Noise-tolerant fair classification

Alexandre Louis Lamy*† Ziyuan Zhong*† Aditya Krishna Menon‡ Nakul Verma†

Abstract

Fair machine learning concerns the analysis and design of learning algorithms that do not exhibit systematic bias with respect to some sensitive feature (e.g., race, gender). This subject has received sustained interest in the past few years, with considerable progress on both devising sensible measures of fairness, and means of achieving them. Typically, the latter involves correcting one's learning procedure so that there is no bias on the training sample. However, all such work has operated 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 privacy concerns. This poses the question of whether one can still learn fair classifiers in the presence of such *noisy* sensitive features.

In this paper, we answer the question in the affirmative for a widely-used measure of fairness and model of noise. 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 the canonical supervised learning problem, concerned with maximally discriminating between a number of pre-defined groups (e.g., determining if an individual is likely to repay a loan, or not). *Fairness-aware* classification concerns the analysis and design of classifiers that do not discriminate with respect to some sensitive feature (e.g., race, gender). Recently, much progress has been made in this subject, both on devising sensible measures of fairness (Calders et al., 2009; Dwork et al., 2011; Feldman, 2015; Kim et al., 2018; Hardt et al., 2016; Zafar et al., 2017b,a; Kusner et al., 2017; Speicher et al., 2018; Zhang & Bareinboim, 2018; Heidari et al., 2019), and on 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; Heidari 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 operate under the assumption that we are able to correctly measure or obtain the sensitive feature. In many real-world cases, one may only have access to noisy observations of the sensitive feature. A simple example is that respondents to a survey may choose to conceal or obfuscate their group identity out of privacy concerns.

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

*Equal contribution
†Department of Computer Science, Columbia University, New York, US, alexandre.l.lamy@columbia.edu, zz2521@columbia.edu, verma@cs.columbia.edu
‡Google, New York, US, adityakmenon@google.com

is lost, then it is as if the sensitive feature was not provided. It is well known that the result can then be unfair (Dwork et al., 2011; Pedreshi et al., 2008). While some progress is still possible (Hashimoto et al., 2018), results for general definitions of fairness are not known (and potentially impossible). The question of what can be done when there is a smaller amount of noise is thus both interesting and non-trivial.

In this paper, we consider two realistic situations where only a noisy version of the sensitive feature would be available:

(1) First, in some cases while the sensitive feature might be able to be measured without noise, it may be necessary or desirable to add noise to it before being able to publish the data. Indeed, one might legally or ethically have to obfuscate sensitive attributes so as to protect the privacy of a study's participants. Thus, being able to design fair classifiers despite being given only noisy versions of the sensitive feature is a way to achieve both privacy and fairness.

(2) Second, we consider cases where the sensitive feature might only be able to be recorded in the *Positive and Unlabelled* (PU) setting (Denis, 1998). That is, rather than being able to tell whether the sensitive feature is positive or negative, we are simply given positive and unlabeled entries. This could exist in historical datasets where the presence of the sensitive feature (disability for example) was occasionally recorded (positive points) but where most entries could have been either positive or negative (thus unlabelled points). Alternatively, in the medical setting 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 not provide truthful responses.

By considering a general measure of fairness and model of noise, we show that fair classification is indeed possible under many (including the above) settings. Our precise contributions (**C1**) and (**C2**) are as follows:

(**C1**) We show that if the sensitive features are subject to noise as per the *mutually contaminated learning model*, and one measures fairness using the *mean-difference score*, then a simple identity (Theorem 1) 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 which the sensitive feature is purposefully obfuscated for privacy reasons, or only available in the PU setting.

The paper proceeds as follows. We review the existing literature on learning fair and noise-tolerant classifiers in §2, and introduce the (to our knowledge novel) problem formulation of noise-tolerant fair learning in §3. We then detail how to address this problem in §4, and provide empirical case-studies confirming the efficacy of our approach in §5.

2 Related work

We review relevant literature from the fields of fair, noise-tolerant, and privacy-preserving machine learning.

2.1 Fair machine learning

Algorithmic fairness has gained significant attention in recent years because of the undesirable social impact caused by bias existing in machine learning algorithms Angwin et al. (2016); Buolamwini & Gebru (2018); Lahoti et al. (2018). There are two central problems: what is an appropriate fairness definition, and how to optimize the original objective while satisfying the chosen fairness definition.

To deal with the first problem, there have been multiple fairness definitions been proposed. These definitions fall into two main categories: individual- and group-level fairness. Individual-level fairness Dwork et al. (2011); Kusner et al. (2017); Kim et al. (2018) asks the treatment of two individuals similar if they are similar according to some distance measure. On the other hand, 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), equalized odds and equality of opportunity (Hardt et al., 2016), and disparate mistreatment (Zafar et al., 2017b); see §3.2 for formal definitions.

Methods achieving the above fairness definitions while optimizing the original objective fall into three main categories:

— 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 embed the representation of data into a new space such that the bias is 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 classical 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 theoretical or practical viability of learning under this model. Prominent noise models include symmetric and class-conditional noise (Angluin & Laird, 1988; Bylander, 1994; Blum et al., 1996; Blum & Mitchell, 1998), constant-partition noise (Decatur, 1997; Ralaivola et al., 2006), and instance-dependent noise (Awasthi et al., 2015). It has also been shown that learning performance degrades as label noise increases (Nettleton et al., 2010).

Class-conditional noise (CCN) has been a particularly well-studied 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 and then use these rates for prediction.

In the important related problem of learning from positive and unlabelled data (PU learning), one has, in lieu of explicit negative samples, a pool of unlabelled data. PU learning in turn has two prominent settings. In the censoring setting (Elkan & Noto, 2008), one imagines first drawing labelled samples, and then placing some fraction of positives in the unlabelled pool. In the case-controlled setting (Ward et al., 2009), on imagines independently drawing samples from the positive and marginal distribution over instances.

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

2.3 Fairness and differential privacy

Recently, Jagielski et al. (2018) explored preserving differential privacy (Dwork, 2006) while at the same time of preserving fairness constraints and achieving good accuracy. 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. Our work differs

from them in five ways: (*1*) rather than modifying specific algorithms, our work can be used before applying a large group of fair algorithms; (*2*) our work focuses on the case of data corruption rather than privacy; (*3*) rather than considering differential privacy, we explore fairness preservation under the framework of MC learning; (*4*) while they only discuss equalized odds, we also deal with demographic parity; (*5*) we verify our theoretical results hold empirically.

