Fair Classification with Group-Dependent Label Noise

Jialu Wang
faldict@ucsc.edu
UC Santa Cruz
Santa Cruz, CA, USA

Yang Liu
yangliu@ucsc.edu
UC Santa Cruz
Santa Cruz, CA, USA

Caleb Levy
cclevy@ucsc.edu
UC Santa Cruz
Santa Cruz, CA, USA

ABSTRACT

This work examines how to train fair classifiers in settings where training labels are corrupted with random noise, and where the error rates of corruption depend both on the label class and on the membership function for a protected subgroup. Heterogeneous label noise models systematic biases towards particular groups when generating annotations. We begin by presenting analytical results which show that naively imposing parity constraints on demographic disparity measures, without accounting for heterogeneous and group-dependent error rates, can decrease both the accuracy *and* the fairness of the resulting classifier. Our experiments demonstrate these issues arise in practice as well. We address these problems by performing empirical risk minimization with carefully defined surrogate loss functions and surrogate constraints that help avoid the pitfalls introduced by heterogeneous label noise. We provide both theoretical and empirical justifications for the efficacy of our methods. We view our results as an important example of how imposing fairness on biased data sets without proper care can do at least as much harm as it does good.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; **Philosophical/theoretical foundations of artificial intelligence**.

KEYWORDS

machine learning, algorithmic fairness, learning with noisy and biased labels

ACM Reference Format:
Jialu Wang, Yang Liu, and Caleb Levy. 2021. Fair Classification with Group-Dependent Label Noise. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442188.3445915

1 INTRODUCTION

Recent work shows that machine learning classifiers can perpetuate and amplify existing systemic injustices in society. Notable examples include discrepancies in allocation of medical care to patients on the basis of race [36] and significant disparities in predicting recidivism rates for African-American defendants [4, 13], and more [9, 38, 42]. A number of techniques have been developed in order to mitigate bias in machine learning classifiers [1, 10, 17, 22, 33, 45]. Typically, these methods consider populations with groups corresponding to a set of protected sensitive attributes, such as race or gender. The classifier is then required to exhibit similar behavior across all groups [13, 22, 24, 45]. This can be done by imposing equality of true positive rate or true negative rate conditioned on group membership. These are called "fairness," or parity constraints.

Many of these methods assume the availability of clean and accurate labels. However, this is often not the case. In fact, bias in data is particularly pertinent to label corruption. To make things worse, the accuracy of available labels is often strongly influenced by whether a person falls within a protected group, and these discrepancies can have significant and often life-altering outcomes. For example, it has been shown that labels for criminal activity generated via crowdsourcing are systematically biased against certain racial groups [14]. As another example, both women and lower-income individuals often receive significantly less accurate diagnoses for cancer and other ailments than men, due to imbalance in the sample population of medical trials [20], and due to bias from doctor treatment [8]. Similar discrepancies arise in the accuracy of mathematical aptitude evaluations for males and females in primary school [27], and it has long been known that an employer's evaluation of a resume will be influenced by the perceived ethnic origin of an applicant's name [6]. Moreover, studies show that people of all races use and sell illegal drugs at remarkably similar rates, but in some states, black male have been admitted to prison on drug charges at rates twenty to fifty times greater than those of white men [2].

The structure and magnitude of group-specific label noise can dramatically affect the performance *and* fairness of a classifier. To see this, we consider the following examples.

Example 1. Enforcing fairness constraints without accounting for group-specific label noise can harm the accuracy of the classifier for the group whose labels have been accurate recorded.

Consider training classifiers using data from two groups $z \in \{A, B\}$ with homogeneous data distributions $\mathbb{P}(Y = +1|X = \boldsymbol{x}, z = A) = \mathbb{P}(Y = +1|X = \boldsymbol{x}, z = B)$, where $\boldsymbol{x} = [x_1, x_2]$, a 2-dimensional feature vector. In this setting, the Bayes-optimal classifiers for A and B (denoted as f^*_A and f^*_B respectively) will obey any parity constraint. However, suppose group A has a set of clean labels, while group B has clean labels when the ground truth is $y = +1$ but there is a 70% chance that corrupting noise will cause the observed label to be flipped from the true value when $y = -1$. In this case, f^*_{fair} trained on both groups achieves perceived equal True Positive Rates (TPR) (50%) between the two groups and is the best one to do so - this indeed hurts group A's prediction performance (as opposed to 100% accuracy before), but the labels in group A are not affected by noise. Although [7] also considers this single-group noise setting and shows that fairness interventions could aid in reducing the

FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445915

error caused by label bias, our observation demonstrates a special case where potential harm occurs.

Example 1: A simple classification problem to illustrate the possibility of harming the clean group when training a fair True Positive Rate (TPR) model over a set of noisy labels.

	Group A			Group B			Pooled		
$(x_1, x_2), y$	$+1$	-1	f_A^*	$+1$	-1	f_B^*	$+1$	-1	f_{fair}^*
$(0,0),\ -1$	0	25	-1	70	30	$+1$	70	55	$+1$
$(0,1),\ -1$	0	25	-1	70	30	$+1$	70	55	-1
$(1,0),\ +1$	25	0	$+1$	100	0	$+1$	125	0	$+1$
$(1,1),\ +1$	25	0	$+1$	100	0	$+1$	125	0	-1

Example 2: A simple classification problem to illustrate the possibility of wrongly perceived fairness due to training on noisy labels.

	Group A			Group B			Pooled		
$(x_1, x_2), (y_A, y_B)$	$+1$	-1	f_A^*	$+1$	-1	f_B^*	$+1$	-1	f_{fair}^*
$(0,0),\ (-1,-1)$	0	100	-1	75	225	-1	75	325	-1
$(0,1),\ (-1,+1)$	0	100	-1	75	25	$+1$	75	125	$+1$
$(1,0),\ (+1,+1)$	100	0	$+1$	75	25	$+1$	175	25	-1
$(1,1),\ (+1,+1)$	100	0	$+1$	75	25	$+1$	175	25	$+1$

Example 2. A classifier may appear to achieve parity when it does not. Furthermore, imposing a parity constraint might actually make everyone worse off.

Consider training classifiers using data from two groups $z \in \{A, B\}$ with heterogeneous data distributions $\mathbb{P}(Y = +1 \mid X = \boldsymbol{x}, z = A) = \mathbb{P}(Y = +1 \mid X = \boldsymbol{x}, z = B)$. Suppose group A has a set of clean labels, while one quarter of group B's labels are incorrect. We denote the Bayes-optimal classifiers for A and B as f_A^* and f_B^* respectively and they obey any parity constraint. The classifier f_{fair}^* trained on the observed corrupted data is subject to equal TPR constraint for both groups. [1] However, f_{fair}^* has a higher TPR (2/3: 200 correct predictions out of 300 true +1 labels) on B than on A (1/2: 100 correct predictions out of 200 true +1 labels) when evaluated on the clean data.

In this paper, we look at the problem of fair classification from data whose labels are corrupted, such that the error rates of corruption are group-dependent. Several recent works deal with fair classification with noisy labels [7, 23, 26]. In particular, it has been shown that fairness constraints on the noisy training labels can be beneficial when the label noise is homogeneous across the different groups that are to be protected [7]. More recently, [18] shows that how the true fairness rates, such as TPR, are related to observed quantities with respect to noise parameters. Our work complements these results: we show that enforcing fairness constraints when training on data with noisy labels produces a classifier that violates the fairness constraints as measured with respect to the clean data. We then provide a fair empirical risk minimization (ERM) framework that handles heterogenous label noise. Our framework uses an estimation procedure that infers the knowledge of group-dependent noise in the training data and applies this knowledge using bias removal techniques, thus eliminating the effects of noisy labels in both the objective function and the fairness constraints (in expectation).

Our main contributions are as follows: (1) We show that imposing fairness constraints on the training process without accounting for bias in the noisy labels can result in classifiers being less accurate *and* less fair (Theorems 1 and 3 of Section 3). (2) We experimentally demonstrate that these harms can indeed occur in practice with real data sets, and show that obliviously enforcing equality of opportunity without awareness of the noise leads to classifiers with no discriminatory power. (3) We design two noise-resistant fair ERM approaches that address these problems (Section 4). The main idea is to construct unbiased estimators of the loss functions and of the fairness constraints. (4) We provide empirical evidence showing that these fair ERM solutions improve both accuracy and fairness guarantees when facing group-dependent label noise (Section 5). (5) Our codes for solving the noise-resistant fairness constrained ERM can be found at https://github.com/Faldict/fair-classification-with-noisy-labels.

