# Noise-tolerant fair classification

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Fairness-aware learning involves designing algorithms that do not discriminate with respect to some sensitive feature (e.g., race or gender). Existing work on the problem operates under the assumption that the sensitive feature available in one's training sample is perfectly reliable. This assumption may be violated in many real-world cases: for example, respondents to a survey may choose to conceal or obfuscate their group identity out of fear of potential discrimination. This poses the question of whether one can still learn fair classifiers given *noisy* sensitive features. In this paper, we answer the question in the affirmative: we show that if one measures fairness using the *mean-difference score*, and sensitive features are subject to noise from the *mutually contaminated learning* model, then owing to a simple identity we only need to change the desired fairness-tolerance. The requisite tolerance can be estimated by leveraging existing noise-rate estimators. We finally show that our procedure is empirically effective on two case-studies involving sensitive feature censoring.

## 1 Introduction

Classification is concerned with maximally discriminating between a number of pre-defined groups. *Fairness-aware* classification concerns the analysis and design of classifiers that do not discriminate with respect to some sensitive feature (e.g., race, gender, age, income). Recently, much progress has been made on devising appropriate measures of fairness (Calders et al., 2009; Dwork et al., 2011; Feldman, 2015; Hardt et al., 2016; Zafar et al., 2017b,a; Kusner et al., 2017; Kim et al., 2018; Speicher et al., 2018; Heidari et al., 2019), and means of achieving them (Zemel et al., 2013; Zafar et al., 2017b; Calmon et al., 2017; Dwork et al., 2018; Agarwal et al., 2018; Donini et al., 2018).

Typically, fairness is achieved by adding constraints which depend on the sensitive feature and by correcting one's learning procedure to achieve these fairness constraints. For example, suppose the data comprises of pairs of individuals and their loan repay status, and the sensitive feature is gender. Then, we may add a constraint that we should predict equal loan repayment for both men and women (see §3.2 for a more precise statement). However, this and similar approaches assume that we are able to correctly measure or obtain the sensitive feature. In many real-world cases, one may only observe noisy versions of the sensitive feature. For example, survey respondents may choose to conceal or obfuscate their group identity out of concerns of potential mistreatment or outright discrimination.

One is then brought to ask whether fair classification in the presence of such *noisy* sensitive features is still possible. Indeed, if the noise is high enough and all original information about the sensitive features is lost, then it is as if the sensitive feature was not provided. Standard learners can then be unfair on such data (Dwork et al., 2011; Pedreshi et al., 2008). Recently, Hashimoto et al. (2018) showed that progress is possible, albeit for specific fairness measures. The question of what can be done under a smaller amount of noise is thus both interesting and non-trivial.

In this paper, we consider two practical scenarios where we may only observe noisy sensitive features:

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

(1) suppose we are releasing data involving human participants. Even if noise-free sensitive features are available, we may wish to *add* noise so as to obfuscate sensitive attributes, so as to protect participant data from potential misuse. Thus, being able to learn fair classifiers under sensitive feature noise is a way to achieve both privacy *and* fairness.

(2) suppose we wish to analyse data where the presence of the sensitive feature is only known for a subset of individuals, while for others the feature value is unknown. For example, patients filling out a form may feel comfortable disclosing that they do not have a pre-existing medical condition; however, some who do have this condition may wish to refrain from responding. This can be seen as a variant of the *positive and unlabelled* (PU) setting (Denis, 1998), where the sensitive feature is present (positive) for some individuals, but absent (unlabelled) for others.

By considering popular measures of fairness and a general model of noise, we show that fair classification is possible under many settings, including the above. Our precise contributions are:

(**C1**) we show that if the sensitive features are subject to noise as per the *mutually contaminated learning model* (Scott et al., 2013a), and one measures fairness using the *mean-difference score* (Calders & Verwer, 2010), then a simple identity (Theorem 2) yields that we only need to change the desired fairness-tolerance. The requisite tolerance can be estimated by leveraging existing noise-rate estimators, yielding a reduction (Algorithm 1) to regular noiseless fair classification.

(**C2**) we show that our procedure is empirically effective on both case-studies mentioned above.

In what follows, we review the existing literature on learning fair and noise-tolerant classifiers in §2, and introduce the novel problem formulation of noise-tolerant fair learning in §3. We then detail how to address this problem in §4, and empirically confirm the efficacy of our approach in §5.

## 2 Related work

We review relevant literature on fair and noise-tolerant machine learning.

### 2.1 Fair machine learning

Algorithmic fairness has gained significant attention recently because of the undesirable social impact caused by bias in machine learning algorithms (Angwin et al., 2016; Buolamwini & Gebru, 2018; Lahoti et al., 2018). There are two central objectives: designing appropriate application-specific fairness criterion, and developing predictors that respect the chosen fairness conditions.