3 Background and notation

We recall the settings of standard and fairness-aware binary classification[1], and establish notation used in the sequel. Our notation is summarized in Table 1.

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 find

$$\begin{aligned} f^* &:= \operatorname*{arg\,min}_{f \in \mathcal{F}} L_D(f) \\ L_D(f) &:= \mathbb{E}_{(X,Y)\sim D}[\ell(f(X), Y)]. \end{aligned}$$

In practice, one only observes samples from D, and minimizes the empirical counterpart to L_D.

3.2 Fairness-aware classification

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

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

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

for user-specified *fairness tolerance* $\tau \geq 0$, and *fairness constraint* $\Lambda_D \colon \mathcal{F} \to \mathbb{R}_+$. Such constrained optimization problems can be solved in various ways, e.g., convex relaxations (Donini et al., 2018), alternating minimization (Zafar et al., 2017b), or by considering randomized classifiers to linearize the constraints (Hardt et al., 2016; Menon & Williamson, 2018).

We note here that f is not assumed to be allowed to use A at test time, and thus only takes X as input. This is in line with some of the literature, since it may be forbidden to use A at test time for legal reasons (Lipton et al., 2018). We emphasise that A can of course be used at training time to find an f which satisfies the

[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.

constraint. While it would be interesting to extend the approach to when A is available as input to f as well, this proves challenging when subsequently considering the effect of noise on A.

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

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

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

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 and $\bar{L}_D(f) := \mathbb{E}_{(X,A,Y)\sim D}[\bar{\ell}(f(X), Y)]$. Intuitively, these measure the difference in the average of the fairness loss incurred among the instances with and without the sensitive feature. To make this concrete, observe that if $\bar{\ell}$ is taken to be $\bar{\ell}(s,y) = \mathbb{1}[\mathrm{sign}(s) \neq 1]$ and the 0-1 loss $\bar{\ell}(s,y) = \mathbb{1}[\mathrm{sign}(s) \neq y]$ respectively, then (2) and (3) simplify to

$$\Lambda_D^{\mathrm{DP}}(f) = \left|\mathbb{P}_{D_{0,\cdot}}[f(X) \geq 0] - \mathbb{P}_{D_{1,\cdot}}[f(X) \geq 0]\right|$$

$$\Lambda_D^{\mathrm{EO}}(f) = \left|\mathbb{P}_{D_{0,1}}[f(X) \geq 0] - \mathbb{P}_{D_{1,1}}[f(X) \geq 0]\right|$$

For $\tau = 0$, these constraints correspond to the standard *demographic parity* (Dwork et al., 2011) and *equality of opportunity* (Hardt et al., 2016) constraints. Thus, for general τ, 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).

Sometimes *equalized odds* is required. This can be viewed as a special case under the same conditions as the second constraint above and, simultaneously, of the following additional constraint:

$$\left|\bar{L}_{D_{0,0}}(f) - \bar{L}_{D_{1,0}}(f)\right| \leq \tau$$

Although we do not explicitly consider this fairness scenario, all of the results for the generalized equality of opportunity constraint also apply for this third constraint, and having two constraints simultaneously poses no issue.

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_{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

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

where $\alpha, \beta \in (0,1)$ are (typically unknown) noise parameters with $\alpha + \beta < 1$. Further, the corrupted base rate $\pi_{\mathrm{corr}} := \mathbb{P}[Y_{\mathrm{corr}} = 1]$ is arbitrary, and may in general have no relationship to the clean base rate $\pi := \mathbb{P}[Y = 1]$.

The MC learning framework captures as special cases learning from class-conditional noise, as well as learning from positive and unlabelled data (Scott et al., 2013b; Menon et al., 2015); thus, it is an appealing model of noise.

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

Table 1: Glossary of commonly used symbols

Symbol	Meaning	Symbol	Meaning
X	instance	D_{corr}	corrupted distribution D
A	sensitive feature	f	score function $f : \mathcal{X} \to \mathbb{R}$
Y	target feature	ℓ	accuracy loss $\ell : \mathbb{R} \times \{0,1\} \to \mathbb{R}_+$
D	distribution $\mathbb{P}(X, A, Y)$	L_D	expected accuracy loss on D
$D_{a,\cdot}$	distribution $\mathbb{P}(X, A, Y \mid A = a)$	$\bar{\ell}$	fairness loss $\bar{\ell} : \mathbb{R} \times \{0,1\} \to \mathbb{R}_+$
$D_{\cdot,y}$	distribution $\mathbb{P}(X, A, Y \mid Y = y)$	$\bar{L}_D$	expected fairness loss on D
$D_{a,y}$	distribution $\mathbb{P}(X, A, Y \mid A = a, Y = y)$	Λ_D	fairness constraint

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. In other words, we show that it so happens that fairness constraints are naturally robust to noise in the sensitive attribute. 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 choose MC learning as our noise model. Our choice is motivated by the fact that MC learning captures both PU and CCN learning as special cases; hence, we automatically obtain results for both of 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), given some (typically unknown) noise parameters $\alpha, \beta \in (0,1)$ with $\alpha + \beta < 1$, we assume that the corrupted class-conditional distributions are given by:

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

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 π_a). In words, 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 is instead such that:

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

for noise parameters $\alpha', \beta' \in (0,1)$. As shown by the following lemma, this is not a different noise assumption but simply a direct implication of (5).

Lemma 1. *If we assume that there is noise in the sensitive attribute only, as given in Equation (5) then Equation (6) holds with*

$$\alpha' = \frac{\alpha \mathbb{P}[Y = 1 \mid A = 0]}{(1-\alpha)\mathbb{P}[Y = 1 \mid A = 1] + \alpha \mathbb{P}[Y = 1 \mid A = 0]}$$

and

$$\beta' = \frac{\beta \mathbb{P}[Y = 1 \mid A = 1]}{(1-\beta)\mathbb{P}[Y = 1 \mid A = 0] + \beta \mathbb{P}[Y = 1 \mid A = 1]}$$

Proof. Suppose that we have noise as given by Equation (5). We denote by A the random variable denoting the value of the true sensitive attribute and by A_{corr} the random variable denoting the value of the corrupted sensitive attribute.

