, it also follows that A perturbs at most η-fraction of samples, and hence, is an η-Hamming adversary.

[8] In this example, $(1/2) \quad (\tau/2)$ is the minimum value of δ_L needed to ensure that $f^\star$, an optimal solution of Program (2), is feasible for [44]'s program with $\eta = 0$.

Fix $\mathcal{X}$ to be any set with at least two distinct points, say x_A and x_B. Let $\mathcal{F}$ be any hypothesis class that shatters the set $\{x_A, x_B\} \times [2]$. We define the distribution $\mathcal{D}$ as follows

$$\Pr_D[X = x, Y = y, Z = z] := \begin{cases} \tau/4 & \text{if } x = x_A, y = 1, z = 1, \\ 1/4 & \text{if } x = x_A, y = 1, z = 2, \\ 1/2 - \tau/4 & \text{if } x = x_B, y = 0, z = 1, \\ 1/4 & \text{if } x = x_B, y = 0, z = 2, \\ 0 & \text{otherwise.} \end{cases} \tag{118}$$

Note that for a sample $(X, Y, Z) \sim \mathcal{D}$, conditioned on X, the value of Y is $\mathtt{I}[X = x_A]$. Thus, the classifier $f^\star(x, z) := \mathtt{I}[X = x_A]$ has 0 predictive error.

We use Lemma D.4 in the proof of Theorem D.3.

Lemma D.4 (**Estimates of statistic on perturbed samples).** *For all $\delta \in (0, 1)$, with probability at least $1 - e^{(\delta^2 N)}$ (over the draw of S), the following bounds hold*

$$\left|\Pr_{\widehat{D}}[f^\star = 1, Z = 1] - \Pr_D[f^\star = 1, Z = 1]\right| \le \delta, \tag{119}$$

$$\left|\Pr_{\widehat{D}}[f^\star = 1, Z = 2] - \Pr_D[f^\star = 1, Z = 2]\right| \le \delta, \tag{120}$$

$$\left|\Pr_{\widehat{D}}[Z = 1] - (\Pr_D[Z = 1] + \eta \cdot \Pr_D[Z = 2])\right| \le \delta, \tag{121}$$

$$\left|\Pr_{\widehat{D}}[Z = 2] - \Pr_D[Z = 2] \cdot (1 - \eta)\right| \le \delta. \tag{122}$$

Equivalently substituting the statistics on $\mathcal{D}$ in Equations (119) *to* (122)*, we get*

$$\left|\Pr_{\widehat{D}}[f^\star = 1, Z = 1] - \frac{\tau}{4}\right| \le \delta, \tag{123}$$

$$\left|\Pr_{\widehat{D}}[f^\star = 1, Z = 2] - \frac{1}{4}\right| \le \delta, \tag{124}$$

$$\left|\Pr_{\widehat{D}}[Z = 1] - \frac{1 + \eta}{2}\right| \le \delta, \tag{125}$$

$$\left|\Pr_{\widehat{D}}[Z = 2] - \frac{1 - \eta}{2}\right| \le \delta. \tag{126}$$

The proof of Lemma D.4 follows by analyzing the algorithm of A and using the Chernoff bound. The proof of Lemma D.4 appears at the end of this section.

Proof of Theorem D.3. Since the distribution $\mathcal{D}$ is supported on 4 points, namely $\{x_A, x_B\} \times [2]$ (see Equation (118)), we only need to consider hypothesis in the restriction of $\mathcal{F}$ on the set $\{x_A, x_B\} \times [2]$; this restriction has 2^4 hypothesis.

The first observation is that $\mathtt{I}[X = x_A]$ is an optimal solution for Program (2) and satisfies. $\mathrm{Err}_D(f^\star) = 0$ and $_D(f^\star) = 1$, where is the statistical rate fairness metric. This is because, $\mathtt{I}[X = x_A]$ satisfies the constraints of Program (2) and has perfect accuracy.

The second observation is that $f^\star$ is not feasible for [14] and [44]'s programs.

$f^\star$ is not feasible for [14]'s program. [14] express their constraints (for statistical rate) in terms of vectors $u(f), w \in [0, 1]^2$ (where $u(f)$ depends on $f \in \mathcal{F}$). They define $u(f)$ and w as follows

$$u(f) := (P^T)^{\ 1} \cdot \begin{bmatrix} \Pr_{\widehat{D}}[f = 1, Z = 1] \\ \Pr_{\widehat{D}}[f = 1, Z = 2] \end{bmatrix}, \tag{127}$$

$$w := (P^T)^{\ 1} \cdot \begin{bmatrix} \Pr_{\widehat{D}}[Z = 1] \\ \Pr_{\widehat{D}}[Z = 2] \end{bmatrix}. \tag{128}$$

[14] impose the following constraint

$$\frac{\min_{\ell 2 [p]} u(f)_\ell / w_\ell}{\max_{\ell 2 [p]} u(f)_\ell / w_\ell} \ge \tau. \tag{129}$$

In our example, $(P^T)^{-1} := \frac{1}{1-\eta} \cdot \begin{bmatrix} 1 & \eta \\ 0 & 1 \end{bmatrix}$. Set

$$\delta := \frac{\eta \cdot \tau}{64}.$$

Substituting the value of $(P^T)^{-1}$ in Equations (127) and (128), and then using Lemma D.4, we get that with probability at least $1 - e^{-(\delta^2 N)}$, $u(f^\star)$ and w satisfy the following bounds

$$\left| u(f^\star)_1 - \left(\frac{\tau}{4} - \frac{\eta}{4(1-\eta)} \right) \right| \leq \frac{\delta}{1-\eta}, \tag{130}$$

$$\left| u(f^\star)_2 - \frac{1}{4(1-\eta)} \right| \leq \delta, \tag{131}$$

$$\left| w_1 - \frac{1}{2} \right| \leq \frac{\delta}{1-\eta}, \tag{132}$$

$$\left| w_2 - \frac{1}{2} \right| \leq \delta. \tag{133}$$

Suppose the Equations (130) to (133) hold. Toward computing the constraint in Equation (129), we compute bounds for $u(f^\star)_1/w_1$ and $u(f^\star)_2/w_2$.

$$\begin{aligned}
\frac{u(f^\star)_1}{w_1} &\leq \frac{\frac{\tau}{4} - \frac{\eta}{4(1-\eta)} - \frac{\delta}{1-\eta}}{\frac{1}{2} + \frac{\delta}{1-\eta}} && \text{(Using Equations (130) and (132))} \\
&\leq \frac{\frac{\tau}{4} - \frac{\eta}{8(1-\eta)}}{\frac{1}{2} \cdot (1 - \frac{2\delta}{1-\eta})} && \text{(Using that } \delta \leq \eta/8\text{)} \\
&= \frac{\tau}{2} \cdot \frac{1 - \frac{\eta}{2\tau(1-\eta)}}{1 - \frac{2\delta}{1-\eta}} \\
&< \frac{\tau}{2}, && \text{(Using that } \delta \leq \eta/(4\tau)\text{)} \quad (134) \\
\frac{u(f^\star)_2}{w_2} &\geq \frac{\frac{1}{4(1-\eta)} - \delta}{\frac{1}{2} + \delta} && \text{(Using Equations (131) and (133))} \\
&= \frac{1}{2(1-\eta)} \cdot \frac{1 - 4\delta(1-\eta)}{1 + 2\delta} \\
&\geq \frac{1}{2(1-\eta)} \cdot (1 - 4\delta(1-\eta)) \cdot (1 - 2\delta) && \text{(Using that for all } x \in \mathsf{R}, \tfrac{1}{1+x} \geq 1 - x.\text{)} \\
&\geq \frac{1}{2(1-\eta)} \cdot (1 - 6\delta) && \text{(Using that } \delta, \eta > 0\text{)} \\
&> \frac{1}{2}. && \text{(Using that } \delta < \eta/6\text{)} \quad (135)
\end{aligned}$$

Substituting Equations (134) and (135) in Equation (129), we get that

$$\frac{\min_{\ell \in [p]} u(f)_\ell / w_\ell}{\max_{\ell \in [p]} u(f)_\ell / w_\ell} \leq \frac{u(f)_1/w_1}{u(f)_2/w_2} \overset{(134),(135)}{<} \tau.$$

Thus, $f^\star$ is not feasible for [14]'s optimization program.