Broadly, fairness objectives can be categorised into individual- and group-level fairness. Individual-level fairness (Dwork et al., 2011; Kusner et al., 2017; Kim et al., 2018) requires the treatment of "similar" individuals to be similar. Group-level fairness asks the treatment of the groups divided based on some sensitive attributes (e.g., gender, race) to be similar. Popular notions of group-level fairness include demographic parity (Calders et al., 2009) and equality of opportunity (Hardt et al., 2016); see §3.2 for formal definitions.

Group-level fairness criteria have been the subject of more algorithmic design and analysis, and are achieved in three possible ways:

- pre-processing methods (Zemel et al., 2013; Louizos et al., 2015; Lum & Johndrow, 2016; Johndrow & Lum, 2017; Calmon et al., 2017; del Barrio et al., 2018; Adler et al., 2018), which usually find a new representation of data where the bias with respect to sensitive feature is explicitly removed.
- methods enforcing fairness during training (Calders et al., 2009; Woodworth et al., 2017; Zafar et al., 2017b; Agarwal et al., 2018), which usually add a constraint that is a proxy of the fairness criteria or add a regularization term to penalise fairness violation.
- post-processing methods (Feldman, 2015; Hardt et al., 2016), which usually apply a thresholding function to make the prediction satisfying the chosen fairness notion across groups.

### 2.2 Noise-tolerant classification

Designing noise-tolerant classifiers is a classic topic of study, concerned with the setting where one's training labels are corrupted in some manner. Typically, works in this area postulate a particular

model of label noise, and study the viability of learning under this model. Class-conditional noise (CCN) (Angluin & Laird, 1988) is one such effective noise model. Here, samples from each class have their labels flipped with some constant (but class-specific) probability. Algorithms that deal with CCN corruption have been well studied (Natarajan et al., 2013; Liu & Tao, 2016; Northcutt et al., 2017). These methods typically first estimate the noise rates, which are then used for prediction. A special case of CCN learning is learning from positive and unlabelled data (PU learning) (Elkan & Noto, 2008), where in lieu of explicit negative samples, one has a pool of unlabelled data.

Our interest in this paper will be the *mutually contaminated* (MC) *learning* noise model (Scott et al., 2013a). This model (described in detail in §3.3) captures both CCN and PU learning as special cases (Scott et al., 2013b; Menon et al., 2015), as well as other interesting noise models.

# 3 Background and notation

We recall the settings of standard and fairness-aware binary classification[1], and establish notation.

## 3.1 Standard binary classification

Binary classification concerns predicting the label or *target feature* $Y \in \{0, 1\}$ that best corresponds to a given instance $X \in \mathcal{X}$. Formally, suppose $D$ is a distribution over (instance, target feature) pairs from $\mathcal{X} \times \{0, 1\}$. Let $f \colon \mathcal{X} \to \mathbb{R}$ be a score function, and $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a user-defined class of such score functions. Finally, let $\ell : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_+$ be a loss function measuring the disagreement between a given score and binary label. The goal of binary classification is to minimise

$$L_D(f) := \mathbb{E}_{(X,Y)\sim D}[\ell(f(X), Y)].$$

## 3.2 Fairness-aware classification

In fairness-aware classification, the goal of accurately predicting the target feature $Y$ remains. However, there is an additional *sensitive feature* $A \in \{0, 1\}$ upon which we do not wish to discriminate. Intuitively, some user-defined fairness loss should be roughly the same regardless of $A$.

Formally, suppose $D$ is a distribution over (instance, sensitive feature, target feature) triplets from $\mathcal{X} \times \{0, 1\} \times \{0, 1\}$. The goal of *fairness-aware* binary classification is to find[2]

$$\begin{aligned} f^* &:= \operatorname*{arg\,min}_{f\in\mathcal{F}} L_D(f), \text{ such that } \Lambda_D(f) \leq \tau \\ L_D(f) &:= \mathbb{E}_{(X,A,Y)\sim D}[\ell(f(X), Y)], \end{aligned} \tag{1}$$

for user-specified *fairness tolerance* $\tau \geq 0$, and *fairness constraint* $\Lambda_D \colon \mathcal{F} \to \mathbb{R}_+$. Such constrained optimisation problems can be solved in various ways, e.g., convex relaxations (Donini et al., 2018), alternating minimisation (Zafar et al., 2017b; Cotter et al., 2018), or linearisation (Hardt et al., 2016).