Then, for any measurable subset of instances U,

$$\begin{aligned}
&\mathbb{P}[X \in U \mid Y = 1, A_{\text{corr}} = 1] \\
&= \frac{\mathbb{P}[X \in U, Y = 1 \mid A_{\text{corr}} = 1]}{\mathbb{P}[Y = 1 \mid A_{\text{corr}} = 1]} \\
&= \frac{\mathbb{P}[X \in U, Y = 1 \mid A_{\text{corr}} = 1]}{(1-\alpha)\mathbb{P}[Y = 1 \mid A = 1] + \alpha\mathbb{P}[Y = 1 \mid A = 0]} \\
&= \frac{(1-\alpha)\mathbb{P}[X \in U, Y = 1 \mid A = 1]}{(1-\alpha)\mathbb{P}[Y = 1 \mid A = 1] + \alpha\mathbb{P}[Y = 1 \mid A = 0]} + \frac{\alpha\mathbb{P}[X \in U, Y = 1 \mid A = 0]}{(1-\alpha)\mathbb{P}[Y = 1 \mid A = 1] + \alpha\mathbb{P}[Y = 1 \mid A = 0]} \\
&= \frac{(1-\alpha)\mathbb{P}[Y = 1 \mid A = 1]\mathbb{P}[X \in U \mid Y = 1, A = 1]}{(1-\alpha)\mathbb{P}[Y = 1 \mid A = 1] + \alpha\mathbb{P}[Y = 1 \mid A = 0]} + \frac{\alpha\mathbb{P}[Y = 1 \mid A = 0]\mathbb{P}[X \in U \mid Y = 1, A = 0]}{(1-\alpha)\mathbb{P}[Y = 1 \mid A = 1] + \alpha\mathbb{P}[Y = 1 \mid A = 0]} \\
&= (1-\alpha')\mathbb{P}[X \in U \mid Y = 1, A = 1] + \alpha'\mathbb{P}[X \in U \mid Y = 1, A = 0],
\end{aligned}$$

where in the last equality we set

$$\alpha' := \frac{\alpha\mathbb{P}[Y = 1 \mid A = 0]}{(1-\alpha)\mathbb{P}[Y = 1 \mid A = 1] + \alpha\mathbb{P}[Y = 1 \mid A = 0]}.$$

Note that the last equality is equivalent to the first equality of Equation (6) with α' as in the lemma.

The proof for β' is exactly the same and simply expands $\mathbb{P}[X \in U \mid Y = 1, A_{\text{corr}} = 0]$ instead of $\mathbb{P}[X \in U \mid Y = 1, A_{\text{corr}} = 1]$. □

One thing to notice is that $\alpha = \alpha'$ when $Y \perp\!\!\!\perp A$, i.e., $\mathbb{P}[Y = 1|A = 0] = \mathbb{P}[Y = 1|A = 0]$. However, in practice this will rarely be the case as such a condition makes achieving fairness trivial (it indicates that there is no intrinsic unfairness).

Furthermore, although the lemma shows that (5) implies (6) and gives a way to calculate α' and β' from α and β, 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 α' and β' is required. Rather than estimating all of $\alpha, \beta, \mathbb{P}[Y = 1|A = 1]$, and $\mathbb{P}[Y = 1|A = 0]$ it is often much easier to estimate α' and β' directly (which can be done in the same way as estimating α and β simply by considering $D_{\cdot,1,\text{corr}}$ rather than D_{corr}).

4.2 Fairness constraints under MC learning

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

Theorem 1. *Assume that we have noise as described above by Equation (5). Then, for the DP-like constraints,*

$$\Lambda_D^{\text{DP}}(f) \leq \tau \iff \Lambda_{D_{\text{corr}}}^{\text{DP}}(f) \leq \tau \cdot (1 - \alpha - \beta).$$

Additionally, for the EO-like constraints,

$$\Lambda_{D_{\cdot,1}}^{\text{EO}}(f) \leq \tau \iff \Lambda_{D_{\text{corr},\cdot,1}}^{\text{EO}}(f) \leq \tau \cdot (1 - \alpha' - \beta'),$$

where α' and β' are as per Equation (6) and Lemma 1.

Proof. For the DP-like constraints simply note that by definition of D_{corr} we have that

$$\bar{L}_{D_{0,\cdot,\text{corr}}}(f) = (1-\beta) \cdot \bar{L}_{D_{0,\cdot}}(f) + \beta \cdot \bar{L}_{D_{1,\cdot}}(f)$$

and similarly,

$$\bar{L}_{D_{1,\cdot,\text{corr}}}(f) = (1-\alpha)\cdot\bar{L}_{D_{1,\cdot}}(f) + \alpha\cdot\bar{L}_{D_{0,\cdot}}(f)$$

Thus we have that

$$\bar{L}_{D_{0,\cdot,\text{corr}}}(f) - \bar{L}_{D_{1,\cdot,\text{corr}}}(f) = (1-\alpha-\beta)\cdot(\bar{L}_{D_{0,\cdot}}(f) - \bar{L}_{D_{1,\cdot}}(f)),$$

which immediately implies the desired result.

The result for the EO constraint is obtained in the exact same way by simply replacing $D_{a,\cdot}$ with $D_{a,1}$, $D_{a,\cdot,\text{corr}}$ with $D_{a,1,\text{corr}}$, and α and β with α' and β'. □

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 the quantity

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

for functions $h_0, h_1 \colon \mathcal{Z} \to \mathbb{R}$. The quantity B_D is a generalised form of balanced error, being the average of false positive and negative rates of a classifier. Then, if $h_0(z) + h_1(z) = 0$, one may show (van Rooyen, 2015, Theorem 4.16) (see also (Blum & Mitchell, 1998; Zhang & Lee, 2008; Menon et al., 2015) that

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

where D_{corr} refers to a corrupted version of D under MC learning with noise parameters α, β. In words, the effect of MC noise on B_D is simply to perform a scaling.

In our context, we may apply the above result with Z corresponding to $X \times Y$, W corresponding to the sensitive feature A, and

$$h_0((x,y)) = +\bar{\ell}(y, f(x))$$
$$h_1((x,y)) = -\bar{\ell}(y, f(x)).$$

With these choices, $B_D = \bar{L}_D(f)$. Thus, (7) immediately implies $\bar{L}_D(f) = (1-\alpha-\beta)\cdot\bar{L}_{D_{\text{corr}}}(f)$, which in turn implies Theorem 1.

4.3 Algorithmic implications

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

Unsurprisingly, τ' depends on the noise parameters α, β. 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 α, β, and plug these in to construct an estimate of τ'.

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 τ 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 α, β.

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** $\mathsf{FairAlg}(S, \mathcal{F}, \Lambda, \tau')$

5 Experiments

We present experimental results validating the preceding theory. Specifically, we demonstrate via two case-studied that it is viable to learn fair classifiers despite the presence of noise in the sensitive feature.