$f^\star$ is not feasible for [44]'s program. For any $f \in \mathcal{F}$, [44] impose the constraint

$$\left| \Pr_{\widehat{D}}[f = 1 \mid Z = 1] - \Pr_{\widehat{D}}[f = 1 \mid Z = 2] \right| \leq \delta_L \cdot (1 - \alpha - \beta), \tag{136}$$

where $\alpha, \beta \in [0, 1]$ are some function of η_1 and η_2. In particular, it holds that if $\eta_1 > 0$ (respectively $\eta_2 > 0$) then $\alpha > 0$ (respectively $\beta > 0$), otherwise $\alpha = 0$ (respectively $\beta = 0$) In our example, $\eta_1 = 0$ and $\eta_2 = \eta > 0$. Thus, $\alpha = 0$ and $\beta > 0$. Recall that $\delta_L := \frac{1}{2} - \frac{\tau}{2}$. To show that $f^\star$ does not

satisfy Equation (136), we bound $\Pr_{\widehat{D}}[f = 1 \mid Z = 1]$ and $\Pr_{\widehat{D}}[f = 1 \mid Z = 2]$.

$$\begin{aligned}
\Pr_{\widehat{D}}[f = 1 \mid Z = 1] &\leq \frac{\frac{\tau}{4} + \delta}{\frac{1+\eta}{2} - \delta} && \text{(Using Equations (123) and (125))} \\
&= \frac{\tau}{2(1+\eta)} \cdot \frac{1 + \frac{4\delta}{\tau}}{1 - \frac{2\delta}{1+\eta}} \\
&\leq \frac{\tau}{2(1+\eta)} \cdot (1 + \frac{\eta}{16}) \cdot \left(1 + \frac{4\delta}{1+\eta}\right) \\
& && \text{(Using that } \delta := \tfrac{\eta\tau}{64} \text{ and } \tfrac{4\delta}{1+\eta} \in [0, 1/2]) \\
&\leq \frac{\tau}{2(1+\eta)} \cdot \left(1 + \frac{\eta}{8}\right)^2 && \text{(Using that } \delta \leq \tfrac{\eta}{4}) \\
&< \frac{\tau}{2}, && \text{(Using that } \eta \leq 1) \quad (137) \\
\Pr_{\widehat{D}}[f = 1 \mid Z = 2] &\geq \frac{\frac{1}{4} - \delta}{\frac{1}{2} - \frac{\eta}{2} + \delta} && \text{(Using Equations (124) and (126))} \\
&= \frac{1}{2(1-\eta)} \cdot \frac{1 - 4\delta}{1 + \frac{2\delta}{1+\eta}} \\
&\geq \frac{1}{2(1-\eta)} \cdot \left(1 - \frac{\eta}{8}\right) \cdot \left(1 - \frac{2\delta}{1+\eta}\right) \\
& \text{(Using that } 4\delta \leq \eta/8 \text{ and for all } x \in \mathbb{R}, (1+x)^{-1} \geq 1 - x) \\
&\geq \frac{1}{2(1-\eta)} \cdot \left(1 - \frac{\eta}{8}\right)^2 && \text{(Using that } 2\delta \leq \eta/8) \\
&\geq \frac{1 - \frac{\eta}{4}}{2(1-\eta)} \\
&> \frac{1}{2}. && \text{(Using that } \eta > 0) \quad (138)
\end{aligned}$$

Thus, combining Equations (137) and (138) and the fact that $\beta > 0$ and $\alpha = 0$, it follows that $f^\star$ is not feasible for Equation (136).

[44]'s and [14]'s frameworks output a classifier with large error. Since $f^\star$ is not feasible for [44]'s and [14]'s programs, they must output some other classifier $f_{\mathsf{Alt}} \in \mathcal{F}$. Toward a contradiction, suppose that $\mathsf{Err}_D(f_{\mathsf{Alt}}) < 1/4$. Consider the set $U := \{x_A, x_B\} \times [2] \setminus \{(x_A, 1)\}$. Each point in the U has probability mass at least $1/4$. Thus, if f_{Alt} has different outcome than $f^\star$ on the set U, then $\mathsf{Err}_D(f_{\mathsf{Alt}}) \geq 1/4$. So we must have $f_{\mathsf{Alt}}(r, s) = f^\star(r, s)$ for all $(r, s) \in U$. Because f_{Alt} is different than $f^\star$, its outcome must differ from $f^\star$ on at least one point in the support of $\mathcal{D}$. The only remaining point is $(x_A, 1)$. Thus, we must have $f_{\mathsf{Alt}}(x_A, 1) = 0$. However, in this case, one can show that $\Pr_{\widehat{D}}[f_{\mathsf{Alt}} = 1, Z = 1] = 0$ and $\Pr_{\widehat{D}}[f_{\mathsf{Alt}} = 1, Z = 1] > 0$. Substituting this in Equations (129) and (136) we get, that f_{Alt} is not feasible for [14]'s and [44]'s optimization programs. This is a contradiction since we assumed that [14] and [44] output f_{Alt}. This proves Equations (115) and (116).

Finally, Equations (117) and the bound on $\mathsf{Err}_D(f_{ET})$ follows from Theorem 4.3 because $f^\star$ satisfies Assumption 1 with constant $\lambda = \tau/4$. □

Remark D.5 (Choice of P). *In Theorem D.3, we fix $P := \begin{bmatrix} 1 & 0 \\ \eta & 1-\eta \end{bmatrix}$. However, we can show that Theorem D.3 holds for $P := \begin{bmatrix} 1-\eta_1 & \eta_1 \\ \eta_2 & 1-\eta_2 \end{bmatrix}$ where $0 \leq \eta_1 < \eta_2 < \eta$. The only change is that the high probability guarantee changes from $1 - e^{-\min\{\tau, \eta\} N}$ to $1 - e^{-\min\{\tau, \eta_1/\eta_2\} N}$. Note that the distribution $\mathcal{D}$ does not change.*

Proof of Lemma D.4.

Proof of Lemma D.4. We give a proof of Equations (119) and (121). The proofs of Equations (120) and (122) follow by replacing $Z = 1$ by $Z = 2$ in the following argument.

Since A only flips samples with feature x_B and $f^\star(x_B, z) = 0$ for all $z \in [2]$, we have that

$$\Pr_{\widehat{D}}[f^\star = 1, Z = 1] = \Pr_D[f^\star = 1, Z = 1] \tag{139}$$

Using the Chernoff bound, it follows that the next inequality holds with probability at least $1 - 2e^{-\frac{16}{3\tau^2}\delta^2 N}$

$$|\Pr_D[f^\star = 1, Z = 1] - \Pr_D[f^\star = 1, Z = 1]| \leq \delta \tag{140}$$

Equation (119) follows from Equations (139) and (140) by using the triangle inequality for the absolute value function. Since A flips η-fraction of the samples with $Z = 2$ to $Z = 1$, we have that

$$\Pr_{\widehat{D}}[Z = 1] = \Pr_D[Z = 1] + \eta \cdot \Pr_D[Z = 2], \tag{141}$$

$$\Pr_{\widehat{D}}[Z = 2] = \Pr_D[Z = 2] \cdot (1 - \eta). \tag{142}$$

Using the Chernoff bound, it follows that the next inequality holds with probability at least $1 - 4e^{-\frac{16}{3}\delta^2 N}$

$$|\Pr_D[Z = 1] - \Pr_D[Z = 1]| \leq \delta \tag{143}$$

Equation (121) follows from Equations (141) and (143) by using the triangle inequality for the absolute value function. □

D.2.3 [59]'s distributionally robust framework can output classifiers with low accuracy

In this section, for any $\eta \in (0, 1/4)$, we give an example where with high probability [59]'s distributionally robust optimization (DRO) framework outputs a classifier $f_{\mathsf{DRO}} \in \mathcal{F}$ whose error is at least $1/2 - \eta/2$ (Theorem D.6) On the same example, an optimal solution of Program (ErrTolerant) has error at most 2η and additively violates the fairness constraint by at most $O(\eta)$.

[59], in their distributionally robust approach, assume that 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. Formally, given a distribution $\mathcal{P}$, for each $\ell \in [p]$, let $\mathcal{P}_\ell$ be the distribution of features and labels in group ℓ when the data is drawn from $\mathcal{P}$, i.e., $\mathcal{P}_\ell$ is the distribution of $(X, Y) \mid Z = \ell$ when $(X, Y, Z) \sim \mathcal{P}$. Let $\mathcal{D}$ be the true distribution of samples and let $\widehat{D}$ be the (empirical) distribution of perturbed samples. Define a vector $\gamma \in [0, 1]^p$ as follows: For all $\ell \in [p]$

$$\gamma_\ell := \mathsf{TV}(\mathcal{D}_\ell, \widehat{D}_\ell). \tag{144}$$

[59] assume that an upper bound on γ is known. One can show that in presence of an η-Hamming adversary a tight upper bound on γ_ℓ is $\frac{\eta}{\Pr[Z=\ell]}$; It is achieved by the adversary that, given N samples, perturbs the protected attribute of $\eta \cdot N$ samples with protected attribute $Z = \ell$.