A number of fairness constraints $\Lambda_D(\cdot)$ have been proposed in the literature. We focus on two important and specific choices in this paper, inspired by Donini et al. (2018):

$$\Lambda_D^{\text{DP}}(f) := \left|\bar{L}_{D_{0,\cdot}}(f) - \bar{L}_{D_{1,\cdot}}(f)\right| \tag{2}$$

$$\Lambda_D^{\text{EO}}(f) := \left|\bar{L}_{D_{0,1}}(f) - \bar{L}_{D_{1,1}}(f)\right|, \tag{3}$$

where we denote by $D_{a,\cdot}$, $D_{\cdot,y}$, and $D_{a,y}$ the distributions over $\mathcal{X} \times \{0, 1\} \times \{0, 1\}$ given by $D_{|A=a}$, $D_{|Y=y}$, and $D_{|A=a,Y=y}$ and $\bar{\ell} : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_+$ is the user-defined fairness loss with corresponding $\bar{L}_D(f) := \mathbb{E}_{(X,A,Y)\sim D}[\bar{\ell}(f(X), Y)]$. Intuitively, these measure the difference in the average of the fairness loss incurred among the instances with and without the sensitive feature.

Concretely, if $\bar{\ell}$ is taken to be $\bar{\ell}(s, y) = \mathbb{1}[\text{sign}(s) \neq 1]$ and the 0-1 loss $\bar{\ell}(s, y) = \mathbb{1}[\text{sign}(s) \neq y]$ respectively, then for $\tau = 0$, (2) and (3) correspond to the *demographic parity* (Dwork et al., 2011) and *equality of opportunity* (Hardt et al., 2016) constraints. Thus, we denote these two relaxed fairness measures *disparity of demographic parity* (DDP) and *disparity of equality of opportunity* (DEO). These quantities are also referred to as the *mean difference score* in Calders & Verwer (2010).

[1] For simplicity, we consider the setting of binary target and sensitive features. However, our derivation and method can be easily extended to the multi-class setting.

[2] Here, $f$ is assumed to not be allowed to use $A$ at test time, which is a common legal restriction (Lipton et al., 2018). Of course, $A$ can be used at training time to find an $f$ which satisfies the constraint.

### 3.3 Mutually contaminated learning

In the framework of learning from mutually contaminated distributions (MC learning) (Scott et al., 2013b), instead of observing samples from the "true" (or "clean") joint distribution $D$, one observes samples from a corrupted distribution $D_{\text{corr}}$. The corruption is such that the observed *class-conditional* distributions are mixtures of their true counterparts. More precisely, let $D_y$ denote the conditional distribution for label $y$. Then, one assumes that

$$\begin{aligned} D_{0,\text{corr}} &= (1-\alpha) \cdot D_1 + \alpha \cdot D_0 \\ D_{1,\text{corr}} &= \beta \cdot D_1 + (1-\beta) \cdot D_0, \end{aligned} \tag{4}$$

where $\alpha, \beta \in (0,1)$ are (typically unknown) noise parameters with $\alpha + \beta < 1$. Further, the corrupted base rate $\pi_{\text{corr}} := \mathbb{P}[Y_{\text{corr}} = 1]$ may be arbitrary. The MC learning framework subsumes CCN and PU learning (Scott et al., 2013b; Menon et al., 2015); thus, it is a flexible and appealing noise model.

## 4 Fairness under sensitive attribute noise

The standard fairness-aware learning problem assumes we have access to the true sensitive attribute, so that we can both measure and control our classifier's unfairness as measured by, e.g., Equation 2. Now suppose that rather than being given the sensitive attribute, we get a noisy version of it. We will show that the fairness constraint on the clean distribution is *equivalent* to a *scaled* constraint on the noisy distribution. This gives a simple reduction from fair machine learning in the presence of noise to the regular fair machine learning, which can be done in a variety of ways as discussed in §2.1.

### 4.1 Sensitive attribute noise model

As previously discussed, we use MC learning as our noise model, as this captures both CCN and PU learning as special cases; hence, we automatically obtain results for both these interesting settings.

Our specific formulation of MC learning noise on the sensitive feature is as follows. Recall from §3.2 that $D$ is a distribution over $\mathcal{X} \times \{0,1\} \times \{0,1\}$. Following (4), for unknown noise parameters $\alpha, \beta \in (0,1)$ with $\alpha + \beta < 1$, we assume that the corrupted class-conditional distributions are:

$$\begin{aligned} D_{1,\cdot,\text{corr}} &= (1-\alpha) \cdot D_{1,\cdot} + \alpha \cdot D_{0,\cdot} \\ D_{0,\cdot,\text{corr}} &= \beta \cdot D_{1,\cdot} + (1-\beta) \cdot D_{0,\cdot}, \end{aligned} \tag{5}$$

and that the corrupted base rate is $\pi_{a,\text{corr}}$ (we write the original base rate, $\mathbb{P}_{(X,A,Y)\sim D}[A=1]$ as $\pi_a$). That is, the distribution over (instance, label) pairs for the group with $A=1$, i.e. $\mathbb{P}(X, Y \mid A=1)$, is assumed to be mixed with the distribution for the group with $A=0$, and vice-versa.