1.1 Related Works

A great deal of research has been devoted to fair classification in general, including fair classification under statistical constraints [1, 17, 22, 45], decoupled training with preference guarantees [10, 16, 28, 41, 44], and preventing gerrymandering [25], among many others [12, 33].

In this work, we specifically focus on fairness in the presence of biased and group-dependent noisy training labels. Our work contributes to the fair classification literature by introducing robust methods for dealing with heterogeneous label noise. We also provide insight into the effects of noise being present in the labels. Our work parallels others' on fair classification with noisy labels [7, 23]. Ours differs primarily in two main respects. First, existing works often assume knowledge of the noise generation process. Second, previous works have only considered noise rates that are homogeneous across different groups. We consider a more realistic setting, where different groups might suffer different levels of bias, and therefore reach very different conclusions. Mitigating bias is substantially more challenging in our setting. Nevertheless, our results could generalized prior work when the noise is assumed constant across groups, or only one group is assumed to have noise.

Both of our fair ERM approaches extend the literature on learning with noisy data [3, 11, 19, 29, 31, 32, 34, 37, 39]. Our first uses surrogate loss functions based on [34] to create unbiased estimators of the fairness constraints. This first approach requires knowledge of the noise parameters. Our second approach relaxes this assumption by extending the work of [30] to account for both biases in the fairness constraints and for group specific label noise.

Recent work on fair classification with imperfect data shows how to emulate noiseless fair classification by appropriately rescaling the fairness tolerance with the noise but is only restricted

[1] Note that f_{fair}^* on the pooled data output +1 for $(0, 1)$ and -1 for $(1, 0)$ because equal TPR constraint is enforced. In this case, the TPRs for both groups are 50%. If the classifier output -1 for $(0, 1)$ and +1 for $(1, 0)$ instead, the TPR for group A is 100% while the TPR for group B is only 50%, which violates the equal TPR constraint.

to class-conditional random noise without considering group difference [26]. Most of the reported results are for the cases with noisy sensitive attributes but not the labels (despite that the authors provided discussions to how the two problems are related). The surrogate fairness constraints in our paper could be viewed as an extension of their method. Nonetheless, our work is more general, as we consider the more sophisticated settings with group-dependent label noise. [21] explores the use of proxy variables when the sensitive attributes are missing. Lastly, [18] also provides some insights on correcting for observed predictive bias might further increase outcome disparities but is concerned with fairness evaluation rather than learning. In contrast with their work, we simplify the assumption on instance-dependent noise into group-dependent, and further develop two fair ERM approaches in terms of the unbiased estimators.

2 PRELIMINARIES

We start with a dataset with n examples $(\boldsymbol{x}_i, y_i, z_i)_{i=1}^n$, where each example consists of a *feature vector* $\boldsymbol{x}_i = (1, x_{i,1}, \dots, x_{i,d}) \in \mathbb{R}^{d+1}$, a *label* $y_i \in \{+1, -1\}$, and a *group attribute* $z_i \in Z$ (e.g., $z_i = \mathbb{1}[\texttt{female}]$). We assume that there are $m = |Z| \geq 2$ groups. We let n_z denote the number of examples in group z, and we use $I_z = \{i \mid z_i = z\}$ and $I = \bigcup_{z \in Z} I_z$ to denote their indices. We assume that each example is drawn iid from a joint distribution $\mathcal{D}$ of random variables (X, Y, Z).

We use the data set to train a classifier $f \in \mathcal{H} : \mathbb{R}^{d+1} \to \{+1, -1\}$, where $\mathcal{H}$ denotes our concept class. To this end, we consider solving a standard risk minimization problem with fairness constraints.

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\mathbb{1}(f(X) \neq Y)\right] \tag{1}$$

$$\text{s.t. } |F_z(f) - F_{z'}(f)| \leq \delta \qquad \forall z, z' \in Z. \tag{2}$$

Here, $F_z(f)$ is some fairness statistic of f for group z given the true labels y, such as *true positive rate* :

$$(\text{TPR}) : F_z(f) = \mathbb{P}(f(X) = +1 | Y = +1, Z = z).$$

Constraint (2) restricts the disparity between z, z' to at most $\delta \geq 0$. A standard approach for performing above constrained minimization is via empirical risk minimization (ERM):

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \mathbb{1}(f(\boldsymbol{x}_i) \neq y_i) \tag{3}$$

$$\text{s.t. } |\widehat{F}_z(f) - \widehat{F}_{z'}(f)| \leq \delta \qquad \forall z, z' \in Z. \tag{4}$$

where $\widehat{F}_z(f)$ is our fairness metric defined using training data. For instance, when using the TPR as a fairness measure:

$$\widehat{F}_z(f) := \frac{\#(f(\boldsymbol{x}_i) = +1, y_i = +1, z_i = z)}{\#(y_i = +1, z_i = z)},$$

where $\#(\cdot)$ is simply a counting function that counts the number of samples that satisfy the specified conditions.

For computational purposes, ERM is performed in practice by minimizing over a classification-calibrated loss function [5] $\ell : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}_+$. This fits:

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i) \tag{5}$$

$$\text{s.t. } |\widehat{F}_z(f) - \widehat{F}_{z'}(f)| \leq \delta \qquad \forall z, z' \in Z. \tag{6}$$

Typical $\ell(\cdot)$s include square loss, logistic loss, cross-entropy loss and more.

We aim to train a classifier using a dataset where the ground truth labels y_i are replaced by *noisy* (or *corrupted*) labels $\tilde{y}_i \sim \tilde{Y}$. A noisy label $\tilde{y}$ corresponds to a true label y that may have been flipped based on noise rate $0 \leq \epsilon_z^+ + \epsilon_z^- < 1$ (as a function of true label y). More precisely, we assume that the noise rates vary based on the true label y as well as the group attribute z:

$$\epsilon_z^+ = \mathbb{P}(\tilde{Y} = -1 \mid Y = +1, Z = z),$$
$$\epsilon_z^- = \mathbb{P}(\tilde{Y} = +1 \mid Y = -1, Z = z)$$

i.e., the training labels are generated as:

$$\tilde{y}_i = \begin{cases} y_i & \text{w.p. } 1 - \epsilon_{z_i}^{\text{sign}(y_i)} \\ -y_i & \text{w.p. } \epsilon_{z_i}^{\text{sign}(y_i)}. \end{cases}$$

This reflects a setting where noise rates are independent of the $\boldsymbol{x}_i$ at fixed values y_i and z_i (e.g., a medical problem where y_i is the presence of a disease, and the disease is diagnosed less reliably for females $z_i = 1$).

In this paper, we mainly focus on two specific fairness constraints: Equal Opportunity and Equal Odds [22]. Equal opportunity requires that each group achieves equal *true positive rate* (TPR) or *false positive rate* (FPR), while equal odds requires both equal TPR and equal FPR. We use the following shorthand to denote different measures of performance, including TPR and FPR, computed for each group using the true labels y and the noisy labels $\tilde{y}$, where $y, \tilde{y} \in \{+1, -1\}$:

$$\text{TPR}_z := \mathbb{P}(f(X) = +1 \mid Y = +1, Z = z)$$
$$\text{FPR}_z := \mathbb{P}(f(X) = +1 \mid Y = -1, Z = z)$$
$$\widetilde{\text{TPR}}_z := \mathbb{P}(f(X) = +1 \mid \tilde{Y} = +1, Z = z)$$
$$\widetilde{\text{FPR}}_z := \mathbb{P}(f(X) = +1 \mid \tilde{Y} = -1, Z = z)$$

$\widetilde{\text{TPR}}_z$ and $\widetilde{\text{FPR}}_z$ are taken with respect to the noisy labels.

3 ENFORCING FAIRNESS CONSTRAINTS ON NOISY LABELS CAN BE HARMFUL

Recent results have established that enforcing fairness constraints improves classifier accuracy when the labels suffer from label noise that is *uniform* across different groups [7]. However, as we shall see, adding fairness constraints can lead to harm when *group-dependent* noise is present in the labels.

3.1 Parity Constraints on Noisy Labels Harms Groups with Clean Labels

The first message that we wish to deliver is that *naively enforcing parity constraints on the noisy labels may harm the accuracy of the classifier for the groups that are not affected by label noise.* Without loss of generality, we present our results in settings where we wish to train a classifier with equal TPR across groups. Similar

Table 3: Label noise harms accuracy: Adult dataset. High FPR implies weak discrimination power. We highlight any high harm the classifier suffers when enforcing equal TPR.