Let $\mathsf{D}(\gamma)$ be the set of all distributions $\mathcal{P}$ which satisfy that:

$$\text{for all } \ell \in [p], \quad \mathsf{TV}(\widehat{D}_\ell, \mathcal{P}_\ell) \leq \gamma_\ell. \tag{145}$$

Then, [59]'s output a classifier $f_{\mathsf{DRO}} \in \mathcal{F}$ which has the highest accuracy on $\widehat{D}$ such that it satisfies their fairness constraints for all distributions $\mathcal{P} \in \mathsf{D}(\gamma)$; for statistical rate, they solve

$$\min_{f \in \mathcal{F}} \ \mathsf{Err}_{\widehat{D}}(f), \quad \text{s.t.,} \quad \text{for all } \mathcal{P} \in \mathsf{D}(\gamma), \quad \mathsf{E}_P[f = 1 \mid Z = \ell] = \mathsf{E}_P[f = 1]. \tag{146}$$

Theorem D.6. *Suppose that there are two protected groups ($p := 2$) and $\mathcal{X}$ contains at least three distinct points. Then, there is a family of hypothesis classes $\mathcal{F}$ such that for all perturbation rates $\eta \in (0, 1/4)$, there is an adversary $A \in \mathcal{A}(\eta)$ and a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\} \times [2]$ that satisfies*

1. *Assumption 1 with $\lambda := 1/4$, and*
2. *$\Pr_D[Z = 1] = \Pr_D[Z = 2] = 1/2$,*

such that, for a draw of $N \in \mathbb{N}$ iid samples S from $\mathcal{D}$, with probability at least $1 - O(e^{-N})$ (over the draw of S), it holds that the optimal solution $f_{\mathsf{DRO}} \in \mathcal{F}$ of Program (146) with parameter $\gamma = (2\eta, 2\eta)$[9] and samples $A(S)$ has error

$$\mathsf{Err}_D(f_{\mathsf{DRO}}) \geq \frac{1}{2} - \eta, \tag{147}$$

[9] Following the fact mentioned earlier, for each $\ell \in [2]$, we set $\gamma_\ell := \frac{\eta}{\Pr[Z=\ell]}$.

and the optimal solution $f_{\mathsf{ET}} \in \mathcal{F}$ of Program (ErrTolerant) with parameters η, λ, and $\tau = 1$ and samples $A(S)$ has error

$$\mathsf{Err}_D(f_{\mathsf{ET}}) \leq 2\eta. \tag{148}$$

While $\_D(f_{\mathsf{DRO}}) = 1$ and $\_D(f_{\mathsf{ET}}) \geq 1 - O(\eta)$.

Proofs for Section D.2.3.

Setting P, $\mathcal{D}$, and A. Fix $\mathcal{X}$ to be any set with at least three distinct points, say x_A, x_B and x_C. Let $\mathcal{F}$ be any hypothesis class that shatters the set $\{x_A, x_B, x_C\} \times [2] \subseteq \mathcal{X} \times [p]$. Define $\mathcal{D}$ as the unique distribution such that for a draw $(X, Y, Z) \sim \mathcal{P}$, Y takes the value $\mathtt{I}[X \neq x_C]$ and that has the following marginal distribution:

	x_A	x_B	x_C
1	η	$1/4$	$1/4 - \eta$
2	0	$1/4$	$1/4$

Where for each $(r, s) \in \{x_A, x_B, x_C\} \times [2]$, the corresponding cell denotes $\mathsf{Pr}_{(X,Y,Z) \sim P}[(X, Z) = (r, s)]$. Because Y takes the value $\mathtt{I}[X \neq x_C]$, the classifier $f^\star(x, z) := \mathtt{I}[X \neq x_C]$ has 0 predictive error. We fix $A \in \mathcal{A}(\eta)$ to be the adversary that does not perturb any samples. (This suffices to prove the claim in Theorem D.6, but one can also other adversaries in $\mathcal{A}(\eta)$.)

Supporting lemmas. We use the following lemmas in the proof of Theorem D.6.

Lemma D.7. *The classifier $f^\star(X, Z) = \mathtt{I}[X = x_B]$ is an optimal solution for Program (2) with $\tau = 1$: $\mathsf{Err}_D(f^\star) = \eta$ and $\_D(f^\star) = 1$, where is the statistical rate fairness metric.*

Proof. One can verify that the classifier with the perfect accuracy, $f_{\mathsf{OPT}} := \mathtt{I}[X \neq x_C]$, has a statistical rate strictly smaller than 1. So it is not feasible for Program (2) for $\tau = 1$. Any feasible classifier $f \in \mathcal{F}$ must differ from f_{OPT} on some point in the support of $\mathcal{D}$. Since (by construction) all points in the support of $\mathcal{D}$ have a probability mass at least η, it follows that any $f \in \mathcal{F}$ feasible Program (2) must have an error at least η. Now the result follows since $f^\star := \mathtt{I}[X = x_A]$ is feasible for Program (2) and has the optimal error, $\mathsf{Err}_D(f^\star) = \eta$. □

Lemma D.8. *Consider the distribution $\mathcal{P}$, such that, for a draw $(X, Y, Z) \sim \mathcal{P}$, $Y := \mathtt{I}[X \neq x_C]$ and that has the following marginal distribution:*

	x_A	x_B	x_C
1	η	$1/4 - \eta$	$1/4$
2	0	$1/4$	$1/4$

Where for each $(r, s) \in \{x_A, x_B, x_C\} \times [2]$, the corresponding cell denotes $\mathsf{Pr}_{(X,Y,Z) \sim P}[(X, Z) = (r, s)]$. Given a draw of N iid samples from $\mathcal{D}$, with probability at least $1 - e^{(N)}$, it holds that $\mathcal{P} \in \mathtt{D}(\gamma)$ and $\mathcal{D} \in \mathtt{D}(\gamma)$.

Proof. Given a sufficient number of iid samples S from $\mathcal{D}$, one can show that with high probability, the empirical distribution D of S satisfies that: $\mathsf{TV}(D_1, \mathcal{D}_1) \leq \eta$ and $\mathsf{TV}(D_2, \mathcal{D}_2) \leq \eta$. Since $\gamma := (2\eta, 2\eta)$, this implies that with high probability $\mathcal{D} \in \mathtt{D}(\gamma)$. Further, the construction in Lemma D.8 ensures that $\mathsf{TV}(\mathcal{P}_1, \mathcal{D}_1) = \eta$ and $\mathsf{TV}(\mathcal{P}_2, \mathcal{D}_2) = \eta$. By using the triangle inequality of the total variation distance, it follows that with high probability, $\mathsf{TV}(\mathcal{P}_1, D_1) \leq 2\eta$ and $\mathsf{TV}(\mathcal{P}_2, D_2) \leq 2\eta$. Since $\gamma := (2\eta, 2\eta)$, it follows that with high probability $\mathcal{P} \in \mathtt{D}(\gamma)$. □

Lemma D.9. *Any $f \in \mathcal{F}$ that satisfies the following equalities*

$$\mathsf{E}_D[f = 1 \mid Z = \ell] = \mathsf{E}_D[f = 1], \tag{149}$$

$$\mathsf{E}_P[f = 1 \mid Z = \ell] = \mathsf{E}_P[f = 1], \tag{150}$$

where $\mathcal{D}$ is true distribution and $\mathcal{P}$ is the distribution from Lemma D.8 must have an error

$$\mathsf{Err}_D(f) \geq \frac{1}{2} - \eta.$$

Using Lemmas D.7 to D.9, Theorem D.6 follows as a corollary.

Proof of Theorem D.6. From Lemma D.8 know that with probability at least $1 - e^{(N)}$, $\mathcal{P} \in \mathsf{D}$ and $\mathcal{D} \in \mathsf{D}(\gamma)$. Suppose that this event happens. Assume that $f_{\mathsf{DRO}} \in \mathcal{F}$ is the optimal solution of Program (146). Since f_{DRO} is feasible for Program (146), it must satisfy that

$$\mathsf{E}_D[f_{\mathsf{DRO}} = 1 \mid Z = \ell] = \mathsf{E}_P[f_{\mathsf{DRO}} = 1],$$
$$\mathsf{E}_P[f_{\mathsf{DRO}} = 1 \mid Z = \ell] = \mathsf{E}_P[f_{\mathsf{DRO}} = 1],$$

Then Lemma D.9 tells us that $\mathsf{Err}_D(f_{\mathsf{DRO}}) \geq \frac{1}{2} - \eta$.