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 section 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 A 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 (τ), i.e., for small enough values of τ 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 τ 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

While our results apply for any instantiation of the MC learning noise framework, in practice this general kind of noise is rarely observed. Instead, our case studies focus on two very common special cases of MC learning: CCN noise and the PU setting.

Under CCN noise the sensitive feature's value is randomly flipped with probability ρ^+ if its value was 1 or with probability ρ^- if its value was 0. As shown in Menon et al. (2015, Appendix C), CCN noise is a special case of MC learning with:

$$\begin{aligned} \pi_{a,\text{corr}} &= (1 - \rho^+) \cdot \pi_a + \rho^- \cdot (1 - \pi_a) \\ \alpha &= \pi_{a,\text{corr}}^{-1} \cdot (1 - \pi_a) \cdot \rho^- \\ \beta &= (1 - \pi_{a,\text{corr}})^{-1} \cdot \pi_a \cdot \rho^+ \end{aligned}$$

Meanwhile, for PU learning we consider the censoring setting (Elkan & Noto, 2008) which is simply a special case of CCN learning where one of ρ^+ and ρ^- is 0. While our results also apply to the slightly different case-controlled setting of PU learning (Ward et al., 2009) the former setting is slightly more natural in our context.

Consequently, in both case studies we only need to estimate ρ^+ and ρ^- (as well as $\pi_{a,\text{corr}}$, which is trivial) to be able to calculate the scaling coefficient $1 - \alpha - \beta$. Since noise parameter estimation is not the aim of our paper, we assume that the values of ρ^+ and ρ^- are known.

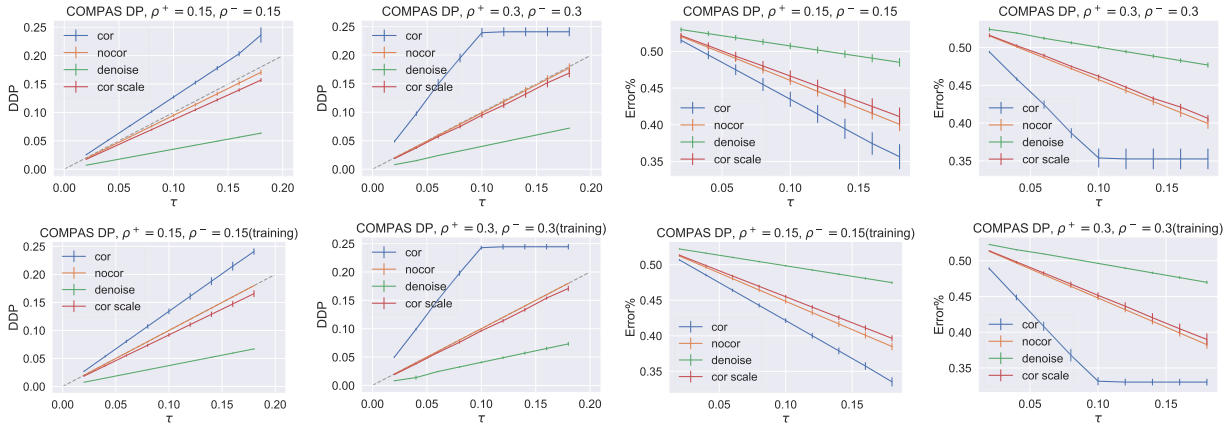

Figure 1: Relationship between input τ and fairness violation/error on the `COMPAS` dataset using DP constraint (both training and testing curves are included). The black dotted line corresponds is the $y = x$ line and represents the ideal fairness violation.

5.2 Benchmarks

For each case study we compare our method (termed cor scale), which scales the input parameter τ using 1 and the known values of ρ^+ and ρ^- which we provide and then uses the fair classifier to perform classification, 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 ρ^+ and ρ^- that we use to apply our scaling so this is a fair comparison to our method.

For both case studies, we compare the relationship between the input parameter τ and the training error, testing error, and fairness violation for all three benchmarks and our method. For simplicity, we only consider the DP constraint.

5.3 Case study: privacy preservation

When data gets released it is often necessary to protect the identities of those that the data pertains to. This is especially true for medical data or data from the social sciences. While this concern pervades all areas of machine learning, it is particularly important in the context of fair learning since those same datasets are often the ones where fairness is a major concern. While anonymization is sometimes enough, there are cases where stronger steps have to be taken in order to protect the participant's privacy.

Surprisingly, our result immediately gives an extraordinarily simple and relatively effective method to protect privacy while still allowing for fair classification. Indeed, consider simply adding CCN noise to the sensitive attribute (and perhaps other features) of the of dataset instances. While this does not protect against multiple stronger privacy attacks, it does give some assurance that an instance's true features (and thus hopefully its identity) cannot be recovered with 100% certainty. Meanwhile our result implies that fair classification would still be possible. Furthermore, it gives a way to achieve fair classification by simply scaling τ as per Theorem

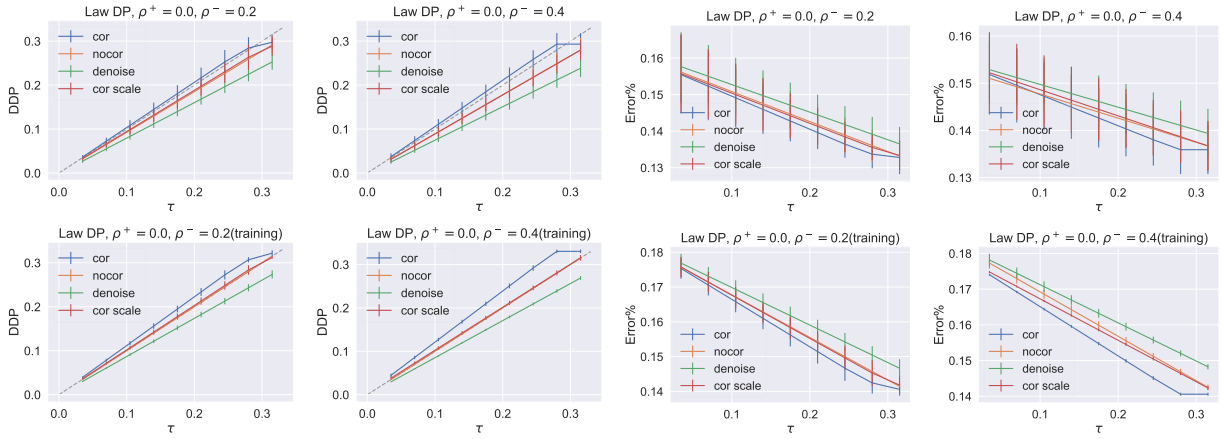

Figure 2: Relationship between input τ and fairness violation/error on the `law school` dataset using DP constraint (both training and testing curves are included). The black dotted line corresponds is the $y = x$ line and represents the ideal fairness violation. Note that in some of the graphs, the red line and the orange line perfectly overlap with each other.