Metrics	Groups	f		f_{fair}
TPR	*female*	97.12%	⇒	96.44%
	male	92.40%	⇒	98.26%
FPR	*female*	53.35%	⇒	78.11%
	male	46.81%	⇒	84.32%
Accuracy	*female*	91.62%	⇒	88.32%
	male	80.39%	⇒	72.97%

derivations hold for other related constraints (e.g., the ones as linear combinations of the entries in the confusion matrix), such as equal FPR, and equal balance error (BER) [33].

Consider a classification problem with two identical groups z and z' where samples from group z have uncorrupted labels while samples from group z' have noisy labels. On the clean data, the parity constraints naturally hold since the data for both groups is drawn from an identical distribution. We next show that the label noise presented in group z' can harm the clean group z when enforcing parity constraints. Formally:

Theorem 1. *Consider a setting with two identical groups $(X, Y, Z = z)$ and $(X, Y, Z = z')$. Group z has clean labels, i.e., $\epsilon_z^+ = \epsilon_z^- = 0$. Group z' suffers from symmetric noise $\epsilon_{z'}^+ = \epsilon_{z'}^- = e > 0$. In this setting, a classifier trained subject to the equal TPR constraint ($\text{TPR}_z = \widetilde{\text{TPR}}_{z'}$) leads to an uninformative classifier that $\text{TPR}_z = \text{FPR}_z$.*

We defer the proof to Section Ommited Proofs. Thus, even if group z is represented with completely uncorrupted labels in the training data, the imposition of equal TPR in the presence of noise for z' will diminish the classifier's predictive accuracy on members of group z.

Case study. We empirically examine the above observation on the Adult dataset from UCI Machine Learning repository [15]. There are two sensitive groups, $Z = \{male, female\}$, in this data set. We inject symmetric noise $\epsilon^+ = \epsilon^- = 0.3$ into labels for members of the *female* group. Then, we train two classifiers: f, which is trained without any fairness constraints, and f_{fair}, which is trained with the imposition of equal TPR using the reduction method [1]. As is shown in Table 3, the empirical results mirror Theorem 1. When the difference between f_{fair}'s TPR for the two groups becomes small (less than 2%), f_{fair}'s TPR and FPR become close together, and the accuracy decreases significantly. The above trends hold even when we try to equalize TPR and FPR together across groups. We notice that the two groups are not strictly identical in the Adult dataset, but our example implies that there exists dangerous cases where enforcing fairness constraints can harm classifier accuracy for the group with uncorrupted labels.

3.2 Violation of Fairness under Perceived Fairness

Our second message is that *training fair classifiers using noisy labels may lead to a false impression of fairness.* This arises when the fairness constraints are satisfied over the noisy labels while being violated over the clean labels. Before proceeding, we require extending Proposition 16 of [32] into the situation with group-dependent label noise. A similar result appears in [40].

Lemma 1. *For each group z we have that*

$$\text{TPR}_z = (1 - \epsilon_z^+) \cdot \widetilde{\text{TPR}}_z + \epsilon_z^+ \cdot \widetilde{\text{FPR}}_z \tag{7}$$

$$\text{FPR}_z = \epsilon_z^- \cdot \widetilde{\text{TPR}}_z + (1 - \epsilon_z^-) \cdot \widetilde{\text{FPR}}_z \tag{8}$$

Proof. Expanding $\mathbb{P}(f(X) = +1 \mid Y = +1, Z = z)$ using law of total probability we have

$$\begin{aligned}
\text{TPR}_z &= \mathbb{P}(f(X) = +1 \mid Y = +1, Z = z)\\
&= \mathbb{P}(f(X) = +1, \tilde{Y} = +1 \mid Y = +1, Z = z)\\
&\quad + \mathbb{P}(f(X) = +1, \tilde{Y} = -1 \mid Y = +1, Z = z)\\
&= \mathbb{P}(\tilde{Y} = +1 | Y = +1, Z = z) \cdot \mathbb{P}(f(X) = +1 \mid \tilde{Y} = +1, Y = +1, Z = z)\\
&\quad + \mathbb{P}(\tilde{Y} = -1 | Y = +1, Z = z) \cdot \mathbb{P}(f(X) = +1 \mid \tilde{Y} = -1, Y = +1, Z = z)\\
&= \mathbb{P}(\tilde{Y} = +1 | Y = +1, Z = z) \cdot \mathbb{P}(f(X) = +1 \mid \tilde{Y} = +1, Z = z)\\
&\quad + \mathbb{P}(\tilde{Y} = -1 | Y = +1, Z = z) \cdot \mathbb{P}(f(X) = +1 \mid \tilde{Y} = -1, Z = z)\\
&= (1 - \epsilon_z^+) \cdot \widetilde{\text{TPR}}_z + \epsilon_z^+ \cdot \widetilde{\text{FPR}}_z
\end{aligned} \tag{9}$$

Note in the above we drop the dependence on Y when conditioning on $\tilde{Y}$. This is because f is trained purely on the noisy labels, and $\tilde{Y}$ encodes all the information f has about Y.

A similar derivation holds for FPR_z. □

We also note that, in the special case where all groups suffer from an identical rate of label corruption, the learner *can* be oblivious to the specific error rates:

Theorem 2. *Consider a classification problem with noisy labels where the noise rates are independent of group membership, so that $\epsilon_z^+ = \epsilon_{z'}^+$ and $\epsilon_z^- = \epsilon_{z'}^-$ $\forall z, z' \in Z$. Then it follows that $\text{TPR}_z = \text{TPR}_{z'}$ $\forall z, z' \in Z$, if equal odds (equalizing both TPR and FPR) on the noisy labels is imposed.*

The proof follows by applying the assumption of equal error rates and equal odds on the noisy labels with Lemma 1. However, things break down in the general case. If we impose equal odds across groups on a learner that is unaware of the labels' noisiness (i.e. whenever $\widetilde{\text{TPR}}_z = \widetilde{\text{TPR}}_{z'}$), then:

Theorem 3. *Assume that a classifier is subject to equal odds in the presence of group-dependent label noise. Then for any two groups $z, z' \in Z$, we have*

$$| \text{TPR}_z - \text{TPR}_{z'} | = |\widetilde{\text{TPR}}_z - \widetilde{\text{FPR}}_z| \cdot |\epsilon_z^+ - \epsilon_{z'}^+|,$$

$$| \text{FPR}_z - \text{FPR}_{z'} | = |\widetilde{\text{TPR}}_z - \widetilde{\text{FPR}}_z| \cdot |\epsilon_z^- - \epsilon_{z'}^-|.$$

Unless the classifier is random on the noisy training data, i.e., $\widetilde{\text{TPR}}_z = \widetilde{\text{FPR}}_z$, it is impossible to satisfy equal odds over the clean data whenever $\epsilon_z^+ \neq \epsilon_{z'}^+$ and $\epsilon_z^- \neq \epsilon_{z'}^-$.

Proof. Noticing that $\widetilde{\text{TPR}}_z = \widetilde{\text{TPR}}_{z'}$ and $\widetilde{\text{FPR}}_z = \widetilde{\text{FPR}}_{z'}$ (equalizing fairness metrics on the noisy data) and applying Lemma 1, we obtain

$$\begin{aligned}|\operatorname{TPR}_z - \operatorname{TPR}_{z'}| &= |((1-\epsilon_z^+)\cdot\widetilde{\operatorname{TPR}}_z + \epsilon_z^+\cdot\widetilde{\operatorname{FPR}}_z) \\ &\quad - ((1-\epsilon_{z'}^+)\cdot\widetilde{\operatorname{TPR}}_{z'} + \epsilon_{z'}^+\cdot\widetilde{\operatorname{FPR}}_{z'})| \\ &= |\epsilon_z^+\cdot(\widetilde{\operatorname{FPR}}_z - \widetilde{\operatorname{TPR}}_z) - \epsilon_{z'}^+\cdot(\widetilde{\operatorname{FPR}}_z - \widetilde{\operatorname{TPR}}_z)| \\ &= |(\epsilon_z^+ - \epsilon_{z'}^+)\cdot(\widetilde{\operatorname{FPR}}_z - \widetilde{\operatorname{TPR}}_z)| \\ &= |\widetilde{\operatorname{TPR}}_z - \widetilde{\operatorname{FPR}}_z|\cdot|\epsilon_z^+ - \epsilon_{z'}^+|\end{aligned}$$