Finally, one can verify that when statistical rate is the fairness metric, $f^\star$ (from Lemma D.7) satisfies Assumption 1 with $\lambda = \frac{1}{2}$. Thus, Equation (148) and the inequality $_D(f_{\mathsf{ET}}) \geq 1 - O(\eta)$ follow from Theorem 4.3. □

Proof of Lemma D.9. Since that both $\mathcal{D}$ and $\mathcal{P}$ are supported on subsets of $\{x_A, x_B, x_C\} \times [2]$, it suffices to consider the restriction of $\mathcal{F}$ on this domain. Consider any classifier $f \in \mathcal{F}$ and define the following variables

$$f_{A1} := f(x_A, 1), \quad f_{B1} := f(x_B, 1), \quad f_{C1} := f(x_C, 1)$$
$$f_{A2} := f(x_A, 2), \quad f_{B2} := f(x_B, 2), \quad f_{C2} := f(x_C, 2),$$

denoting the predictions of f on $\{x_A, x_B, x_C\} \times [2]$. Since f satisfies Equation (149), we must have

$$\begin{aligned} 2 \cdot \left(\eta f_{A1} + \frac{1}{4} f_{A2} + \left(\frac{1}{4} - \eta\right) f_{A3}\right) &= \mathsf{E}_D[f = 1 \mid Z = \ell] \\ &= \mathsf{E}_D[f = 1] \\ &= \eta f_{A1} + \frac{1}{4} f_{A2} + \left(\frac{1}{4} - \eta\right) f_{A3} + \frac{1}{4} f_{B2} + \frac{1}{4} f_{B3}. \end{aligned} \tag{151}$$

Similarly, since f satisfies Equation (150), we must have

$$\begin{aligned} 2 \cdot \left(\eta f_{A1} + \left(\frac{1}{4} - \eta\right) f_{A2} + \frac{1}{4} f_{A3}\right) &= \mathsf{E}_P[f = 1 \mid Z = \ell] \\ &= \mathsf{E}_P[f = 1] \\ &= \eta f_{A1} + \left(\frac{1}{4} - \eta\right) f_{A2} + \frac{1}{4} f_{A3} + \frac{1}{4} f_{B2} + \frac{1}{4} f_{B3}. \end{aligned} \tag{152}$$

Combining Equations (151) and (152), we get

$$\begin{aligned} \eta f_{A1} + \left(\frac{1}{4} - \eta\right) f_{A2} + \frac{1}{4} f_{A3} &= \frac{1}{4} f_{B2} + \frac{1}{4} f_{B3} \\ &= \eta f_{A1} + \frac{1}{4} f_{A2} + \left(\frac{1}{4} - \eta\right) f_{A3}. \end{aligned} \tag{153}$$

On canceling the like terms in the LHS and RHS, and using that $\eta > 0$, we get

$$f_{A2} = f_{A3}.$$

We consider two cases.

(Case A) $f_{A2} = f_{A3} = 1$**:** Substituting $f_{A2} = f_{A3} = 1$ in Equation (153), we get

$$\eta f_{A1} + \frac{1}{2} - \eta = \frac{1}{4} f_{B2} + \frac{1}{4} f_{B3}.$$

Here, the RHS can only take values $\{0, 1/4, 1/2\}$ and the LHS can only take values $\{1/2 - \eta, 1/2\}$. Thus, the unique solution is $f_{A1} = f_{B2} = f_{B3} = 1$. One can verify that the unique resulting classifier has error $\mathsf{Err}_D(f) = 1/2 - \eta$.

(Case B) $f_{A2} = f_{A3} = 0$**:** Substituting $f_{A2} = f_{A3} = 0$ in Equation (153), we get

$$\eta f_{A1} = \frac{1}{4} f_{B2} + \frac{1}{4} f_{B3}.$$

Here, the RHS can only take values $\{0, 1/4, 1/2\}$ and the LHS can only take values $\{0, \eta\}$. Thus, the unique solution is $f_{A1} = f_{B2} = f_{B3} = 0$. One can verify that the unique resulting classifier has error $\mathsf{Err}_D(f) = 1/2$.

Thus, any $f \in \mathcal{F}$ satisfying Equations (149) and (150), must have an error at least $1/2 - \eta$. □

E Implementation details and additional empirical results

In this section, we give an implementation of our optimization framework using the logistic loss function (Section E.1.1), list the all hyper-parameters used for our approach (Section E.1.1) and baselines (Section E.1.4), and present some additional empirical results (Section E.2).

Code. The code for all the simulations is available at `https://github.com/AnayMehrotra/Fair-classification-with-adversarial-perturbations`.

E.1 Implementation details

E.1.1 Implementation of our framework using logistic regression

As an illustration, we implement our optimization framework (Program (ErrTolerant+)) using the logistic regression framework. For some $\theta \in \mathsf{R}^d$, let f_θ be the linear-classifier that given an input $x \in \mathsf{R}^d$ predicts $\mathtt{I}\left[1/\left(1 + e^{\ \hbar x, \theta i}\right) \geq 0.5\right]$ (or equivalently that predicts $\mathtt{I}[\langle x, \theta\rangle \geq 0]$). Then, the logistic regression framework considers the following hypothesis class: $\mathcal{F}_{\mathsf{LR}} := \left\{f_\theta \mid \theta \in \mathsf{R}^d\right\}$; see [54] for more details.

Several baselines (e.g., **CHKV** and **LMZV**) do not use protected attributes for prediction. For a fair comparison, in this implementation we do not use protected attributes for prediction. Let the domain of the features $\mathcal{X}$ satisfy $\mathcal{X} \subseteq \mathsf{R}^t$ for some $t \in \mathsf{N}$. In this case, we have $d := t$, where d is the dimension of θ (which parameterizes the hypothesis class $\mathcal{F}_{\mathsf{LR}}$). (To use protected attributes for prediction, one can set $d := t + 1$ and append the protected attribute to the features.)

Recall that Program (ErrTolerant+) takes the following values as input: the perturbation rate $\eta \in [0, 1]$, fairness threshold $\tau \in [0, 1]$, and for each $\ell \in [p]$, $\lambda_\ell \in [0, 1]$ and $\gamma_\ell \in [0, 1]$. Given these, we solve the following problem and initialize s to be its solution

$$\min_{\eta_1, \eta_2, \ldots, \eta_p \ \ 0} \ \min_{\ell, k 2 [p]} \frac{1 - \eta_\ell/\lambda_\ell}{1 + (\eta_k - \eta_\ell)/\gamma_\ell} \cdot \frac{1 + (\eta_\ell - \eta_k)/\gamma_k}{1 + \eta_\ell/\lambda_k}, \quad \text{s.t.,} \quad \sum_{\ell 2 [p]} \eta_\ell \leq \eta + \quad . \tag{154}$$

We solve Program (154) once to initialize s, then the same value s is used for all runs of Program (ErrTolerant+). Let $\mathcal{E}(f)$ and $\mathcal{E}^0(f)$ denote the events defining the relevant linear-fractional fairness metric (Definition 3.1). Let $S := \{(x_i, y_i, z_i)\}_{i 2 [N]}$ be the perturbed samples. Then we solve the following constrained optimization program

$$\min_{\theta 2 \mathsf{R}^d} \quad \frac{1}{N} \cdot \sum_{i=1}^{N} y_i \cdot \log f_\theta(x_i) + (1 - y_i) \cdot \log\left(1 - f_\theta(x_i)\right) \qquad \text{(ErrTolerant+)} \tag{155}$$

$$\text{s.t.,} \qquad \min_{\ell, k 2 [p]} \frac{q_\ell(f)}{q_k(f)} \geq \tau \cdot s$$

$$\forall\, \ell \in [p], \quad \frac{1}{N} \sum_{i 2 [N]} \mathtt{I}[\mathcal{E}(f(x_i)), \mathcal{E}^0(f(x_i)), Z = \ell] \geq \lambda_\ell - \eta - \quad ,$$

where for each $\ell \in [p]$,

$$q_\ell(f) := \frac{\sum_{i 2 [N] : z_i = \ell} \mathtt{I}[\mathcal{E}(f(x_i)), \mathcal{E}^0(f(x_i))]}{\sum_{i 2 [N] : z_i = \ell} \mathtt{I}[\mathcal{E}^0(f(x_i))]}.$$

In particular, for statistical rate, for each $i \in [N]$,

$$\mathcal{E}(f(x_i)) := \mathtt{I}[f(x_i) = 1] \text{ and } \mathcal{E}^0(f(x_i)) := 1,$$

and for false-positives rate, for each $i \in [N]$,

$$\mathcal{E}(f(x_i)) := \mathtt{I}[f(x_i) = 1] \text{ and } \mathcal{E}^0(f(x_i)) := \mathtt{I}[y_i = 0].$$

By substituting the appropriate $\mathcal{E}$ and $\mathcal{E}^0$, one can extend this implementation to any linear-fractional fairness metric (Definition 3.1).