Now, when interested in the EO constraint, it can be simpler to assume that the noise instead satisfies

$$\begin{aligned} D_{1,1,\text{corr}} &= (1-\alpha') \cdot D_{1,1} + \alpha' \cdot D_{0,1} \\ D_{0,1,\text{corr}} &= \beta' \cdot D_{1,1} + (1-\beta') \cdot D_{0,1}, \end{aligned} \tag{6}$$

for noise parameters $\alpha', \beta' \in (0,1)$. As shown by the following, this is not a different assumption.

**Lemma 1.** *If we assume that there is noise in the sensitive attribute only, as given in Equation* (5)*, then there exists $\alpha', \beta'$ such that Equation* (6) *holds.*

Although the lemma gives a way to calculate $\alpha', \beta'$ from $\alpha, \beta$, in practice it may be useful to consider (6) independently. Indeed, when one is interested in the EO constraints we will show below that only knowledge of $\alpha', \beta'$ is required. It is often much easier to estimate $\alpha', \beta'$ directly (which can be done in the same way as estimating $\alpha, \beta$ simply by considering $D_{\cdot,1,\text{corr}}$ rather than $D_{\text{corr}}$).

### 4.2 Fairness constraints under MC learning

We now show that fairness constraints are automatically robust to MC learning noise in $A$.

**Theorem 2.** *Assume that we have noise as per Equation* (5)*. Then,*

$$\begin{aligned} \Lambda^{\text{DP}}_{D}(f) \leq \tau &\iff \Lambda^{\text{DP}}_{D_{\text{corr}}}(f) \leq \tau \cdot (1-\alpha-\beta) \\ \Lambda^{\text{EO}}_{D_{\cdot,1}}(f) \leq \tau &\iff \Lambda^{\text{EO}}_{D_{\text{corr},\cdot,1}}(f) \leq \tau \cdot (1-\alpha'-\beta'), \end{aligned}$$

*where $\alpha'$ and $\beta'$ are as per Equation* (6) *and Lemma 1.*

The above can be seen as a consequence of the immunity of the *balanced error* (Chan & Stolfo, 1998; Brodersen et al., 2010; Menon et al., 2013) to corruption under the MC model. Specifically, consider a distribution $D$ over an input space $\mathcal{Z}$ and label space $\mathcal{W} = \{0, 1\}$. Define

$$B_D := \mathbb{E}_{Z|W=0}[h_0(Z)] + \mathbb{E}_{Z|W=1}[h_1(Z)]$$

for functions $h_0, h_1 \colon \mathcal{Z} \to \mathbb{R}$. Then, if for every $z \in \mathbb{R}$ $h_0(z) + h_1(z) = 0$, we have (van Rooyen, 2015, Theorem 4.16), (Blum & Mitchell, 1998; Zhang & Lee, 2008; Menon et al., 2015)

$$B_{D_{\text{corr}}} = (1 - \alpha - \beta) \cdot B_D, \tag{7}$$

where $D_{\text{corr}}$ refers to a corrupted version of $D$ under MC learning with noise parameters $\alpha, \beta$. That is, the effect of MC noise on $B_D$ is simply to perform a scaling. Observe that $B_D = \bar{L}_D(f)$ if we set $Z$ to $X \times Y$, $W$ to the sensitive feature $A$, and $h_0((x, y)) = +\bar{\ell}(y, f(x))$, $h_1((x, y)) = -\bar{\ell}(y, f(x))$. Thus, (7) implies $\bar{L}_D(f) = (1 - \alpha - \beta) \cdot \bar{L}_{D_{\text{corr}}}(f)$, and thus Theorem 2.

### 4.3 Algorithmic implications

Theorem 2 has an important algorithmic implication. Suppose we pick a fairness constraint $\Lambda_D$ and seek to solve Equation 1 for a given tolerance $\tau \geq 0$. Then, given samples from $D_{\text{corr}}$, it suffices to simply change the tolerance to $\tau' = \tau \cdot (1 - \alpha - \beta)$.

Unsurprisingly, $\tau'$ depends on the noise parameters $\alpha, \beta$. In practice, these will be unknown; however, there have been several algorithms proposed to estimate these from noisy data alone (Scott et al., 2013b; Menon et al., 2015; Liu & Tao, 2016; Ramaswamy et al., 2016; Northcutt et al., 2017). Thus, we may use these to construct estimates of $\alpha, \beta$, and plug these in to construct an estimate of $\tau'$.