The argument for FPR is symmetrical:

$$\begin{aligned}|\operatorname{FPR}_z - \operatorname{FPR}_{z'}| &= |(\epsilon_z^-\cdot\widetilde{\operatorname{TPR}}_z + (1-\epsilon_z^-)\cdot\widetilde{\operatorname{FPR}}_z) \\ &\quad - (\epsilon_{z'}^-\cdot\widetilde{\operatorname{TPR}}_{z'} + (1-\epsilon_{z'}^+)\cdot\widetilde{\operatorname{FPR}}_{z'})| \\ &= |\epsilon_z^-\cdot(\widetilde{\operatorname{TPR}}_z - \widetilde{\operatorname{FPR}}_z) - \epsilon_{z'}^-\cdot(\widetilde{\operatorname{TPR}}_z - \widetilde{\operatorname{FPR}}_z)| \\ &= |(\epsilon_z^- - \epsilon_{z'}^-)\cdot(\widetilde{\operatorname{TPR}}_z - \widetilde{\operatorname{FPR}}_z)| \\ &= |\widetilde{\operatorname{TPR}}_z - \widetilde{\operatorname{FPR}}_z|\cdot|\epsilon_z^- - \epsilon_{z'}^-|\end{aligned}$$

Therefore

$$|\operatorname{TPR}_z - \operatorname{TPR}_{z'}| > 0, |\operatorname{FPR}_z - \operatorname{FPR}_{z'}| > 0,$$

when $\widetilde{\operatorname{TPR}}_z \neq \widetilde{\operatorname{FPR}}_z, \epsilon_z^+ \neq \epsilon_{z'}^+, \epsilon_z^- \neq \epsilon_{z'}^-$. □

The proof follows by a direct application of Lemma 1. Theorem 3 implies that the true fairness violation is proportional to the difference in error rates across the different sub-groups. We offer two remarks. First, if the error rates are systematically biased towards a particular group, then a perceived fair classifier will lead to unequal odds. Second, the above bias will be reinforced when the trained model is more accurate on noisy data; a more accurate model will lead to a larger difference in $|\widetilde{\operatorname{TPR}}_z - \widetilde{\operatorname{FPR}}_z|$.

4 FAIR ERM WITH NOISY LABELS

In this section, we describe two noise-tolerant and fair ERM solutions that address the combined challenges of heterogeneous and group-dependent label noise. Both the surrogate loss and group-weighted peer loss approaches for handling noisy labels rely on estimations of the label noise. Our procedure for estimating the noise parameters, detailed in Section 4.3, is an adaptation of [35]. Section 4.3 also offers discussion of the impacts of noisy estimates.

4.1 A Surrogate Loss Approach

As we shall see, training an unmodified loss function using the noisy labels $\tilde{y}_i$ corrupts the model in a manner that cannot be addressed via post-hoc correction. Thus, a natural resolution is to modify the loss function itself. This modified loss function is called a *surrogate loss.*

Bias removal surrogate loss functions. Bias removal via a surrogate loss is a popular approach to handling label noise [34]. The original loss function $\ell(\cdot)$ is replaced with a surrogate loss function $\tilde{\ell}(\cdot)$ that "corrects" for noise in the labels in expectation. Formally, the surrogate loss is chosen so that the cost of mis-classifying an element $\boldsymbol{x}_i$ with true label y_i is equivalent to the expected loss value that arises from using noisy label $\tilde{y}_i$. Thus, we want to find a surrogate loss $\tilde{\ell}$ such that:

Table 4: Surrogate constraints for surrogate loss.

Metric	$\widehat{F}_z(f)$
TPR	$(1-\epsilon_z^+)\cdot\widehat{\operatorname{TPR}}_z + \epsilon_z^+\cdot\widehat{\operatorname{FPR}}_z$
FPR	$\epsilon_z^-\cdot\widehat{\operatorname{TPR}}_z + (1-\epsilon_z^-)\cdot\widehat{\operatorname{FPR}}_z$
Equal Odds	both TPR and FPR

Table 5: Surrogate constraints for group weighted peer loss

Metric	$\widehat{F}_z(f)$
TPR	$\mathbb{P}(f(X)=+1\|Z=z) + \frac{\Delta_z}{2}(\widehat{\operatorname{TPR}}_z - \widehat{\operatorname{FPR}}_z)$
FPR	$\mathbb{P}(f(X)=+1\|Z=z) - \frac{\Delta_z}{2}(\widehat{\operatorname{TPR}}_z - \widehat{\operatorname{FPR}}_z)$
Equal Odds	both TPR and FPR

$$\ell(f(\boldsymbol{x}), y) = \mathbb{E}_{\tilde{Y}}[\tilde{\ell}(f(\boldsymbol{x}), \tilde{Y}) \mid Y = y] \tag{10}$$

for all $\boldsymbol{x}$ and y. When the noise depends on the label value, the function given by

$$\tilde{\ell}(f(\boldsymbol{x}_i), \tilde{y}_i = +1) := \frac{(1-\epsilon_{z_i}^-)\ell(f(\boldsymbol{x}_i), +1) - \epsilon_{z_i}^+\ell(f(\boldsymbol{x}_i), -1)}{1-\epsilon_{z_i}^+ - \epsilon_{z_i}^-}, \tag{11}$$

$$\tilde{\ell}(f(\boldsymbol{x}_i), \tilde{y}_i = -1) := \frac{(1-\epsilon_{z_i}^+)\ell(f(\boldsymbol{x}_i), -1) - \epsilon_{z_i}^-\ell(f(\boldsymbol{x}_i), +1)}{1-\epsilon_{z_i}^+ - \epsilon_{z_i}^+}. \tag{12}$$

satisfies the above property, as shown by Lemma 1 in [34]. A classifier f minimizing the surrogate loss on noisy data $\tilde{\ell}(X, \tilde{Y})$ will minimize the loss on clean data $\ell(X, Y)$ in expectation. This property allows us to perform model selection on a noisy validation set, and one could choose the model that performs better on the validation set to deploy.

Surrogate fairness constraints. We will also need to modify the fairness constraints to account for the effects of noise. Our method of doing so is inspired by the surrogate loss that we need to work with an unbiased estimate of the fairness constraints. For the case of binary classification, we can express the surrogate measures of group-based fairness constraints using Lemma 1.

We use Equation (11) and Equation (12) to define our surrogate loss functions $\tilde{\ell}_z(f(\boldsymbol{x}_i), \tilde{y}_i = +1)$, and $\tilde{\ell}_z(f(\boldsymbol{x}_i), \tilde{y}_i = -1)$. Furthermore, define the empirical TPR and FPR over the noisy labels as follows:

$$\widehat{\operatorname{TPR}}_z(f) = \frac{\#(f(\boldsymbol{x}_i) = +1, \tilde{y}_i = +1, z_i = z)}{\#(\tilde{y}_i = +1, z_i = z)} \tag{13}$$

$$\widehat{\operatorname{FPR}}_z(f) = \frac{\#(f(\boldsymbol{x}_i) = +1, \tilde{y}_i = -1, z_i = z)}{\#(\tilde{y}_i = -1, z_i = z)} \tag{14}$$

We then define our surrogate fairness measures $\widehat{F}_z(f)$ using only noisy data, as detailed in Table 4. Our noise-resistant fair ERM states as follows:

$$\begin{aligned}&\min_{f\in\mathcal{H}} \sum_{i=1}^{n} \tilde{\ell}(f(\boldsymbol{x}_i), \tilde{y}_i) \\ &\text{s.t. } |\widehat{F}_z(f) - \widehat{F}_{z'}(f)| \leq \delta, \ \forall z, z'.\end{aligned} \tag{15}$$