Hyper-parameters. As a heuristic, given $\mathcal{E}$ and $\mathcal{E}^0$, in our simulations for each $\ell \in [p]$, we set $\gamma_\ell = \lambda_\ell := \mathsf{Pr}_{(X,Y,\widehat{Z}) \quad \widehat{D}}[\mathcal{E}^0(Y), \widehat{Z} = \ell]$, where $\widehat{D}$ is the empirical distribution of $\widehat{S}$ and Y is the label in perturbed data $\widehat{S}$. We find that these estimates suffice, and expect that a more refined approach would only improve the performance of **Err-Tol**. For all simulations, we set $:= 10^{-2}$.

E.1.2 SLSQP parameters

For simplicity, we do not implement the algorithm mentioned in Section C, and instead use existing optimization packages in our implementation. We solve both Program (ErrTolerant+) and Program (7) using the SLSQP solver [41] in SciPy [57]. For each optimization problem, we run the solver for 1000 iterations with parameters $\mathtt{ftol} = 10^{-4}$ and $\mathtt{eps} = 10^{-4}$, starting at a point chosen uniformly at random. If the solver fails to find a feasible solution, we rerun the solver for up to 10 iterations. If it does not find a feasible solution after 10 iterations, we return the infeasible point reached.

E.1.3 Implementation details of the adversaries

Across our simulations we consider three η-Hamming adversaries (which we call A_{TN}, A_{FN}, and A_{FP}). Each adversary has access to the true samples S, the fairness metric , and the desired fairness threshold τ. Using these, the adversary computes the "optimal fair classifier" $f^\star$ that has the highest accuracy (on S) subject to satisfying $(f, S) \geq \tau$. $f^\star$ is an optimal solution of Program (2); note that Program (2) is a special case of Program (ErrTolerant) (with $\lambda, \eta, \to 0$). In practice, we compute $f^\star$ by solving Program (2) on the unperturbed data S, using the SLSQP solver in the SciPy package to heuristically solve Program (2) (with the same parameters as described in Section E.1.2).

After computing $f^\star$, 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} and A_{FP} are identical, except that they consider the set of all false negatives and false positives of $f^\star$ respectively.

E.1.4 Baseline parameters and implementation

LMZV. We use the implementation of the **LMZV** at `https://github.com/AIasd/noise_fairlearn` provided by [44]; where the base classifier is by [1]. **LMZV** takes group-specific perturbation-rates, for each $\ell \in [p]$, $\eta_\ell := \mathsf{Pr}_D[\widehat{Z} \neq Z \mid Z = \ell]$, as input, and controls for additive statistical rate. The desired level of fairness is controlled by $\delta_L \in [0, 1]$, where smaller δ_L corresponds to higher fairness. We refer the reader to [44] for a description of these parameters. In our simulations, we vary δ_L over $\{10^{-2}, 4 \cdot 10^{-2}, 10^{-1}\}$. We fix all other hyper-parameters to the ones suggested by the authors for COMPAS.

AKM. We use the implementation of **AKM** at `https://github.com/matthklein/equalized_odds_under_perturbation` provided by [6]; **AKM** is the equalized-odds postprocessing method of [32]. It takes the unconstrained optimal classifier (**Uncons**) as input, and post-processes its outputs to control for equalized-odds constraints.

WGN+DRO. This is the distributionally robust framework of [59]; we use the implementation of **WGN+DRO** at `https://github.com/wenshuoguo/robust-fairness-code` provided by [59], which controls for additive false-positive rate. It takes true and perturbed protected attributes as input and computes the required bound on the total variation distance. We use the following learning rates for **WGN+DRO:** $\eta_\theta \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, $\eta_\lambda \in \{1/4, 1/2, 1, 2\}$, and $\eta_{\widetilde{p}_j} \in \{10^{-3}, 10^{-2}, 10^{-1}\}$; these are the same as the learning rates used by the authors (see [59, Table 2]). We refer the reader to [59] for the details of the parameters. The implementation runs **WGN+DRO** for all combinations of learning rates and outputs the classifier that has the best training objective and satisfies their constraints.

WGN+SW. This is the "soft-weights" framework of [59]; we use the implementation of **WGN+SW** at `https://github.com/wenshuoguo/robust-fairness-code` provided by [59]. It takes true and perturbed protected attributes as input and controls for additive false-positive rate. We use the following learning rates for **WGN+SW:** $\eta_\theta \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, $\eta_\lambda \in \{1/4, 1/2, 1, 2\}$, and $\eta_w \in \{10^{-3}, 10^{-2}, 10^{-1}\}$; these are the same as the learning rates used by the authors (see [59, Table 2]). See [59] for a discussion on the parameters. Their implementation runs **WGN+SW** for all combinations of learning rates and outputs the classifier that has the best training objective and satisfies their constraints.

CHKV and CHKV-FPR. We use the implementation of the [14]'s framework at `https://github.com/vijaykeswani/Noisy-Fair-Classification` provided by [14]. We use the implementations for statistical rate and false-positive rate, which we refer to these as **CHKV** and **CHKV-FPR** respectively. Both implementations take group specific perturbation-rates, for each $\ell \in [p]$, $\eta_\ell := \Pr_D[\widehat{Z} \neq Z \mid Z = \ell]$, as input. The desired level of fairness is controlled by $\tau \in [0, 1]$, where a larger τ corresponds to higher fairness. In our simulations, we vary τ over $\{0.7, 0.8, 0.9, 0.95, 1.0\}$; other hyper-parameters were the same as those suggested by the authors for COMPAS.

KL. This is the framework of [40] which controls for true-positive rate. It takes the perturbation rate η and for each $\ell \in [p]$, the probability $p_{1\ell} := \Pr_D[Z = \ell, Y = 1]$ as input; where D is the empirical distribution of S. [40] do not provide an implementation of **KL**. We implement **KL** using the logistic loss function. In particular, we solve the following optimization problem

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{N} \cdot \sum_{i=1}^{N} y_i \cdot \log f_\theta(x_i) + (1 - y_i) \cdot \log\left(1 - f_\theta(x_i)\right) \tag{156}$$

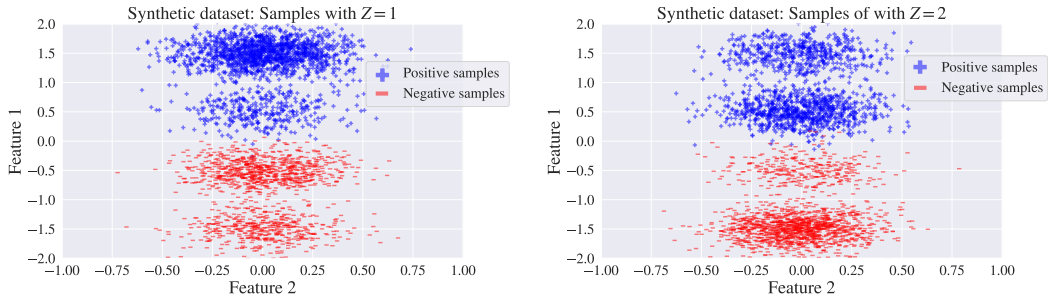
$$\text{s.t.,} \qquad \forall\, \ell \in [p], \quad \frac{\sum_{i \in [N]} \mathbb{I}[f(x_i) = 0, y_i = 1, \widehat{z}_i = \ell]}{\sum_{i \in [N]} \mathbb{I}[y_i = 1, \widehat{z}_i = \ell]} \leq \frac{6\eta}{\min_{\ell \in [p]} p_{1\ell} + 3\eta}$$

We solve Problem (156) using the standard implementation of the SLSQP solver in SciPy [57]; with the same parameters Section E.1.2.

E.1.5 Computational resources used

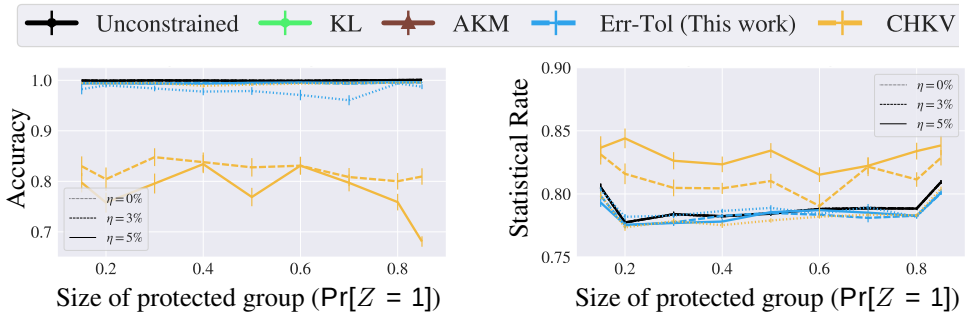
All simulations were run on a `t3a.2xlarge` instance, with 8 vCPUs and 32 Gb RAM, on Amazon's Elastic Compute Cloud (EC2).