1 before calling any of a large group of downstream fair classifiers with several options as to which group fairness constraint to use.

Specifically, in this case study, we look at `COMPAS`: a dataset from Propublica (Angwin et al., 2016) which is one of the most widely used datasets in the fairness and machine learning community. 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 add CCN noise with $\rho^+ = \rho^- \in \{0.15, 0.3\}$ to the sensitive attribute. We try performing fair classification on this noisy data using our method and compare the results to the three benchmarks described above (we apply the DP fairness constraint).

Figure 1 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. It can be seen from the graph that as expected the naïve method cor violates the fairness constraint while our method cor scale reaches a degree of fairness violation which is approximately the maximum amount allowed (as shown by black dotted line). This is both expected and ideal, and it matches what happens when there is no noise nocor. We can also see that when compared with the ideal noiseless nocor method, the accuracy of our method suffers slightly. This is expected as noise will inevitably lead to some loss of information. Although the method denoise seems to also achieve the desired fairness level, it has large fluctuations and sacrifices much more accuracy than our method.

We also tried using the EO constraint as well as other datasets. The results obtained showed the same trends that are observable in Figure 1 and are included in Appendix B for completeness.

5.4 Case study: PU learning

There are many real world cases in which a PU setting for the sensitive attribute is very realistic. Indeed, one may be able to clearly identify some instances of the minority group while being unable to tell whether the rest of the instances were in the majority or minority groups. In this case study, we consider the law 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.

While in our case we have access to 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 lie about their group membership due to fears of discrimination. Note that there would be no similar incentive if belonging to the majority group. Thus any individual identified as belonging to the minority 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 majority group (these would therefore be part of the unlabelled instances).

To model such a scenario we added CCN noise to the dataset with $\rho^+ = 0$ and $\rho^- \in \{0.2, 0.4\}$. Again we try performing fair classification on this noisy dataset with DP as the constraint. We use our method as well as the three benchmarks described above. Figure 2 shows the average result over three runs each with a random 80-20 training-testing split. As before, it can be seen from the graph that the naive method cor violates the fairness constraint while our method cor scale reaches a degree of fairness violation which very close to the level reached by the fair classifier when there is no noise (nocor). Although the method denoise seems to also achieve the desired fairness level, it does so by sacrificing much more accuracy than our method.

These results are consistent with our derivation and show that our method cor scale helps achieve the desired degree of fairness while minimizing loss of accuracy.

We tried running similar experiments on other datasets and observed the same trend. These results are included in Appendix C for completeness.

6 Conclusion

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 τ. Our method can be applied before any downstream classifiers (that takes in a parameter τ that is proportional to mean-difference score of some fairness measures). We also provide applications in PU-learning and privacy preservation.

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). Further, considering the case of handling noisy A when the latter is used as input to the classifier would also be of interest.

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., Dudk, 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.

Blum, A., Frieze, A., Kannan, R., and Vempala, S. A polynomial-time algorithm for learning noisy linear threshold functions. In *Foundations of Computer Science (FOCS)*, pp. 330–338, Oct 1996.

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.

Bylander, T. Learning linear threshold functions in the presence of classification noise. In *Conference on Learning Theory (COLT)*, pp. 340–347, 1994.

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.

Decatur, S. E. PAC learning with constant-partition classification noise and applications to decision tree induction. In *International Conference on Machine Learning (ICML)*, pp. 83–91, 1997.

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., Ferrari, C., Gummadi, K. P., and Krause, A. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems 31*, pp. 1273–1283, 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.

Menon, A. K. M. and Williamson, R. C. The cost of fairness in classification. In *Conference on Fairness, Accountability, and Transparency*, 2018.

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

Nettleton, D. F., Orriols-Puig, A., and Fornells, A. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, 33(4):275–306, Apr 2010. ISSN 1573-7462.

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.

Ralaivola, L., Denis, F., and Magnan, C. N. CN = CPCN. In *International Conference on Machine Learning (ICML)*, pp. 721–728, 2006.

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. 489511, 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.

Zhang, J. and Bareinboim, E. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems 31*, pp. 3675–3685, 2018.

Appendices

A Relationship between Mean-difference score and the constraint used in Agarwal et al. (2018)

Agarwal et al. (2018) adopts slightly different fairness constraints than ours. Using our notation and letting $c_f(X) = \text{sign}(f(X))$, instead of bounding $\Lambda_D^{\text{DP}}(f)$ by τ, they bound

$$\max_{a\in\{0,1\}} \left|\mathbb{E}_{D_{a,\cdot}}[c_f(X)] - \mathbb{E}_D[c_f(X)]\right|$$

and

$$\max_{a\in\{0,1\}} \left|\mathbb{E}_{D_{a,1}}[c_f(X)] - \mathbb{E}_{D_{\cdot,1}}[c_f(X)]\right|$$

for DP and EO respectively by τ. The two have the following relationship.

Theorem 2. *Under the setting of fair binary classification with a single binary sensitive attribute and using* $\bar{\ell}(s,y) = \mathbb{1}[\text{sign}(s)]$ *we have that*

$$\max_{a\in\{0,1\}} \left|\mathbb{E}_{D_{a,\cdot}}[c_f(X)] - \mathbb{E}_D[c_f(X)]\right| = \max_{a\in\{0,1\}} (\mathbb{P}[A=0], \mathbb{P}[A=1])\Lambda_D^{\text{DP}}(f)$$

and

$$\max_{a\in\{0,1\}} \left|\mathbb{E}_{D_{a,1}}[c_f(X)] - \mathbb{E}_{D_{\cdot,1}}[c_f(X)]\right| = \max_{a\in\{0,1\}} (\mathbb{P}[A=0 \mid Y=1], \mathbb{P}[A=1 \mid Y=1])\Lambda_D^{\text{EO}}(f)$$