4.2 Group Weighted Peer Loss Approach

The recently developed *peer loss* function partially circumvents the issue of noise rate estimation [30]. The peer loss requires less prior knowledge of the noise rates for each class. It is defined as:

$$\ell_{peer}(f(\boldsymbol{x}_i), \tilde{y}_i) := \ell\left(f(\boldsymbol{x}_i), \tilde{y}_i\right) - \alpha \cdot \ell\left(f(\boldsymbol{x}_{i_1}), \tilde{y}_{i_2}\right), \tag{16}$$

where

$$\alpha = 1 - (1 - \epsilon^- - \epsilon^+) \cdot \frac{\mathbb{P}(Y = +1) - \mathbb{P}(Y = -1)}{\mathbb{P}(\tilde{Y} = +1) - \mathbb{P}(\tilde{Y} = -1)}$$

is a parameter to balance the instances for each label, and where i_1 and i_2 are uniformly and randomly selected samples from $I_z/\{i\}$ (i.e., “peer" samples which inspired the name peer loss as noted in [30]). Although the noise parameters explicitly appear in the definition of α, only the knowledge of $\Delta := 1 - \epsilon^- - \epsilon^+$ is needed. In practice, we could tune α as a hyper-parameter during training. This loss function has the following important property, proven in Lemma 3 of [30]:

$$\mathbb{E}_{\tilde{\mathcal{D}}_z}[\ell_{peer}(f(X), \tilde{Y})] = \Delta_z \cdot \mathbb{E}_{\mathcal{D}_z}[\ell_{peer}(f(X), Y)], \tag{17}$$

where $\tilde{\mathcal{D}}_z$ denotes the noisy distribution for group z and $\Delta_z = 1 - \epsilon_z^- - \epsilon_z^+$. Adapting the peer loss function to labels with group dependent noise requires accounting for the differing values of Δ_z. We do so by re-weighting Equation (16) to obtain our *group-weighted peer loss* ℓ_{gp}:

$$\ell_{gp}(f(\boldsymbol{x}_i), \tilde{y}_i) := \frac{1}{\Delta_{z_i}}\left(\ell(f(\boldsymbol{x}_i), \tilde{y}_i) - \alpha \cdot \ell\left(f(\boldsymbol{x}_{i_1}), \tilde{y}_{i_2}\right)\right). \tag{18}$$

When class is balanced for every group z, i.e., $\mathbb{P}_{Z=z}(Y = +1) = \mathbb{P}_{Z=z}(Y = -1) = \frac{1}{2}$, the parameter α is exactly 1. In this case, the expected group-weighted peer loss on the noisy distribution $\tilde{\mathcal{D}}$ is the same as the expected uncorrected loss ℓ on the true distribution $\mathcal{D}$. More precisely:

Theorem 4. *For all group dependent noise rates ϵ_z^- and ϵ_z^+ satisfying $\epsilon_z^- + \epsilon_z^+ < 1$, taking $\ell(\cdot)$ as the 0-1 loss $\mathbb{1}(\cdot)$ and when $\mathbb{P}_{Z=z}(Y = +1) = \mathbb{P}_{Z=z}(Y = -1) = \frac{1}{2}$,*

$$\mathbb{E}_{\tilde{\mathcal{D}}}[\ell_{gp}(f(X), \tilde{Y})] = \mathbb{E}_{\mathcal{D}}[\ell(f(X), Y] - \frac{1}{2}. \tag{19}$$

Proof. Observe that

$$\ell_{gp}(f(\boldsymbol{x}_i), \tilde{y}) = \frac{1}{\Delta_{z_i}}\ell_{peer}(f(\boldsymbol{x}_i), \tilde{y})$$

Taking expectations over noisy data, we have

$$\begin{aligned}
&\mathbb{E}_{\tilde{\mathcal{D}}}[\ell_{gp}(f(X), \tilde{Y})] \\
&= \frac{1}{|I|} \cdot \sum_{z \in Z} |I_z| \cdot \mathbb{E}_{\tilde{\mathcal{D}}_z}[\ell_{gp}(f(X_z), \tilde{Y}_z)] \\
&= \frac{1}{|I|} \cdot \sum_{z \in Z} \frac{|I_z|}{\Delta_z} \cdot \mathbb{E}_{\tilde{\mathcal{D}}_z}[\ell_{peer}(f(X_z), \tilde{Y}_z)] \\
&= \frac{1}{|I|} \cdot \sum_{z \in Z} \frac{|I_z|}{\Delta_z} \cdot \Delta_z \mathbb{E}_{\mathcal{D}_z}[\ell_{peer}(f(X_z), Y_z)] \qquad \text{(by Equation 17)} \\
&= \mathbb{E}_{\mathcal{D}}[\ell_{peer}(f(X), Y)]
\end{aligned} \tag{20}$$

Notice that $\alpha = 1$ when $\mathbb{P}(Y = +1) = \mathbb{P}(Y = -1) = \frac{1}{2}$, the definition of peer loss function gives

$$\mathbb{E}_{X,Y}[\ell_{peer}(f(X), Y)] = \mathbb{E}_{X,Y}[\ell(f(X), Y)] - \mathbb{E}_X\mathbb{E}_Y[\ell(f(X), Y)] \tag{21}$$

Using the assumption that $\mathbb{P}(Y = +1) = \mathbb{P}(Y = -1) = \frac{1}{2}$ and the fact that ℓ is 0-1 loss function,

$$\begin{aligned}
\mathbb{E}_X\mathbb{E}_Y[\ell(f(X), Y)] &= \mathbb{P}(Y = +1) \cdot \mathbb{E}_X[\ell(f(X), +1)] + \\
&\quad \mathbb{P}(Y = -1) \cdot \mathbb{E}_X[\ell(f(X), -1)] \\
&= \frac{1}{2} \cdot \ell(f(X), +1) + \frac{1}{2} \cdot \ell(f(X), +1) \\
&= \frac{1}{2} \cdot \mathbb{1}(f(X) \neq +1) + \frac{1}{2} \cdot \mathbb{1}(f(X) \neq -1) \\
&= \frac{1}{2}\mathbb{P}(f(X) = -1) + \frac{1}{2} \cdot \mathbb{P}(f(X) = +1) \\
&= \frac{1}{2}
\end{aligned} \tag{22}$$

Combining Eq. (20), Eq. (21) and Eq. (22), we complete the proof

$$\mathbb{E}_{\tilde{\mathcal{D}}}[\ell_{gp}(f(X), \tilde{Y})] = \mathbb{E}_{\mathcal{D}}[\ell(f(X), Y] - \frac{1}{2}$$

□

Peer-based surrogate fairness constraints. We acquire the following result in order to create group-aware surrogate constraints:

Lemma 2. *True* TPR *and* FPR *relate to* $\widetilde{\mathrm{TPR}}_z, \widetilde{\mathrm{FPR}}_z$ *defined on the noisy labels as follows:*

$$\mathrm{TPR}_z = \mathbb{P}(f(X) = +1 | Z = z) + \Delta_z \cdot (\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z) \cdot \mathbb{P}(Y = -1 | Z = z) \tag{23}$$

$$\mathrm{FPR}_z = \mathbb{P}(f(X) = +1 | Z = z) - \Delta_z \cdot (\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z) \cdot \mathbb{P}(Y = +1 | Z = z) \tag{24}$$

Proof. Following Lemma 1 we have,

$$\mathrm{TPR}_z - \mathrm{FPR}_z = (1 - \epsilon_z^+ - \epsilon_z^-)(\widetilde{\mathrm{TPR}} - \widetilde{\mathrm{FPR}}) = \Delta_z \cdot (\widetilde{\mathrm{TPR}} - \widetilde{\mathrm{FPR}})$$

Notice that

$$\begin{aligned}
\mathbb{P}(f(X) = +1 \mid Z = z) &= \mathbb{P}(Y = +1 \mid Z = z) \cdot \mathbb{P}(f(X) = +1 \mid Y = +1, Z = z) \\
&\quad + \mathbb{P}(Y = -1 | Z = z) \cdot \mathbb{P}(f(X) = +1 \mid Y = -1, Z = z) \\
&= \mathbb{P}(Y = +1 | Z = z) \cdot \mathrm{TPR}_z + \mathbb{P}(Y = -1 | Z = z) \cdot \mathrm{FPR}_z
\end{aligned}$$

Solving the two equations above we complete the proof. □

Lemma 2 allows us to derive the appropriate surrogate fairness constraints for the peer loss, displayed in Table 5. Note that we have assumed that the datasest is balanced for each group; i.e., $\forall z \in Z \quad \mathbb{P}(Y = +1 | Z = z) = \frac{1}{2}$. If the data is imbalanced, we will require knowing the marginal prior $\mathbb{P}(Y = +1 | Z = z)$. We note that it is straightforward to get the estimated marginal priors as given by Equation (27) in Section 4.3.

We merely require knowledge of Δ_z for each z in order to define ℓ_{gp} and $\widehat{F}_z(f)$. This is a weaker requirement compared to knowing the error rates (which will carry estimation of two parameters for each group). We indeed see our group peer loss approach performs more stably as compared to the surrogate loss approach introduced in last subsection when using noisy estimates of the noise rates. With group-weighted peer loss function and surrogate fairness constraints, we are able to perform a fair ERM as detailed in Equaltion (15) by replacing $\tilde{\ell}$ with ℓ_{gp} and the corresponding $\widehat{F}_z(f)$ term.