E.2 Visualization of synthetic data



(a) Samples with protected attribute $Z = 1$ (b) Samples with protected attribute $Z = 2$

Figure 4: *Samples from the synthetic data (Section 5).* Figure 4(a) shows the samples with protected attribute $Z = 1$ and Figure 4(b) shows the samples with protected attribute $Z = 2$. We consider synthetic data with 1,000 samples with 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 four 2D Gaussians. The distribution of the labels and features is such that 1) the protected group $Z = 1$ has a higher likelihood of a positive label than the protected group $Z = 2$, and 2) **Uncons** has a near-perfect accuracy ($> 99\%$) and a statistical rate of 0.8 on S.



(a) A_{TN} with $\eta = 1.75\%$.

(b) A_{FN} with $\eta = 1.75\%$.

Figure 5: *Simulation varying the size of the protected groups on the synthetic data (see Section E.3.1):* We vary the perturbation rate η over $\{0, 0.03, 0.05\}$ and the size α of the protected group denoted by $Z = 1$ from 15% to 85% (see Section E.3.1). For each pair of α and η, we run **CHKV** and **Err-Tol** with $\tau = 0.8$ and report the average accuracy and statistical rate plots Figures 5(a) and 5(b) (respectively). (In the plot, the color of the line identifies the algorithm and its style identifies the value of η). The y-axis depicts accuracy or statistical rate and the x-axis depicts α; both the accuracy and the statistical rate are computed over the unperturbed test set (we refer the reader to Section 5 for further details). We observe that prior approaches can fail to satisfy their guarantees under the η-Hamming model, and their deviation from the accuracy guarantee increases as the size of one protected group decreases (i.e., when α approaches 0.85). Error bars represent the standard error of the mean over 100 iterations.

E.3 Additional empirical results

E.3.1 Simulation varying the size of the protected groups on synthetic data

In this simulation, we study the effect of varying the relative sizes of the protected groups in the synthetic data on the results of the simulation in Section 5. We vary the size of one protected group α from 0.15 to 0.85 (and the size of the other from 0.85 to 0.15.[10] Recall that the synthetic data in Section 5 is generated by sampling 500 samples with $Z = 1$ from a distribution $\mathcal{D}_1$ and sampling 500 samples with $Z = 2$ from different distribution $\mathcal{D}_2$ (where both $\mathcal{D}_1$ and $\mathcal{D}_2$ are mixtures of four 2D Gaussians). (See Figure 4 for a plot of samples from $\mathcal{D}_1$ and $\mathcal{D}_2$). To vary the size of the protected groups, given an $\alpha \in [0, 1]$, we draw $1000 \cdot \alpha$ samples with $Z = 1$ from $\mathcal{D}_1$ and $1000 \cdot (1 - \alpha)$ samples with $Z = 2$ from $\mathcal{D}_2$. For each value of α, we rerun the simulation from Section 5 on the resulting data.

Results. We observe that varying α from from 15% to 70% does not change the accuracy and statistical rate of the algorithms significantly (the accuracy changed by <5% and the statistical rate changed by <1%). However, as α approaches 85%, we observe that **CHKV's** accuracy reduced by 15% (to ≈0.68) and its statistical rate increased by 4% (to ≈0.86). In contrast, **Err-Tol** continued to have high accuracy (>0.99) and statistical rate (≥0.77), without large changes in either (its accuracy changed by <1% and statistical rate changed by <2%). Overall, we observe that prior approaches can fail to satisfy their guarantees under the η-Hamming model, and their deviation from the accuracy guarantee increases as the size of one protected group decreases (i.e., when α approaches 0.85).

E.3.2 Simulations with stochastic perturbations on the COMPAS data

In this simulation, we evaluate our framework under stochastic perturbations on the COMPAS data, and show that it has a similar statistical rate and accuracy trade-off as approaches tailored for stochastic perturbations (e.g., [44] and [14]). Concretely, we consider a binary protected attribute and the perturbation model studied by [14]: Suppose we have a single binary protected attribute (i.e., $p = 2$). Given values $\eta_1, \eta_2 \in [0, 1]$, the protected attributes of each item with protected attribute $Z = 1$ change to $\widehat{Z} = 2$ with probability η_1 (independently), and similarly, the protected attributes of each item with protected attribute $Z = 2$ change to $\widehat{Z} = 1$ with probability η_2 (independently). We consider the COMPAS data as preprocessed by [9], and consider gender (coded as binary) as the protected attribute. We consider four values of (η_1, η_2): $(0\%, 0\%)$, $(0\%, 3.5\%)$, $(3.5\%, 0\%)$, and $(3.5\%, 3.5\%)$.

[10] 15% and 85% are (roughly) the smallest and largest group sizes for which $f^\star$ has a sufficient number of true negatives to use A_{TN} with $\eta = 5\%$.

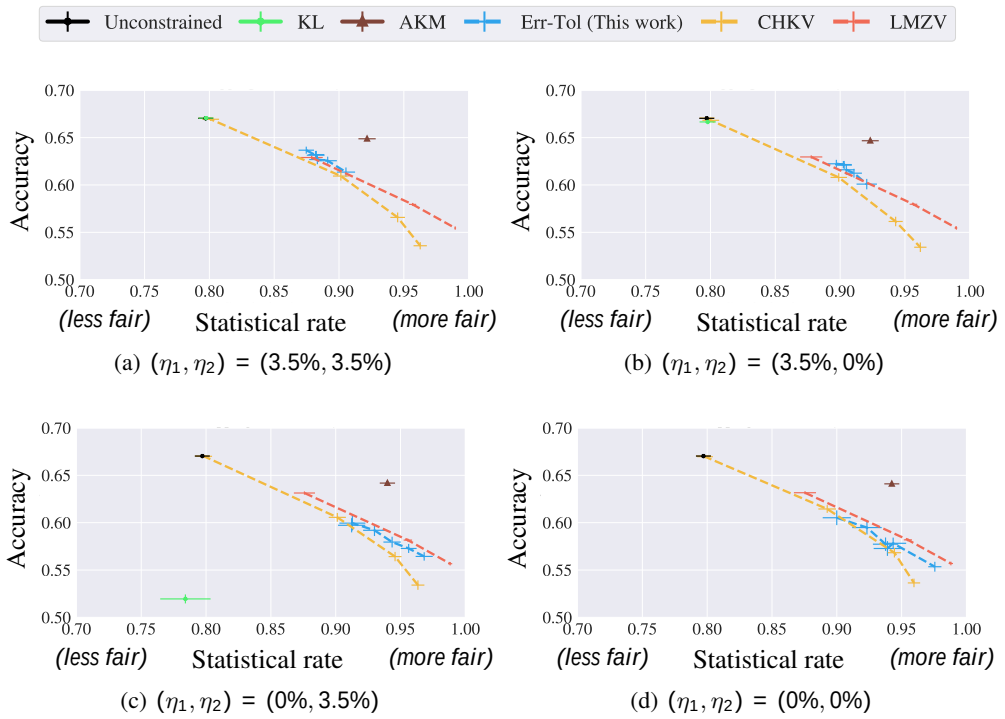


Figure 6: *Simulations on COMPAS data with flipping noise (see Section E.3.2):* Perturbed data is generated using the flipping noise model of [14, Definition 2.3] with η_1 and η_2 mentioned with each subfigure. All algorithms are run the perturbed data varying the fairness parameters ($\tau \in [0.7, 1]$ and $\delta_L \in [0, 0.1]$). The y-axis depicts the accuracy and the x-axis depicts statistical rate; both values are computed over the unperturbed test set. We observe that in each case, our approach, **Err-Tol**, has a similar fairness-accuracy trade-off as approaches tailored for flipping noise [14, 44] Error bars represent the standard error of the mean over 100 iterations.