Proof. For the DP case,

$$\begin{aligned}
&\left|\mathbb{E}_{D_{1,\cdot}}[c_f(X)] - \mathbb{E}_D[c_f(X)]\right| \\
&= \left|\mathbb{E}_{D_{1,\cdot}}[c_f(X)] - (\mathbb{P}[A=1]\mathbb{E}_{D_{1,\cdot}}[c_f(X)] + \mathbb{P}[A=0]\mathbb{E}_{D_{0,\cdot}}[c_f(X)])\right| \\
&= \left|(1-\mathbb{P}[A=1])\mathbb{E}_{D_{1,\cdot}}[c_f(X)] - \mathbb{P}[A=0]\mathbb{E}_{D_{0,\cdot}}[c_f(X)]\right| \\
&= \left|\mathbb{P}[A=0]\mathbb{E}_{D_{1,\cdot}}[c_f(X)] - \mathbb{P}[A=0]\mathbb{E}_{D_{0,\cdot}}[c_f(X)]\right| \\
&= \mathbb{P}[A=0]\left|(\mathbb{E}_{D_{1,\cdot}}[c_f(X)] - \mathbb{E}_{D_{0,\cdot}}[c_f(X)])\right| \\
&= \mathbb{P}[A=0]\left|\bar{L}_{D_{0,\cdot}}(f) - \bar{L}_{D_{1,\cdot}}(f)\right| \\
&= \mathbb{P}[A=0]\Lambda_D^{\text{DP}}(f)
\end{aligned}$$

and similarly

$$\left|\mathbb{E}_{D_{0,\cdot}}[c_f(X)] - \mathbb{E}_D[c_f(X)]\right| = \mathbb{P}[A=1]\Lambda_D^{\text{DP}}(f)$$

so the theorem holds.

The result for the EO case is proved in exactly the same way by simply replacing $\mathbb{P}[A=0], \mathbb{P}[A=1]$, $D_{a,\cdot}$ and D with $\mathbb{P}[A=0 \mid Y=1], \mathbb{P}[A=1 \mid Y=1]$, $D_{a,1}$ and $D_{\cdot,1}$ respectively.

□

We then have the following as an immediate corollary.

Corollary 1. *Assuming that we have noise as described above by Equation (5) and that we take* $\bar{\ell}(s,y) = \mathbb{1}[\text{sign}(s)]$ *then we have that if* $\max_{a\in\{0,1\}}(\mathbb{P}_D[A=0], \mathbb{P}_D[A=1]) = \max_{a\in\{0,1\}}(\mathbb{P}_{D_{\text{corr}}}[A=0], \mathbb{P}_{D_{\text{corr}}}[A=1])$ *then:*

$$\max_{a\in\{0,\cdot\}} \left|\mathbb{E}_{D_{a,\cdot}}[c_f(X)] - \mathbb{E}_D[c_f(X)]\right| < \tau \iff \max_{a\in\{0,1\}} \left|\mathbb{E}_{D_{a,\cdot,corr}}[c_f(X)] - \mathbb{E}_{D_{corr}}[c_f(X)]\right| < \tau \cdot (1-\alpha-\beta).$$

And if $\max_{a\in\{0,1\}}(\mathbb{P}_{D_{\cdot,1}}[A=0],\mathbb{P}_{D_{\cdot,1}}[A=1]) = \max_{a\in\{0,1\}}(\mathbb{P}_{D_{\cdot,1,\text{corr}}}[A=0],\mathbb{P}_{D_{\cdot,1,\text{corr}}}[A=1])$ *then:*

$$\max_{a\in\{0,1\}}\left|\mathbb{E}_{D_{a,1}}[c_f(X)]-\mathbb{E}_{D_{\cdot,1}}[c_f(X)]\right|<\tau \iff \max_{a\in\{0,1\}}\left|\mathbb{E}_{D_{a,1,corr}}[c_f(X)]-\mathbb{E}_{D_{\cdot,1,corr}}[c_f(X)]\right|<\tau\cdot(1-\alpha'-\beta').$$

Even if the noise does not satisfy these new assumptions, we can still bound the constraint. Note that both $\max_{a\in\{0,1\}}(\mathbb{P}[A=0],\mathbb{P}[A=1])$ and $\max_{a\in\{0,1\}}(\mathbb{P}[A=0\mid Y=1],\mathbb{P}[A=1\mid Y=1])$ have values between 0.5 and 1. Thus,

$$\frac{1}{2}\Lambda_D^{\text{DP}}(f)\le\max_{a\in\{0,1\}}\left|\mathbb{E}_{D_{a,\cdot}}[c_f(X)]-\mathbb{E}_D[c_f(X)]\right|\le\Lambda_D^{\text{DP}}(f)$$

$$\frac{1}{2}\Lambda_D^{\text{EO}}(f)\le\max_{a\in\{0,1\}}\left|\mathbb{E}_{D_{a,1}}[c_f(X)]-\mathbb{E}_{D_{\cdot,1}}[c_f(X)]\right|\le\Lambda_D^{\text{EO}}(f),$$

and therefore the following corollary holds:

Corollary 2. *Assuming that we have noise as described above by Equation (5) and that we take* $\bar{\ell}(s,y)=\mathbb{1}[\text{sign}(s)]$ *then we have that:*

$$\max_{a\in\{0,1\}}\left|\mathbb{E}_{D_{a,\cdot,corr}}[c_f(X)]-\mathbb{E}_{D_{corr}}[c_f(X)]\right|<\frac{1}{2}\tau\cdot(1-\alpha-\beta)\Rightarrow\max_{a\in\{0,1\}}\left|\mathbb{E}_{D_{a,\cdot}}[c_f(X)]-\mathbb{E}_D[c_f(X)]\right|<\tau$$

and,

$$\max_{a\in\{0,1\}}\left|\mathbb{E}_{D_{a,1,corr}}[c_f(X)]-\mathbb{E}_{D_{\cdot,1,corr}}[c_f(X)]\right|<\frac{1}{2}\tau\cdot(1-\alpha'-\beta')\Rightarrow\max_{a\in\{0,1\}}\left|\mathbb{E}_{D_{a,1}}[c_f(X)]-\mathbb{E}_{D_{\cdot,1}}[c_f(X)]\right|<\tau.$$

In addition to giving a simple way to use the classifier of Agarwal et al. (2018) without any modification, these results seem to indicate that with small modifications our scaling method can apply to an even wider range of fair classifiers than formally shown.

B More results for the privacy case study

In this section we give some additional results for the privacy case study. First, 3 shows the results under the EO constraint for the `COMPAS` dataset. I.e. the dataset and setting is the same as described in section 5.3 but with the EO constraint instead of the DP constraint. We see that the trends are the same.