4.3 Error Rates Estimation and its Impact

We employ "confident learning" to perform noise rate estimation in our experiments [35]. The first step is to pre-train a classifier f_{pre} over the noisy labels directly and learn a noisy predicted probability

$$\hat{p}(y; \boldsymbol{x}, z) = \mathbb{P}(f_{pre}(\boldsymbol{x}) = y | Z = z).$$

Then, for each pair of classes $k, l \in \{+1, -1\}$, we define the subset of samples:

$$\widehat{X}_{\tilde{y}=k,z} := \{\boldsymbol{x}_i | \tilde{y}_i = k, i \in I_z\},$$
$$\widehat{X}_{\tilde{y}=k,y=l,z} := \{\boldsymbol{x}_i | \tilde{y}_i = k, \hat{p}(y = l; \boldsymbol{x}_i, z) \geq t_{l,z}, i \in I_z\},$$

where

$$t_{l,z} = \frac{1}{|\widehat{X}_{\tilde{y}=l,z}|} \sum_{\boldsymbol{x} \in \widehat{X}_{\tilde{y}=l,z}} \hat{p}(\hat{y} = l; \boldsymbol{x}, z)$$

is the *expected self-confidence probability* for class l and group z. Using the above quantities, we estimate the group-aware joint probability $\widehat{Q}_{\tilde{y}=k,y=l,z} = \mathbb{P}(\widetilde{Y} = k, Y = l, Z = z)$ over the noisy labels $\tilde{y}$ and clean labels y with:

$$\widehat{Q}_{\tilde{y}=k,y=l,z} = \frac{\frac{|\widehat{X}_{\tilde{y}=k,y=l,z}|}{\sum_l |\widehat{X}_{\tilde{y}=k,y=l,z}|} \cdot |\widehat{X}_{\tilde{y}=k,z}|}{\sum_{k,l} \left(\frac{|\widehat{X}_{\tilde{y}=k,y=l,z}|}{\sum_l |\widehat{X}_{\tilde{y}=k,y=l,z}|} \cdot |\widehat{X}_{\tilde{y}=k,z}| \right)} \tag{25}$$

We use the marginals of estimated joint to compute the noise parameter estimates for each group z:

$$\hat{\epsilon}_z^+ = \frac{\widehat{Q}_{\tilde{y}=-1,y=+1,z}}{\widehat{Q}_{\tilde{y}=-1,y=+1,z} + \widehat{Q}_{\tilde{y}=+1,y=+1,z}},$$
$$\hat{\epsilon}_z^- = \frac{\widehat{Q}_{\tilde{y}=+1,y=-1,z}}{\widehat{Q}_{\tilde{y}=+1,y=-1,z} + \widehat{Q}_{\tilde{y}=-1,y=-1,z}} \tag{26}$$

To estimate Δ_z, we simply substitute $\hat{\epsilon}_z^-$ and $\hat{\epsilon}_z^+$ for ϵ_z^- and ϵ_z^+ in the equation for Δ_z. As a byproduct, we could estimate the marginal priors $\mathbb{P}(Y = +1 | Z = z)$ by

$$\frac{\widehat{Q}_{\tilde{y}=+1,y=+1,z} + \widehat{Q}_{\tilde{y}=-1,y=+1,z}}{\widehat{Q}_{\tilde{y}=+1,y=+1,z} + \widehat{Q}_{\tilde{y}=-1,y=+1,z} + \widehat{Q}_{\tilde{y}=-1,y=+1,z} + \widehat{Q}_{\tilde{y}=-1,y=-1,z}} \tag{27}$$

Effects of noisy estimates. It is important to quantify the impact of the noise rate estimation error on the accuracy and fairness of the resulting classifier. We first note that, for any $\eta, \tau > 0$, the law of large numbers implies that taking sufficiently many samples from $\mathcal{D}$ will ensure that the following holds for all z with probability at least $1 - \eta$:

$$\max \Bigg\{ \left|\hat{\epsilon}_z^+ - \epsilon_z^+\right|, \left|\frac{\hat{\epsilon}_z^+}{1 - \hat{\epsilon}_z^+ - \hat{\epsilon}_z^-} - \frac{\epsilon_z^+}{1 - \epsilon_z^+ - \epsilon_z^-}\right|,$$
$$\left|\hat{\epsilon}_z^- - \epsilon_z^-\right|, \left|\frac{1 - \hat{\epsilon}_z^-}{1 - \hat{\epsilon}_z^+ - \hat{\epsilon}_z^-} - \frac{1 - \epsilon_z^-}{1 - \epsilon_z^+ - \epsilon_z^-}\right| \Bigg\} \leq \tau. \tag{28}$$

Denote by $\hat{\ell}(\cdot)$ the surrogate loss function defined using the estimated noises $\{\hat{\epsilon}_z^+, \hat{\epsilon}_z^-\}$, and let

$$\hat{f}^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^N \hat{\ell}(f(\boldsymbol{x}_i), \tilde{y}_i), \qquad \tilde{f}^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^N \tilde{\ell}(f(\boldsymbol{x}_i), \tilde{y}_i)$$

We have the following result and defer the proof to Section **??**:

Theorem 5. *For every $\eta, \tau > 0$ there exists N such that*

$$\frac{1}{N} \cdot \sum_{i=1}^N \tilde{\ell}(\hat{f}^*(\boldsymbol{x}_i), \tilde{y}_i) - \frac{1}{N} \cdot \sum_{i=1}^N \tilde{\ell}(\tilde{f}^*(\boldsymbol{x}_i), \tilde{y}_i) \leq 4\tau \cdot \bar{\ell} \tag{29}$$

with probability at least $1 - \eta$, where $\bar{\ell} := \max \ell$.

Because the fairness constraints $\widehat{F}_z(f)$ are linear in $\epsilon_z^+, \epsilon_z^-$s, the additional fairness violations incurred due to the noisy estimates of the error rates will also be linear in τ too. Similar observations hold when using the estimated $\tilde{\Delta}_z$ in the peer loss.

5 EXPERIMENTS

Due to the difficulty of acquiring real world datasets with known label corruption characteristics, we artificially synthesize the datasets with a noise generation step. These controlled experiments help us understand the robustness of our approaches under different noise scenarios.

5.1 Experimental Setup

Dataset. We evaluate our methods as well as other baseline methods on five datasets:

- `Adult`, the Adult dataset from the UCI ML Repository with males and females as the protected groups [15].
- `Arrest` and `Violent`, the COMPAS recidivism dataset for arrest and violent crime statistics, with race (restricted to white and black) and gender as the sensitive attributes [4].
- `German`, the German credit dataset from UCI ML Repository with gender as the sensitive attribute [15].
- `Law`, a subset of the original data set from LSAC with race (restricted to black and white) as the sensitive attribute [43].

Table 6 describes the dataset statistics and parameters used in the experiments. We chose to apply a diverse set of noise parameters to the different subgroups. The fairness tolerance δ and noise parameters ϵ for `Adult`, `German` and `Law` data sets are identical, but they are different from `Arrest` and `Violent` data sets because `Arrest` and `Violent` data sets contain more protected groups. We make this choice mainly for the baseline models to obtain meaningful results to compare with.

Noise generation. We randomly split the clean dataset $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^n$ into a training set and a test set in a ratio of 80 to 20. We add asymmetric label noises to the training dataset, and leave the test data untouched for verification purposes. For each sensitive group $z \in Z$, we randomly flip the clean label y with probability ϵ_z^- if its value is -1, and we flip the clean label with probability ϵ_z^+ if it's +1. After injecting this noise, we use the same training set and test set to benchmark all the methods.

Methods. For all of the methods above, we use logistic regression to perform classification and leverage the reduction approach as proposed in [1] for solving our constrained optimization problem. We evaluate the performance of several methods:

- `Clean`, in which the classifier is trained on the clean data subject to the equal odds constraint
- `Corrupt`, which directly trains the classifier on the corrupted data subject to the equal odds fairness constraint

Table 6: Dataset statistic and parameters.