In sum, we may tackle fair classification in the presence of noisy $A$ by suitably combining *any* existing fair classification method (that takes in a parameter $\tau$ that is proportional to mean-difference score of some fairness measures), and *any* existing noise estimation procedure. This is summarised in Algorithm 1. Here, FairAlg is any existing fairness-aware classification method that solves Equation 1, and NoiseEst is any noise estimation method that estimates $\alpha, \beta$.

**Algorithm 1** Reduction-based algorithm for fair classification given noisy $A$.

**Input:** Training set $S = \{(x_i, y_i, a_i)\}_{i=1}^n$, scorer class $\mathcal{F}$, fairness tolerance $\tau \geq 0$, fairness constraint $\Lambda(\cdot)$, fair classification algorithm FairAlg, noise estimation algorithm NoiseEst
**Output:** Fair classifier $f^* \in \mathcal{F}$

1: $\hat{\alpha}, \hat{\beta} \leftarrow \mathsf{NoiseEst}(S)$
2: $\tau' \leftarrow (1 - \hat{\alpha} - \hat{\beta}) \cdot \tau$
3: **return** $\mathsf{FairAlg}(S, \mathcal{F}, \Lambda, \tau')$

### 4.4 Connection to differential privacy

While Algorithm 1 gives a way of achieving fair classification on an already noisy dataset such as the use case described in example (2) of §1, it can also be used to simultaneously achieve fairness and privacy. As described in example (1) of §1, the very nature of the sensitive attribute makes it likely that even if noiseless sensitive attributes are available one might want to add noise to guarantee some form of privacy. Note that simply removing the feature does not suffice, because it would prohibit researchers from developing fairness-aware classifiers for the dataset. Formally, we can give the following privacy guarantee by adding CCN noise to the sensitive attribute.

**Lemma 3.** *To achieve $(\epsilon, \delta = 0)$ differential privacy on the sensitive attribute we can add CCN noise with $\rho^+ = \rho^- = \rho \geq \frac{1}{\exp(\epsilon)+1}$ to the sensitive attribute.*

Thus if a desired level of differential privacy is required before releasing a dataset, one could simply add the required amount of CCN noise to the sensitive attributes, publish this modified dataset as well as the noise level, and researchers could use Algorithm 1 (without even needing to estimate the noise rate) to do fair classification as usual.

Recently, Jagielski et al. (2018) explored preserving differential privacy (Dwork, 2006) while maintaining fairness constraints. The authors proposed two methods: one adds Laplace noise to training

data and apply the post-processing method in Hardt et al. (2016), while another modifies the method in Agarwal et al. (2018) using the exponential mechanism as well as Laplace noise. Our work differs from them in three major ways: (*1*) our work allows for fair classification to be done using a *any* fairness-aware classifier, whereas the method of Jagielski et al. (2018) requires the use of a particular classifier. (*2*) our focus is on designing fair-classifiers with noise-corrupted sensitive attributes; by contrast, the main concern in Jagielski et al. (2018) is achieving differential privacy. (*3*) we deal with not only equalized odds, but also demographic parity.

# 5 Experiments

We demonstrate that it is viable to learn fair classifiers given noisy sensitive features. As our underlying fairness-aware classifier, we use a modified version of the classifier implemented in Agarwal et al. (2018) with the DDP and DEO constraints which, as discussed in §3.2, are special cases of our more general constraints (2) and (3). The classifier's original constraints can also be shown to be noise-invariant but in a slightly different way (see Appendix C for a discussion). An advantage of this classifier is that it is shown to reach levels of fairness violation that are very close to the desired level ($\tau$), i.e., for small enough values of $\tau$ it will reach the constraint boundary.

While we had to choose a particular classifier, our method can be used before using any downstream fair classifier as long as it can take in a parameter $\tau$ that controls the strictness of the fairness constraint and that its constraints are special cases of our very general constraints (2) and (3).

## 5.1 Noise setting

Our case studies focus on two common special cases of MC learning: CCN and PU learning. Under CCN noise the sensitive feature's value is randomly flipped with probability $\rho^+$ if its value was 1 or with probability $\rho^-$ if its value was 0. As shown in Menon et al. (2015, Appendix C), CCN noise is a special case of MC learning. For PU learning we consider the censoring setting (Elkan & Noto, 2008) which is a special case of CCN learning where one of $\rho^+$ and $\rho^-$ is 0. While our results also apply to the case-controlled setting of PU learning (Ward et al., 2009), the former setting is more natural in our context. Note that from $\rho^+$ and $\rho^-$ one can obtain $\alpha$ and $\beta$ as described in Menon et al. (2015).