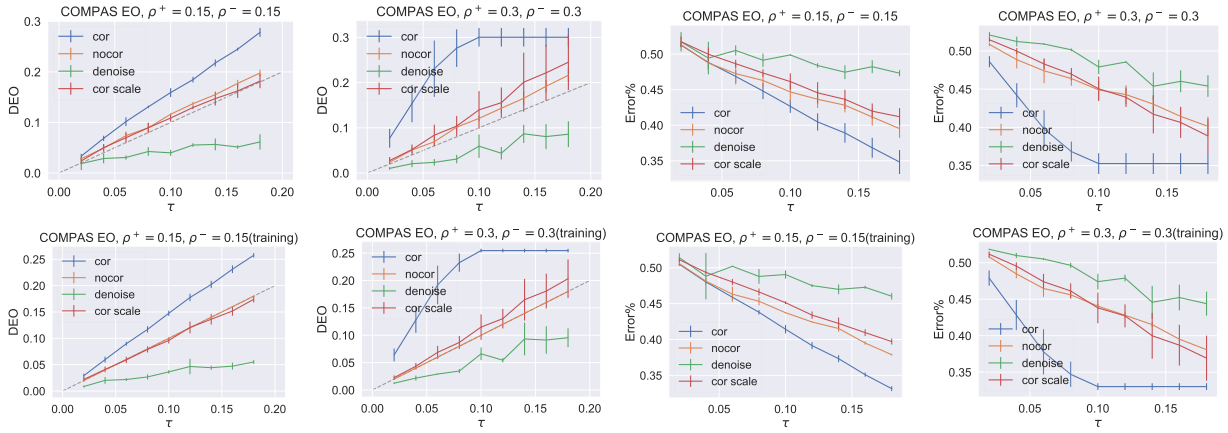

Figure 3: (EO)(testing and training) Relationship between input τ and fairness violation/error on the `COMPAS` dataset.

Figures 4 and Figure 5 show results on the `bank` dataset (Ban) with the DP and EO constraints respectively. This dataset is a subset of the original Bank Marketing dataset from the UCI repository (Dheeru & Karra Taniskidou, 2017). The task is to predict if a client subscribes a term deposit. The sensitive attribute is if a person is middle aged(i.e. has an age between 25 and 60). The data comprises 11162 examples and 17 features. Again we note that the trends are the same.

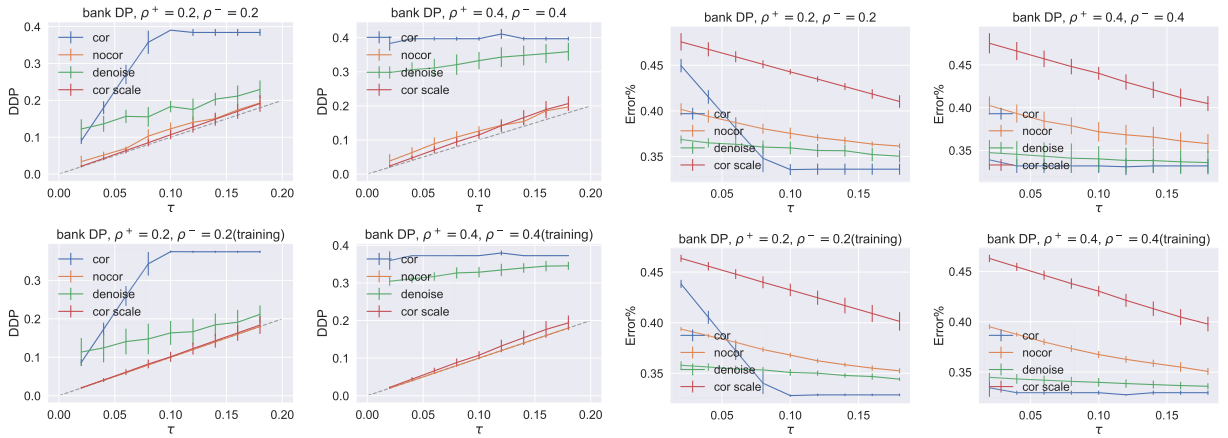

Figure 4: (DP)(testing and training) Relationship between input τ and fairness violation/error on the `bank` dataset.

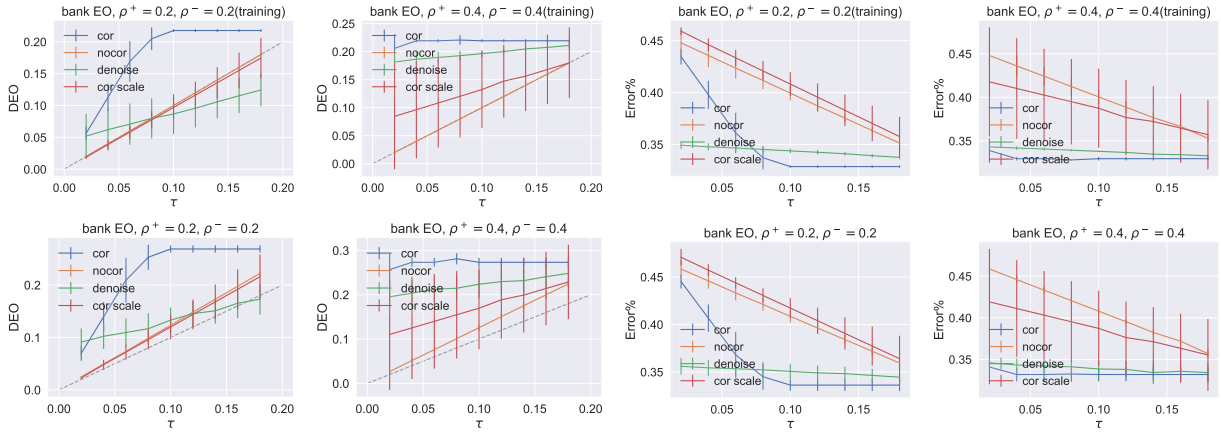

Figure 5: (EO)(testing and training) Relationship between input τ and fairness violation/error on the `bank` dataset.

C More results for the PU case study

In this section we give some additional results for the PU case study. Figure 6 shows the results under PU noise on the `german` dataset, which is another dataset from the UCI repository (Dheeru & Karra Taniskidou, 2017). The task is to predict if one has good credit and the sensitive attribute is whether a person is foreign. The data comprises 1000 examples and 20 features. The trends are similar to those for the `law school` dataset.