Dataset	Source	Number of data examples n	Fairness Tolerance δ	Sensitive Groups	Noise Rates	
					ϵ^-	ϵ^+
adult	UCI [15]	32561	2%	*female* *male*	0.45 0.35	0.15 0.55
arrest	COMPAS [4]	6644	5%	*white* *black*	0.40 0.15	0.30 0.25
arrest	COMPAS [4]	6644	5%	*white* male *black male* *white female* *black female*	0.45 0.10 0.35 0.55	0.10 0.35 0.45 0.25
violent	COMPAS [4]	5278	5%	*white male* *black male* *white female* *black female*	0.45 0.10 0.35 0.55	0.10 0.35 0.45 0.25
German	UCI [15]	1000	2%	*female* *male*	0.45 0.35	0.15 0.55
law	LSAC [43]	18692	2%	*white* *black*	0.45 0.35	0.15 0.55

- Surrogate Loss, which uses the surrogate loss approach described in Section 4.1
- Group Peer Loss, which uses the group weighted peer loss approach described in Section 4.2 to train a fair classifier on the corrupted training set.

The Corrupt baseline gives us a sense about the harm caused by the unawareness of the labels' noise, and the clean baseline shows the biases contained in the datasets.

We set the same maximum fairness violation δ for all the methods on the same dataset during training. As there are more sensitive groups on arrest and violent datasets, we set $\delta = 5\%$ on these datasets and $\delta = 2\%$ on the other datasets. We report metrics for each of the above methods averaged over five runs.

Computing Infrastructure. We conducted all the experiments on a 3 GHz 6-Core Intel Core i5 CPU. The running time for Surrogate Loss is about 10 minutes, while the running time for Group Peer Loss could be over 30 minutes.

Tuning α in Peer Loss. The performance of our group weighted peer loss is highly influenced by the hyperparameter α. Recall that

$$\mathbb{E}_{\tilde{\mathcal{D}}_z}[\ell_{gp}(f(X),\tilde{Y})] = \mathbb{E}_{\mathcal{D}_z}[\ell_{gp}(f(X),Y)]$$

We split 10% of data examples in the train set for validation and found the optimal α using grid search. The range of α we searched varied between 0.0 to 2.0. We observed that both the accuracy and fairness violation on the validation set exhibit the same trends on the test set. In practice, the group weighted peer loss with $\alpha = 0.3$ achieves the best performance on the Adult dataset.

5.2 Results

We present an overview of the performance for each method on the test set in Table 7. We compare the two fair ERM approaches using both the true and estimated noise rates. The metrics we report include *violation*, the maximum difference in TPR and FPR between groups $z, z' \in Z$, and *accuracy*, the accuracy achieved on test set.

We make the following observations about our results. First, both of the two fair ERM approaches in Section 4 produce classifiers that are more effective at mitigating unfairness than a classifier that is naively trained on the corrupted data.

In particular, the group weighted peer loss approach achieves almost 0% violation on the German and law data sets, when given the true noise parameters. The only noticeable worse case arises when applying the surrogate loss approach to the German dataset. This may be due to the high variance of the German dataset, which has fewer than 1000 samples.

Second, as expected, models trained using our proposed fair ERM methods do not achieve the same level of accuracy as a model that is fit using clean labels. However, our models are typically more accurate than the model fit directly to the corrupted data. For example, on the arrest data set with four protected groups, the surrogate loss approach achieves a similar accuracy to the classifier trained on clean data while incurring an even smaller fairness violation. Third, Our methods perform similarly well when trained using both the true and with the estimated noise parameters, indicating that the noise estimation procedures are effective. On arrest and violent datasets, our methods with estimated noise parameters even perform better than those with true parameters. This is probably due to the biases and noise in these datasets. Finally, our fair ERM frameworks adapt well to multiple sensitive groups, as demonstrated by the good performance on the Arrest and Violent data sets.

5.3 Impact of noise levels on classifier performance.

We present the results of varying noise rate on the adult data set (with two groups) in Table 8. We only add symmetric noise to $female$ group and keep the *male* group clean. ERM is generally robust to symmetric noises when a significant subset of the data is clean (one group in our example), so we do not expect significant accuracy improvement from our methods. We focus on how fairness violation reduces. Observe that, comparing to training with clean

Table 7: Overview of group-based performance metrics for all methods on 5 data sets. We highlight the best values achieved for fairness violation and accuracy in green and the worst in red. m is the number of sensitive groups, $\bar{\epsilon}$ is the average of error rates over all the groups and all label classes $\epsilon_z^+, \epsilon_z^-$s. *true* indicates training with true noise parameters and *estimated* indicates training with estimated noise parameters. The values after ± are the standard deviation.

					Surrogate Loss		Group Peer Loss	
Dataset	**Metrics**	**Avg. $\bar{\epsilon}$**	Clean	Corrupt	*true*	*estimated*	*true*	*estimated*
Adult	*violation*	0.38	0.47%	8.36 ± 1.36%	1.46 ± 0.50%	1.39 ± 0.80%	1.18 ± 0.63%	1.69 ± 0.86%
$m = 2$	*accuracy*		83.76%	76.08 ± 2.49%	81.16 ± 3.41%	75.99 ± 7.45%	77.00 ± 2.52%	75.13 ± 5.15%
Arrest	*violation*	0.28	2.27%	2.98 ± 0.74%	0.54 ± 0.27%	0.36 ± 0.24%	1.78 ± 0.89%	1.05 ± 0.55%
$m = 2$	*accuracy*		65.16%	60.72 ± 0.66%	61.7 ± 3.23%	62.3 ± 5.30%	63.81 ± 3.35%	65.31 ± 3.41%
Arrest	*violation*	0.34	5.89%	12.93 ± 0.95%	0.88 ± 0.27%	2.48 ± 1.42%	1.36 ± 0.69%	1.40 ± 0.36%
$m = 4$	*accuracy*		66.0%	53.7 ± 1.82%	65.7 ± 2.92%	58.8 ± 4.96%	60.27 ± 2.90%	57.56 ± 2.96%
Violent	*violation*	0.34	0.37%	7.16 ± 0.80%	4.81 ± 0.70%	7.76 ± 1.02%	2.06 ± 0.81%	0.68 ± 0.28%
$m = 4$	*accuracy*		60.18%	52.2 ± 0.23%	53.14 ± 4.91%	55.4 ± 0.71%	55.64 ± 4.88%	52.7 ± 0.57%
German	*violation*	0.38	0.68%	2.68 ± 0.32%	11.79 ± 3.87%	11.08 ± 2.16%	0.00 ± 0.00%	1.64 ± 0.32%
$m = 2$	*accuracy*		74.5%	70.5 ± 0.00%	68.5 ± 4.27%	71.5 ± 2.53%	70.0 ± 0.71%	70.5 ± 2.53%
Law	*violation*	0.38	0.6%	2.74 ± 0.12%	0.36 ± 0.08%	1.98 ± 1.16%	0.03 ± 0.02%	0.57 ± 0.12%
$m = 2$	*accuracy*		90.67%	90.16 ± 0.79%	90.26 ± 0.48%	89.92 ± 2.86%	90.32 ± 0.10%	90.29 ± 0.20%

Table 8: We show how different levels of symmetric noise $\epsilon^- = \epsilon^+ = \epsilon$ affect the classifiers' performance on adult dataset. SL: Surrogate Loss. GPL: Group Peer Loss. We highlight substantial improvement of fairness in green and sever violation in red.

Noise ϵ	Metric	Clean	Corrupt	SL	GPL
0.1	*violation*	0.47%	3.91%	5.15%	1.41%
	accuracy	83.76%	83.22%	82.73%	82.71%
0.2	*violation*	0.47%	3.83%	3.98%	1.49%
	accuracy	83.75%	82.08%	82.54%	82.16%
0.3	*violation*	0.47%	7.23%	3.63%	1.22%
	accuracy	83.76%	81.36%	82.01%	81.24%
0.4	*violation*	0.47%	5.14%	1.13%	3.1%
	accuracy	83.76%	79.58%	80.62%	80.21%

data, training on corrupted data substantially increases fairness violations, even for relatively low noise rates. The SL and GPL columns show that our fair ERM approaches can effectively mitigate the biases. This holds true even when increasing the noise rate.

5.4 Insights on running on data directly, without adding additional noise

We evaluate our algorithm on the clean adult and arrest datasets as shown in Table 9. On the arrest dataset, our methods achieve a similar performance of accuracy compared with the Clean baseline, but we do observe a consistent drop of fairness violations on the arrest dataset. The fairness violation of our methods on adult dataset is not as good as that of Clean baseline. This fact may imply the possibility that the arrest dataset contains more human biases in labels than the adult dataset. The small drop in accuracy and (sometimes) in fairness is due to the additional noise estimation step, which introduces another layer of complication - this is the price we pay for dealing with potentially highly noisy labels.