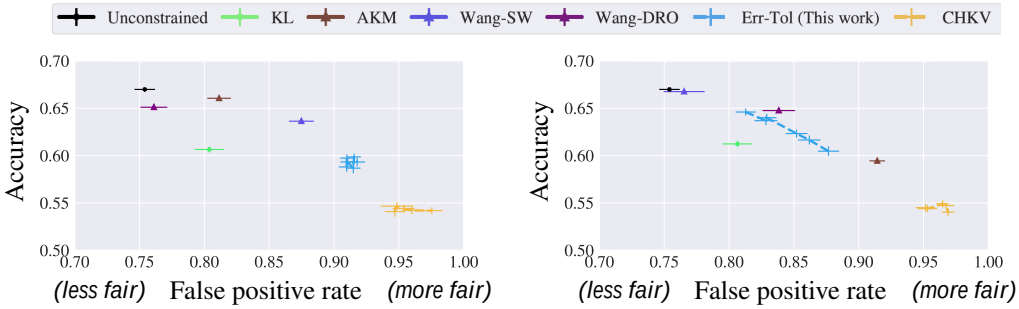
Results. The accuracy and statistical rate of **Err-Tol** and baselines for $\tau \in [0.7, 1]$ and $\delta_L \in [0, 0.1]$, averaged over 100 iterations, are reported in Figure 6. For all settings of η_1 and η_2, **Err-Tol** attains a better statistical rate than the unconstrained classifier (**Uncons**) for a small trade-off in accuracy. Further, **Err-Tol** has a similar statistical rate and accuracy trade-off as **CHKV** and **LMZV.** In all cases, **AKM** a better statistical rate and accuracy trade-off than all other approaches. Understanding why **AKM** has a better trade-off than other approaches with flipping noise requires further study. But it is likely because **AKM** does not need to make pessimistic assumptions on the data (as it does not account for perturbations) and outputs a classifier from a richer hypothesis class compared to other approaches. However, we recall that, when the perturbations are adversarial, **Err-Tol** has a better accuracy and statistical rate trade-off than **AKM** (see Figure 1).

E.3.3 Simulations with adversarial perturbations and false-positive rate on COMPAS data

In this simulation, we evaluate our framework for false-positive rate fairness metric with adversarial perturbations against state-of-the-art fair classification frameworks for false-positive rate under stochastic perturbations: **WGN+SW** [59], **WGN+DRO** [59], and **CHKV-FPR** [14]. We also compare against **KL** [40], which controls for true-positive rate (TPR) in the presence of a Malicious adversary, and **AKM** [6] that is the post-processing method of [32] and controls for equalized-odds fairness constraints.

Similar to the simulation with statistical rate fairness metric (see Section 5), **Err-Tol** is given the perturbation rate η. To advantage the baselines in our comparison, we provide them with more information as needed by their approaches:

1. **WGN+SW** is given both the true and perturbed protected attributes as input; it internally generates the auxiliary data needed by [59]'s "soft-weights" approach.
2. **WGN+DRO** is given both the true and perturbed protected attributes as input; it internally computes the total variation distances needed by [59]'s distributionally robust approach.
3. **CHKV-FPR** is given group-specific perturbation rates: $\forall\, \ell \in [p]$, $\eta_\ell := \mathsf{Pr}_D[\widehat{Z} \neq Z \mid Z = \ell]$.
4. **KL** is given η and for each $\ell \in [p]$, the probability $\mathsf{Pr}_D[Z = \ell, Y = 1]$; where D is the empirical distribution of S.



(a) A_{TN} with $\eta = 1.75\%$. (b) A_{FN} with $\eta = 1.75\%$.

Figure 7: *Simulations on COMPAS with false-positive rate (see Section E.3.3):* Perturbed data is generated using adversaries A_{TN} (a) and A_{FN} as described in Section E with $\eta = 1.75\%$. All algorithms are run on the perturbed data by varying the fairness parameters ($\tau \in [0.7, 1]$). The y-axis depicts accuracy and the x-axis depicts false-positive rate computed over the unperturbed test set. We observe that for all adversaries our approach, **Err-Tol**, attains a better fairness than the unconstrained classifier (**Uncons**) with a natural trade-off in accuracy. Further, **Err-Tol** achieves a higher fairness than each baseline except **CHKV** on at least one of (a) or (b). **CHKV** attains a higher fairness but has a low accuracy (55%) (because it outputs the always-positive classifier). Error bars represent the standard error of the mean over 100 iterations.

Err-Tol implements Program (ErrTolerant+) which requires estimates of λ_ℓ and γ_ℓ for all $\ell \in [p]$. As a heuristic, we set $\gamma_\ell = \lambda_\ell := \Pr_{\widehat{D}}[Y = 1, Z = \ell]$, where $\widehat{D}$ is the empirical distribution of $\widehat{S}$. More generally, for a general linear-fractional fairness metric, given by $\mathcal{E}$ and $\mathcal{E}^0$, the heuristic is to set $\gamma_\ell = \lambda_\ell := \Pr_{\widehat{D}}[\mathcal{E}^0(Y), Z = \ell]$, where $\widehat{D}$ is the empirical distribution of $\widehat{S}$ and Y is the label in perturbed data $\widehat{S}$.

Adversaries. We consider the same adversaries as in Section 5 (which we call A_{TN} and A_{FN}). We consider a perturbation rate of $\eta = 1.75\%$. Again, 1.75% 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 false-positive rate less than the false-positive rate of **Uncons** on the perturbed data. This "threshold" perturbation rate is smaller for false-positive rate than for statistical rate because to the number of false positives of a classifier f is smaller than the number of positive predictions of f; hence, an adversary perturbing the same number of samples, perturbs a larger fraction of false positives than the fraction of positive predictions of f.

Results. The accuracy and statistical rate of **Err-Tol** and baselines for $\tau \in [0.7, 1]$ are reported in Figure 7. For both adversaries, **Err-Tol** attains a better false-positive rate than the unconstrained classifier (**Uncons**) for a small trade-off in accuracy. For adversary A_{TN} (Figure 7(a)), **Uncons** has false-positive rate (0.75) and accuracy (0.65). In contrast, **Err-Tol** achieves a significantly higher false-positive rate (0.92) with accuracy (0.60). In comparison, **CHKV-FPR** has a higher false-positive rate (0.97) but lower accuracy (0.55); this accuracy is close to the accuracy of the all always-positive classifier. Compared to **Err-Tol**, **WGN+SW** has a higher accuracy (0.64) but a lower false-positive rate (0.87), and other baselines have an even lower false-positive rate ($\leq$ 0.82) with accuracy comparable to **WGN+SW.** For adversary A_{FN} (Figure 1(b)), **Uncons** has false-positive rate (0.75) and accuracy (0.67), while **Err-Tol** has a high higher false-positive rate (0.87) and accuracy (0.61). This significantly outperforms **WGN+SW** which has false-positive rate (0.76) and accuracy (0.64). **AKM** achieves the higher false-positive rate (0.92) with a natural reduction in accuracy to 0.59. **CHKV** achieves the higher false-positive rate (0.94) but with a lower accuracy (0.55). Meanwhile, **WGN+DRO** has a comparable false-positive rate (0.84) and a comparable accuracy at the same false-positive rate (0.65), and **KL** has a lower false-positive rate (0.81) and lower accuracy (0.62).

E.3.4 Simulations with adversarial perturbations on the Adult data

In this simulation, we evaluate our framework on the Adult data [23] with the statistical rate fairness metric. The Adult data consists of rows corresponding to approximately 45,000 individuals, with 18 binary features and a binary class label that is 1 is the individual has an income greater than $50,000 USD and 0 otherwise. Among the binary features, we use gender as the protected attribute.

Baselines. Like the simulation with the statistical rate fairness metric on the COMPAS data (Section 5), we compare our framework with:

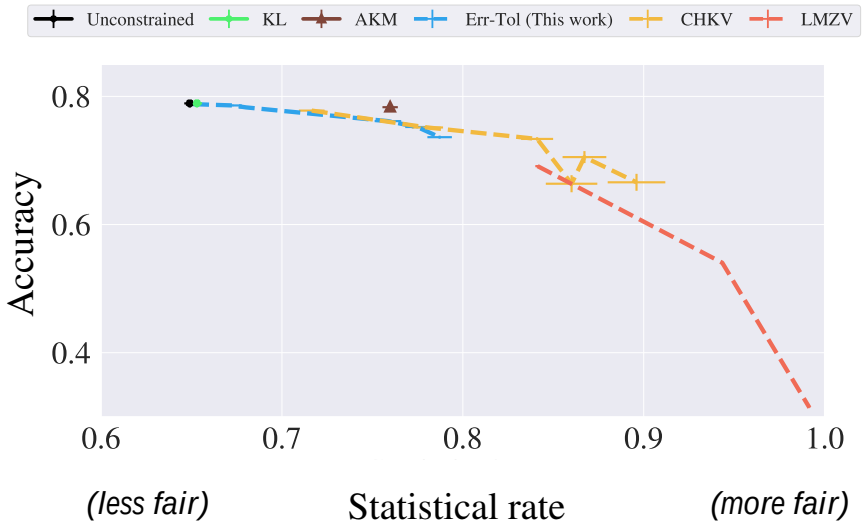


Figure 8: *Simulations on Adult data with gender as the protected attribute:* Let A_{FP} be the adversary that is the same as A_{TN}, except that instead of perturbing the true negatives of $f^\star$, it perturbs the false positives of $f^\star$. The perturbed data is generated using adversary A_{FP} (as described in Section E.3.4) with $\eta = 1\%$. 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. (Error bars represent the standard error of the mean over 100 iterations.) We observe that for our approach **Err-Tol** attains a better statistical rate than the unconstrained classifier **Uncons** with a small natural trade-off in accuracy. Further, **Err-Tol** achieves a better fairness-accuracy trade-off than **KL** and achieves a similar fairness-accuracy trade-off as **AKM, CHKV, LMZV.**