## 5.2 Benchmarks

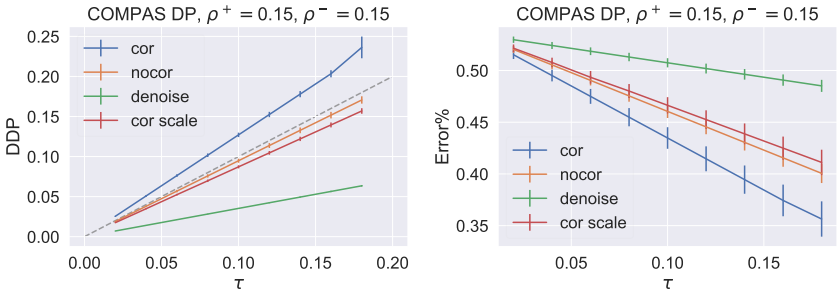
For each case study, we evaluate our method (termed cor scale); recall this scales the input parameter $\tau$ using Theorem 2 and the values of $\rho^+$ and $\rho^-$, and then uses the fair classifier to perform classification. We compare our method with three different baselines. The first two trivial baselines are applying the fair classifier directly on the non-corrupted data (termed nocor) and on the corrupted data (termed cor). While the first baseline is clearly the ideal, it won't be possible when only the corrupted data is available. The second baseline should show that there is indeed an empirical need to deal with the noise in some way and that it cannot simply be ignored.

The third, non-trivial, baseline (termed denoise) is to first denoise $A$ and then apply the fair classifier on the denoised distribution. This denoising is done by applying the RankPrune method in Northcutt et al. (2017). Note that we provide the RankPrune method with the same known values of $\rho^+$ and $\rho^-$ that we use to apply our scaling so this is a fair comparison to our method. Compared to denoise, we do *not* explicitly infer individual sensitive feature values; thus, our method does not compromise privacy.

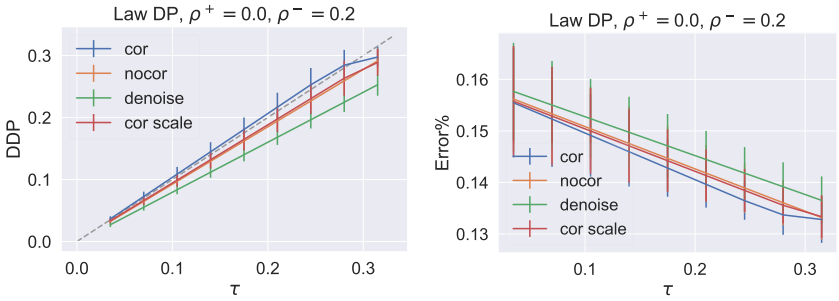
For both case studies, we study the relationship between the input parameter $\tau$ and the testing error and fairness violation. For simplicity, we only consider the DP constraint.

## 5.3 Case study: privacy preservation

In this case study, we look at `COMPAS`, a dataset from Propublica (Angwin et al., 2016) that is widely used in the study of fair algorithms. Given various features about convicted individuals, the task is to predict recidivism and the sensitive attribute is race. The data comprises 7918 examples and 10 features. In our experiment, we assume that to preserve differential privacy, CCN noise with $\rho^+ = \rho^- = 0.15$ is added to the sensitive attribute. As per Lemma 3, this guarantees $(\epsilon, \delta = 0)$ differential privacy with $\epsilon = 1.73$. We assume that the noise level $\rho$ is released with the dataset (and




**(a)** `COMPAS` dataset (privacy case study).




**(b)** `law school` dataset (PU learning case study).

**Figure 1:** Relationship between input fairness tolerance $\tau$ versus DP fairness violation (left panels), and versus error (right panels). Our method (cor scale) achieves approximately the ideal fairness violation (indicated by the gray dashed line in the left panels), with only a mild degradation in accuracy compared to training on the uncorrupted data (indicated by the nocor method). Baselines that perform no noise-correction (cor) and explicitly denoise the data (denoise) offer suboptimal tradeoffs by comparison; for example, the former achieves slightly lower error rates, but does so at the expense of greater fairness violation.

is thus known). We performed fair classification on this noisy data using our method and compare the results to the three benchmarks described above.

Figure 1a shows the average result over three runs each with a random 80-20 training-testing split. (Note that fairness violations and errors are calculated with respect to the true uncorrupted features.) We draw two key insights from this graph:

(i) in terms of fairness violation, our method (cor scale) approximately achieves the desired fairness tolerance (shown by the gray dashed line). This is both expected and ideal, and it matches what happens when there is no noise (nocor). By contrast, the naïve method cor strongly violates the fairness constraint.