Table 9: We examine the performance of our methods on the clean adult and arrest datasets. Clean: train a fair classifier directly with equal odds constraint. SL: Surrogate Loss with estimated noise parameters. GPL: Group Peer Loss with estimated noise parameters. The values after ± are the standard deviation.

	adult		arrest	
Method	*accuracy*	*violation*	*accuracy*	*violation*
Clean	83.76 ± 0.0	0.47 ± 0.0	65.46 ± 0.0	4.46 ± 0.0
SL	76.97 ± 0.24	3.51 ± 0.24	63.07 ± 0.44	2.90 ± 0.72
GPL	81.20 ± 0.19	3.76 ± 0.19	64.98 ± 0.40	1.85 ± 0.36

6 CONCLUDING REMARKS, LIMITATIONS AND FUTURE WORKS

We have demonstrated, both theoretically and empirically, that naively enforcing parity constraints without taking noisy labels into consideration can indeed do harm. Our results show the importance of accounting for group-dependent label-noise when performing ERM subject to fairness constraints. In realistic applications, such as criminal justice and evaluating loan applications, labels are often contaminated by human biases against a certain protected group. The insights gained from this work forewarn decision-makers that improperly mitigating unfairness might do harm on the clean groups. Our two fairness-aware ERM frameworks are an important step toward addressing this problem.

Our work extends a growing body of methods for training classifiers to provide equal opportunity to members of different subgroups within a population. Our new contribution is to address situations where feature and label information for one or more of the subgroups has been recorded less faithfully than for members of other subgroups. Just one example of this, discussed in the text, is the significant disparity in the quality of evaluations for males and females which occur in both medical and academic contexts. These disparities can and do have significant impacts on the quality of life for members of each group, and are well worth addressing.

This work shows how applying existing techniques for mitigating bias in classifiers can actually increase inequality in outcomes, if disparities in the accuracy of training data are not accounted for. We offer new methods for addressing these problems as well. We believe that applying our methods *thoughtfully* will improve existing methods of bias mitigation in machine learning. Our technical solutions and solvers should be of interests to machine learning practitioners/researchers, as well as to policy makers when decided to use classification tools but face a training data with low-quality annotations.

Our work has limitations. Our selection of data sets is limited: we rely on synthetic training data corruption in order to test our methods. This limitation arises from the unavailability of such sensitive data sets for the broader research community. Both this research, and the methods whose shortcomings we have attempted to address, should be re-examined as richer data sets become available for studying disparities in the quality of information recording between members of different subgroups. The lack of relevant data for studying unfairness in machine learning, and the concerns about how to acquire such data while preserving the privacy of people concerned, is itself an important question in this area, although we do not address it in this work.

It is also possible that blind and uncareful application of our approach (by improperly attempting to correct otherwise accurate labels) may in fact create classifiers that produce even greater inequality, or lead to other problems that we have not foreseen. The temptation to apply our methods simply for the purpose of making existing models seem “more fair,” especially to unsuspecting downstream users, is very real. We very much discourage the use of our research in this fashion.

Both the limitations and the insights gained through this work underscore an important underlying message: that blind application of bias mitigation techniques in machine learning may do more harm than good.

ACKNOWLEDGMENTS

The authors would like to thank Berk Ustun for many inspiring early discussions on the practical scenarios when group-dependent noisy labels can cause harm. The authors also thank anonymous reviewers for their constructive comments. This work is partially supported by the National Science Foundation (NSF) under grants IIS-2007951, CCF-2023495, and CCF-1740850.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A Reductions Approach to Fair Classification. In *ICML (Proceedings of Machine Learning Research)*, Jennifer G. Dy and Andreas Krause (Eds.), Vol. 80. PMLR, 60–69. http://proceedings.mlr.press/v80/agarwal18a.html

[2] Michelle Alexander. 2012. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* New Press, New York.

[3] Dana Angluin and Philip Laird. 1988. Learning from noisy examples. *Machine Learning* 2, 4 (1988), 343–370.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias.

[5] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. 2006. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* 101, 473 (2006), 138–156.

[6] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (2004), 991–1013.

[7] Avrim Blum and Kevin Stangl. 2019. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? arXiv:cs.LG/1912.01094

[8] Brain Tumour Charity. 2016. Finding Myself in Your Arms: The Reality of Brain Tumour Treatment and Care.

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency (FAT).* ACM, 77–91.

[10] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).* Association for Computing Machinery, New York, NY, USA, 319–328. https://doi.org/10.1145/3287560.3287586

[11] Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. 2019. On Symmetric Losses for Learning from Corrupted Labels. In *International Conference on Machine Learning.* 961–970.

[12] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 3539–3550. http://papers.nips.cc/paper/7613-why-is-my-classifier-discriminatory.pdf

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

[14] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580.

[15] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[16] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 119–133.

[17] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Conference on Knowledge Discovery and Data Mining (KDD).* ACM, 259–268.

[18] Riccardo Fogliato, Alexandra Chouldechova, and Max G'Sell. 2020. Fairness Evaluation in Presence of Biased Noisy Labels *(Proceedings of Machine Learning Research)*, Silvia Chiappa and Roberto Calandra (Eds.), Vol. 108. PMLR, Online, 2325–2336. http://proceedings.mlr.press/v108/fogliato20a.html

[19] Benoit Frenay and Michel Verleysen. 2014. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 5 (2014), 845–869.

[20] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine* 178, 11 (2018), 1544.

[21] Maya R. Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy Fairness. *CoRR* abs/1806.11212 (2018). arXiv:1806.11212 http://arxiv.org/abs/1806.11212

[22] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16).* Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[23] Heinrich Jiang and Ofir Nachum. 2019. Identifying and Correcting Label Bias in Machine Learning. *CoRR* abs/1901.04966 (2019). arXiv:1901.04966 http://arxiv.org/abs/1901.04966

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

[25] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2017. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *arXiv preprint arXiv:1711.05144* (2017).

[26] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. 2019. Noise-tolerant fair classification. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 294–306. http://papers.nips.cc/paper/8322-noise-tolerant-fair-classification.pdf

[27] Sara M. Lindberg, Janet Shibley Hyde, Jennifer L. Petersen, and Marcia C. Linn. 2010. New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin* 136, 6 (2010), 1123–1135.

[28] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 8135–8145.

[29] Tongliang Liu and Dacheng Tao. 2016. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence* 38, 3 (2016), 447–461.

[30] Yang Liu and Hongyi Guo. 2020. Peer Loss Functions: Learning from Noisy Labels without Knowing Noise Rates. arXiv:cs.LG/1910.03231

[31] N. Manwani and P. S. Sastry. 2013. Noise Tolerance Under Risk Minimization. *IEEE Transactions on Cybernetics* 43, 3 (2013), 1146–1151.

[32] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. 2015. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*. 125–134.

[33] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification *(Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 107–118. http://proceedings.mlr.press/v81/menon18a.html

[34] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with Noisy Labels. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1196–1204. http://papers.nips.cc/paper/5073-learning-with-noisy-labels.pdf

[35] Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2019. Confident Learning: Estimating Uncertainty in Dataset Labels. *ArXiv* abs/1911.00068 (2019).

[36] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[37] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2233–2241. https://doi.org/10.1109/CVPR.2017.240

[38] Alice B Popejoy and Stephanie M Fullerton. 2016. Genomics is failing on diversity. *Nature News* 538, 7624 (2016), 161.

[39] Clayton Scott. 2015. A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels *(Proceedings of Machine Learning Research)*, Guy Lebanon and S. V. N. Vishwanathan (Eds.), Vol. 38. PMLR, San Diego, California, USA, 838–846. http://proceedings.mlr.press/v38/scott15.html

[40] Clayton Scott, Gilles Blanchard, and Gregory Handy. 2013. Classification with Asymmetric Label Noise: Consistency and Maximal Denoising. *ArXiv* abs/1303.1208 (2013).

[41] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without Harm: Decoupled Classifiers with Preference Guarantees *(Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 6373–6382. http://proceedings.mlr.press/v97/ustun19a.html

[42] Effy Vayena, Alessandro Blasimme, and I Glenn Cohen. 2018. Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine* 15, 11 (2018), e1002689.

[43] Linda F. Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series.

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

[45] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, Vol. 54. Proceedings of Machine Learning Research, 962–970.