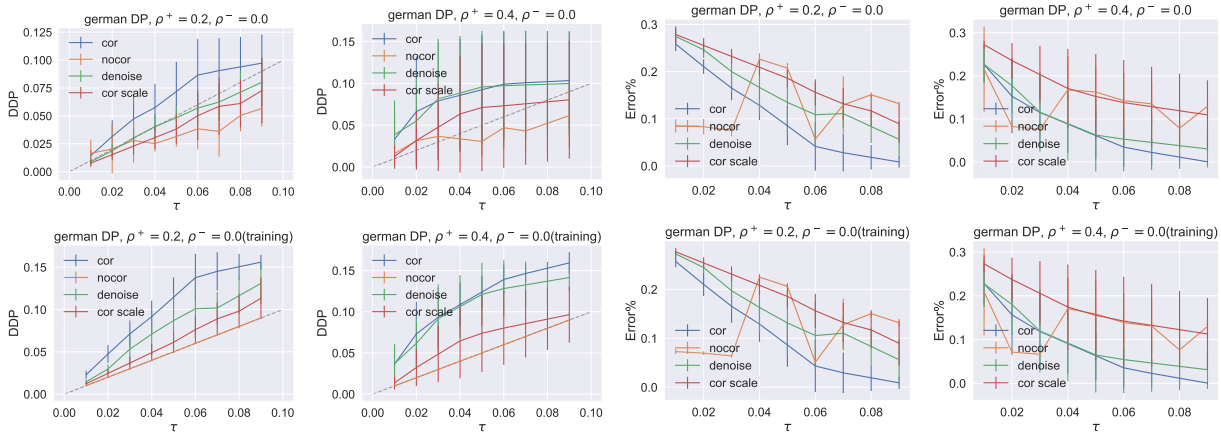

Figure 6: (DP)(training and testing) Relationship between input τ and fairness violation/error on the `german` dataset.

D The influence of different noise level

Figure 7 explores the influence of the noise level on the trends and relationships between our method's performance and that of the benchmarks. We run these experiments on the UCI `adult` dataset, which is another dataset from the UCI repository (Dheeru & Karra Taniskidou, 2017). The task is to predict if one has income more than 50K and gender is the sensitive attribute. The data comprises 48842 examples and 14 features. We run these experiments with the DP constraint under different CCN noise levels ($\rho^+ = \rho^- \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.48\}$). We include both training and testing curves for completeness. As

we can see, as the noise increases the gap between the corrupted data curves and the uncorrupted data curve increases. It becomes very hard to get close to the non-corrupted case when noise becomes too high.

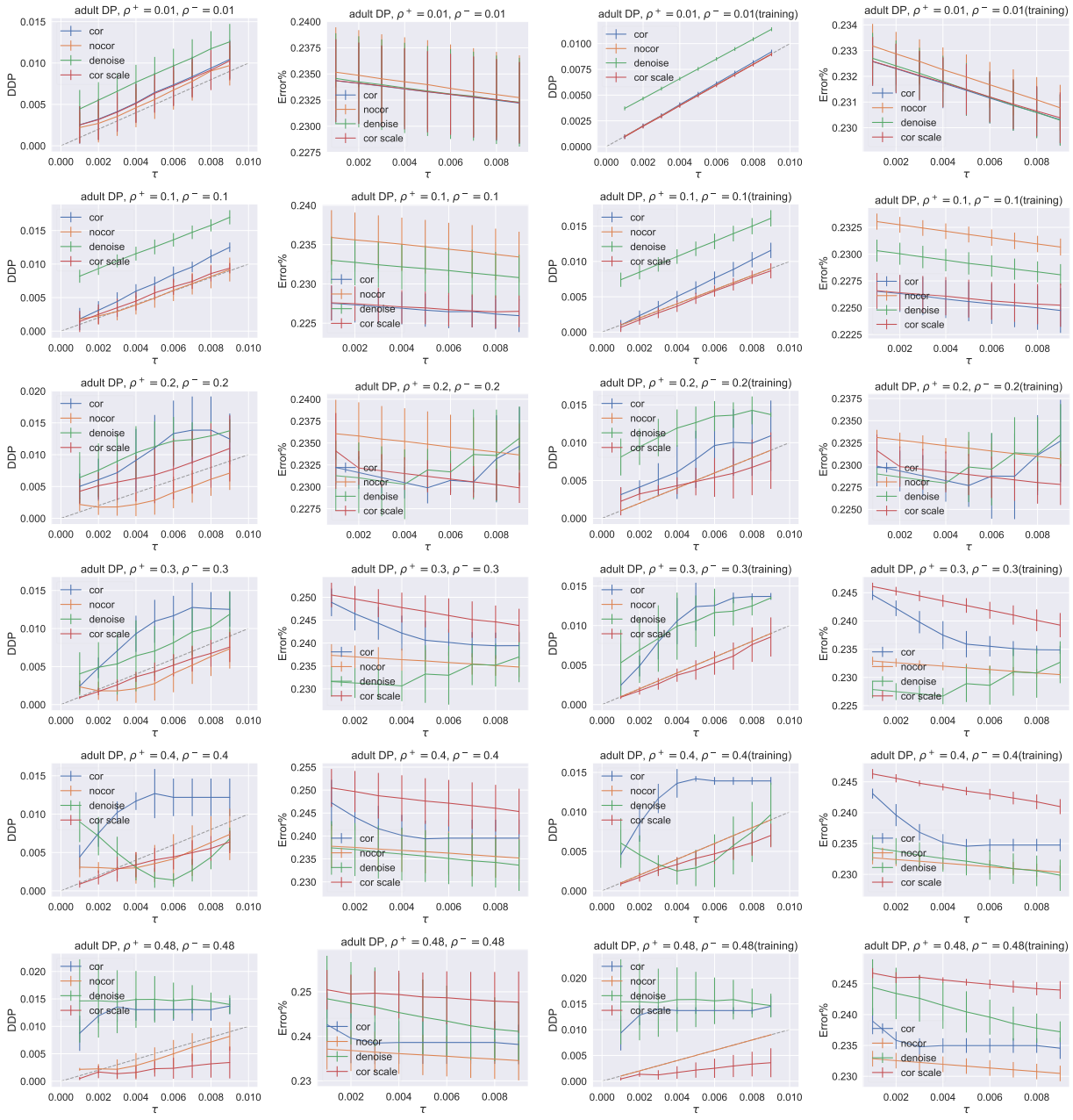

Figure 7: Relationship between input τ and fairness violation/error on the `adult` dataset for various noise levels. From left to right: testing fairness violation, testing error, training fairness violation, and training error. Different noise levels from top to bottom.