1. State-of-the-art fair classification frameworks for statistical rate under stochastic perturbations: **LMZV** [44] and **CHKV** [14].
2. **KL** [40], which controls for true-positive rate in the presence of a Malicious adversary.
3. **AKM** [6] that is the post-processing method of [32] and controls for equalized-odds fairness.
4. The optimal unconstrained classifier, **Uncons;** this is the same as [11]'s algorithm for PAC-learning in the Nasty Sample Noise Model without fairness constraints.

Adversaries and implementation details. We set $\eta = 1\%$ and consider an adversary A_{FP} that is the same as A_{TN}, except that instead of perturbing the true negatives of $f^\star$, it perturbs the false positives of $f^\star$.[11] Note that positive labels are rare in the Adult data (e.g., less than 4% of the total samples are positive and annotated as women). Thus, an adversary with $\eta \geq 4\%$ can remove all positive samples from one protected group–thereby, changing the statistical rate on the perturbed data by an arbitrary amount. This suggests that corrupting the positive labels is a hard case for learning fair classifiers on the Adult data. The adversary tries to reduce the performance of $f^\star$ on $Z = 1$ (which is the rarer than $Z = 2$) in $\widehat{S}$ by removing the samples that $f^\star$ predicts as positive. Thus, decreasing $f^\star$'s statistical rate on $\widehat{S}$. All other implementation details were identical to the simulation with the COMPAS data in Section 5.

Observations. 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 are reported in Figure 8. We observe that **Err-Tol** attains better fairness than the unconstrained classifier **Uncons** at a small tradeoff to accuracy. Further, **Err-Tol** has a fairness-accuracy tradeoff that is better than **KL** and at least as good as **AKM, CHKV,** and **LMZV.** These observations are consistent with our observations on the COMPAS data in Section 5.

Remark E.1. *We also explored the effect of changing the perturbation rate η over $\{0.5\%, 1\%, 1.5\%\}$. We report the results from this simulation in Figure 9. We observe that for all values of $\eta \in \{0.5\%, 1\%, 1.5\%\}$,* **Err-Tol** *achieves a higher statistical rate than the unconstrained classifier*

[11] We were unable to implement the analogous adversary A_{TP}, that perturbs the true positives of $f^\star$, because on the Adult data, $f^\star$ does not have sufficient number of true positives with $Z = 1$.

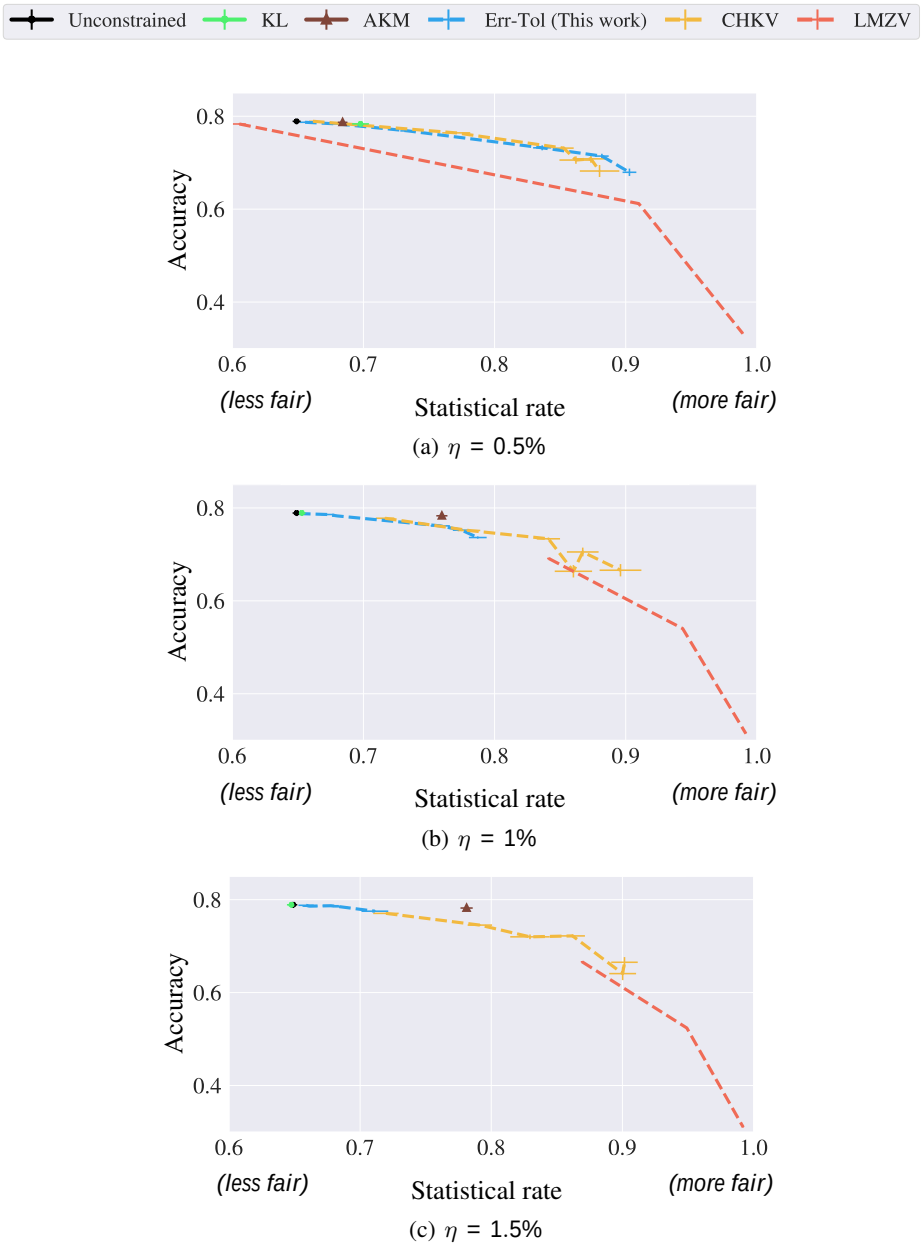


Figure 9: *Simulations on Adult data with adversarial noise on varying η (see Section E.3.2):* In this the results obtained by repeating the simulation in Section E.3.4 by varying η over $\{0.5\%, 1\%, 1.5\%\}$. The y-axis depicts the accuracy and the x-axis depicts statistical rate; both values are computed over the unperturbed test set. (Error bars represent the standard error of the mean over 50 iterations.) We observe that for all values of $\eta \in \{0.5\%, 1\%, 1.5\%\}$, **Err-Tol** achieves a higher statistical rate than the unconstrained classifier **Uncons** at a natural tradeoff to accuracy and **Err-Tol** has a similar (or better) fairness-accuracy tradeoff than other baselines. The only exception is **AKM**, which has a better fairness-accuracy tradeoff than **Err-Tol** at η = 1.5%. For further discussion of the results, we refer the reader to Remark E.1.

Uncons *at a natural tradeoff to accuracy and* **Err-Tol** *has a similar (or better) fairness-accuracy tradeoff than other baselines. The only exception is* **AKM**, *which has a better fairness-accuracy tradeoff than* **Err-Tol** *at η = 1.5%. Further, we observe that the highest statistical rate of achieved by* **Err-Tol** *decreases as η increases and this decrease is larger than the corresponding decrease in the statistical rate of* **AKM, CHKV,** *and* **LMZV.** *We believe this is because A_{FP} is not the worst-case adversary (for this data), and hence, our approach, which "protects" against the worst-case adversary, outputs a "more robust" classifier which happens to have a low statistical rate on Adult data. In contrast, the prior works do not correct for the worst-case adversaries and are able to output "less robust" classifiers which happen to have a high statistical rate for this adversary, but may perform poorly with the worst-case adversary.*