(ii) in terms of accuracy, our method only suffers mildly compared with the ideal noiseless method (nocor); some degradation is expected as noise will lead to some loss of information. By contrast, denoise sacrifices much more predictive accuracy than our method.

In light of both the above, our method is seen to achieve the best overall tradeoff between fairness and accuracy. Experimental results with EO constraints, and other commonly studied datasets in the fairness literature (`adult`, `german`), show similar trends as in Figure 1a, and are included in Appendix D for completeness.

### 5.4 Case study: PU learning

In this case study, we consider the dataset `law school`, which is a subset of the original dataset from LSAC (Wightman, 1998). In this dataset, one is provided with information about various individuals (grades, part time/full time status, age, etc.) and must determine whether or not the individual passed the bar exam. The sensitive feature is race; we only consider black and white. After prepossessing the data by removing instances that had missing values and those belonging to other ethnicity groups (neither black nor white) we were left with 3738 examples each with 11 features.

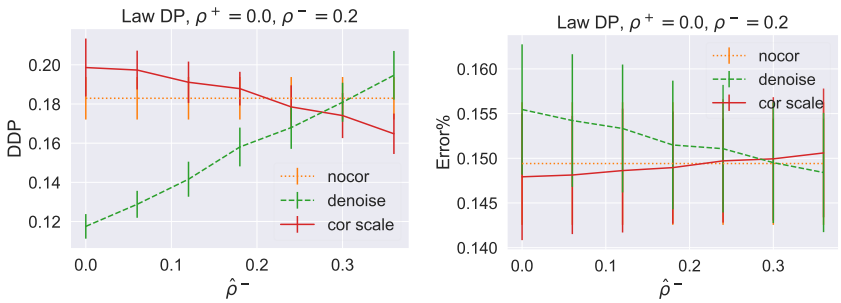



**Figure 2:** Relationship between the estimated noise level $\hat{\rho}^-$ and fairness violation/error on the `law school` dataset using DP constraint (testing curves), with $\hat{\rho}^+ = 0$ and $\tau = 0.2$. Our method (cor scale) is not overly sensitive to imperfect estimates of the noise rate, evidenced by its fairness violation and accuracy closely tracking that of training on the uncorrupted data (nocor) as $\hat{\rho}^-$ is varied. That is, red curve in the left plot closely tracks the yellow reference curve. By contrast, the baseline that explicitly denoises the data (denoise) deviates strongly from nocor, and is sensitive to small changes in $\hat{\rho}^-$. This illustrates that our method performs well even when noise rates must be estimated.

While the data ostensibly provides the true values of the sensitive attribute, one may imagine having access to only PU information. Indeed, when the data is collected one could imagine that individuals from the minority group would have a much greater incentive to conceal their group membership due to fear of discrimination. Thus, any individual identified as belonging to the majority group could be assumed to have been correctly identified (and would be part of the positive instances). On the other hand, no definitive conclusions could be drawn about individuals identified as belonging to the minority group (these would therefore be part of the unlabelled instances).

To model a PU learning scenario, we added CCN noise to the dataset with $\rho^+ = 0$ and $\rho^- = 0.2$. We initially assume that the noise rate is known. Figure 1b shows the average result over three runs under this setting each with a random 80-20 training-testing split. We draw the same conclusion as before: our method achieves the highest accuracy while respecting the specified fairness constraint.

Unlike in the privacy case, the noise rate in the PU learning scenario is usually unknown in practice, and must be estimated. Such estimates will inevitably be approximate. We thus evaluate the impact of the error of the noise rate estimate on all methods. In Figure 2, we consider a PU scenario where we only have access to an estimate $\hat{\rho}^-$ of the negative noise rate, whose true value is $\rho^- = 0.2$. Figure 2 shows the impact of different values of $\hat{\rho}^-$ on the fairness violation and error. We see that that as long as this estimate is reasonably accurate, our method performs the best in terms of being closest to the case of running the fair algorithm on uncorrupted data.

In sum, these results are consistent with our derivation and show that our method cor scale can achieve the desired degree of fairness while minimising loss of accuracy. Appendix E includes results for different settings of $\tau$, noise level, and on other datasets showing similar trends.

## 6 Conclusion and future work

In this paper, we showed both theoretically and empirically that even under the very general MC learning noise model (Scott et al., 2013a) on the sensitive feature, fairness can still be preserved by scaling the input unfairness tolerance parameter $\tau$. In future work, it would be interesting to consider the case of categorical sensitive attributes (as applicable, e.g., for race), and the more challenging case of instance-dependent noise (Awasthi et al., 2015).

# References


Bank marketing dataset (a sample taken from UCI). https://www.kaggle.com/rouseguy/bankbalanced/kernels.

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., and Venkatasubramanian, S. Auditing black-box models for indirect influence. *Knowl. Inf. Syst.*, 54(1):95–122, January 2018. ISSN 0219-1377.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 60–69, Stanford, CA, 2018. JMLR.

Angluin, D. and Laird, P. Learning from noisy examples. *Machine Learning*, 2(4):343–370, Apr 1988. ISSN 1573-0565.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. 2016.

Awasthi, P., Balcan, M.-F., Haghtalab, N., and Urner, R. Efficient learning of linear separators under bounded noise. In *Conference on Learning Theory (COLT)*, volume 40, pp. 167–190, 2015.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Conference on Learning Theory (COLT)*, pp. 92–100, 1998.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. The balanced accuracy and its posterior distribution. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ICPR '10, pp. 3121–3124, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4109-9.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Calders, T. and Verwer, S. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, Sep 2010. ISSN 1573-756X.

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

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems 30*, pp. 3992–4001, 2017.

Chan, P. K. and Stolfo, S. J. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, pp. 164–168, 1998.

Cotter, A., Gupta, M. R., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B. E., and You, S. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. *CoRR*, abs/1807.00028, 2018. URL http://arxiv.org/abs/1807.00028.

del Barrio, E., Gamboa, F., Gordaliza, P., and Loubes, J.-M. Obtaining fairness using optimal transport theory. *arXiv e-prints*, art. arXiv:1806.03195, June 2018.

Denis, F. PAC Learning from Positive Statistical queries, 1998.

Dheeru, D. and Karra Taniskidou, E. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems 31*, pp. 2796–2806, 2018.

Dwork, C. Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I. (eds.), *Automata, Languages and Programming*, pp. 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-35908-1.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 119–133, 2018.

Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220, 2008.

Feldman, M. Computational fairness: Preventing machine-learned discrimination. 2015.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3323–3331, USA, 2016. ISBN 978-1-5108-3881-9.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*, 2018.

Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. A moral framework for understanding of fair ml through economic models of equality of opportunity. *ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*)*, January 2019.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. Differentially private fair learning. 2018. URL `https://arxiv.org/pdf/1812.02696.pdf`.

Johndrow, J. E. and Lum, K. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv e-prints*, art. arXiv:1703.04957, March 2017.

Kim, M. P., Reingold, O., and Rothblum, G. N. Fairness through computationally-bounded awareness. *CoRR*, abs/1803.03239, 2018. URL `http://arxiv.org/abs/1803.03239`.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30*, pp. 4066–4076, 2017.

Lahoti, P., Weikum, G., and P. Gummadi, K. ifair: Learning individually fair data representations for algorithmic decision making. 06 2018.

Lipton, Z., McAuley, J., and Chouldechova, A. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pp. 8136–8146, 2018.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:447–461, 2016.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair auto encoder. 11 2015.

Lum, K. and Johndrow, J. E. A statistical framework for fair predictive algorithms. *CoRR*, abs/1610.08077, 2016.

Menon, A. K., Narasimhan, H., Agarwal, S., and Chawla, S. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Menon, A. K., van Rooyen, B., Ong, C. S., and Williamson, R. C. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Natarajan, N., Tewari, A., Dhillon, I. S., and Ravikumar, P. Learning with noisy labels. In *Neural Information Processing Systems (NIPS)*, dec 2013.

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

Pedreshi, D., Ruggieri, S., and Turini, F. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 560–568. ACM, 2008.

Ramaswamy, H. G., Scott, C., and Tewari, A. Mixture proportion estimation via kernel embedding of distributions. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 2052–2060. JMLR.org, 2016.

Scott, C., Blanchard, G., , and Handy, G. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on Learning Theory (COLT)*, volume 30, pp. 489–511, 2013a.

Scott, C., Blanchard, G., and Handy, G. Classification with asymmetric label noise: Consistency and maximal denoising. In Shalev-Shwartz, S. and Steinwart, I. (eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 489–511, Princeton, NJ, USA, 12–14 Jun 2013b. PMLR.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 2239–2248, 2018. ISBN 978-1-4503-5552-0.

van Rooyen, B. *Machine Learning via Transitions*. PhD thesis, The Australian National University, 2015.

Ward, G., Hastie, T., Barry, S., Elith, J., and Leathwick, J. R. Presence-Only Data and the {EM} Algorithm. *Biometrics*, 65(2):554–563, 2009.

Wightman, L. national longitudinal bar passage study, 1998.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pp. 1920–1953, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., Gummadi, K., and Weller, A. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems 30*, pp. 229–239, 2017a.

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

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

Zhang, D. and Lee, W. S. Learning classifiers without negative examples: A reduction approach. In *2008 Third International Conference on Digital Information Management*, pp. 638–643, Nov 2008.