Can Less be More? When Increasing-to-Balancing Label Noise Rates Considered Beneficial

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we answer the question when inserting label noise (*less informative labels*) can instead return us *more* accurate and fair models. We are primarily inspired by two observations that 1) increasing a certain class of instances' label noise to balance the noise rates (increasing-to-balancing) results in an easier learning problem; 2) Increasing-to-balancing improves fairness guarantees against label bias. In this paper, we will first quantify the trade-offs introduced by increasing a certain group of instances' label noise rate w.r.t. the learning difficulties and performance guarantees. We analytically demonstrate when such an increase proves to be beneficial, in terms of either improved generalization errors or the fairness guarantees. Then we present a method to leverage our idea of inserting label noise for the task of learning with noisy labels, either without or with a fairness constraint. The primary technical challenge we face is due to the fact that we would not know which data instances are suffering from higher noise, and we would not have the ground truth labels to verify any possible hypothesis. We propose a detection method that informs us which group of labels might suffer from higher noise, without using ground truth information. We formally establish the effectiveness of the proposed solution and demonstrate it with extensive experiments.

1 Introduction

The presence of training label noise is generally considered harmful. Typically the goal of learning with noisy labels is to improve the training by reducing the amount of noise in the training data [11, 5]. This paper discusses the feasibility of the opposite strategy and shows cases where adding label noise, leading to a scenario with *less* informative labels, will result in *more* accurate and fair models. We are primarily motivated by two observations:

Observation I: increasing a certain class of instances' noise rate to balance the noise rates results in an easier learning problem. When label noise is class-dependent (label class $Y = +1$ v.s. $Y = -1$), popular noise-tolerant learning algorithms typically require the knowledge of noise rates [22]. Mis-specified noise rates introduce additional errors, especially when the label noise is asymmetric. On the other hand, although increasing the noise rate of the lower class one to balanced error rates reduces the informativeness of the training labels, the balanced situation is generally regarded as an easier case to handle.

Observation II: Increasing-to-balancing improves fairness guarantees against label bias. When the label noise rates are group-dependent (data for the young group v.s. the old group), it has been reported in the literature [27] that imposing fairness constraints directly on the noisy and biased labels would instead reinforce the unfairness. Fixing the fairness constraints again requires the knowledge of the error rates. We will show that increasing and balancing noise rates allows us to impose fairness guarantees on the noisy training data directly, without knowing the noise parameters.

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

This paper first quantifies the trade-offs introduced by increasing a certain group of instances' label noise rate with respect to the loss of informative labels and the gained benefits for doing so. We analytically demonstrate when such an increase proves beneficial in terms of either improved generalization error or fairness guarantees. Then we present a method to leverage our idea of inserting label noise for the task of learning with noisy labels, either without or with a fairness constraint.

We face the primary technical challenge since we would not know which data instances are suffering from higher noise, and again we would not have the ground truth labels to verify any possible hypothesis. We propose a detection method that informs us which group of labels might suffer from higher noise without using ground truth information. The core discovery is a couple of quantities that check the agreements of noisy labels among local neighbors, which we will show to be sufficient to inform us of the class of labels with a higher noise rate. These quantities can be easily estimated using the noisy labels only. With this knowledge, we propose an algorithm (NOISE+) to gradually insert noise into the particular lower noise class of labels. The primary contributions of our paper summarize as follows:

- Our paper prototypes the idea of pre-processing noisy training labels for both the constrained and unconstrained learning problems. We demonstrate the possibility of improving the training model accuracy and fairness guarantees by increasing a certain class of instances' noise rates (increase-to-balance).
- To enable the deployment of our idea, we propose a detection algorithm to identify the class of instances with a higher label noise rate without using any ground truth label information.
- Our solution contributes to the learning with noisy labels literature by adding another tool that is robust to noise rate (noise transition matrix) estimation errors.

All omitted proofs can be found in the Appendix. The code for reproducing the experimental results is available in the supplementary material.

1.1 Related work

Our work is a contribution to the well-established literature of learning with noisy labels [25, 22, 24, 19, 11, 26, 30]. The classical solutions within the literature leveraged the knowledge of noise rates to perform loss correction [25, 22, 24], label correction [24, 20], loss reweighting [17], among many other treatments. The importance of knowing the correct noise rate parameters is clearly established. Follow-up works have provided tools for estimating the noise rates without accessing the ground truth label. These include the confident learning approach [23], anchor point [6], and more recently clustering based ones [32]. Nonetheless, the possibility of unintended harms due to mis-specified noise rates remains. More recent works have looked into robust loss functions that wouldn't require the knowledge of noise rate, hoping to improve the resistance to the mis-specifications [13, 31, 18]. Our work can be viewed as a contribution to this specific line of literature. Of independent interests, our noise increasing step that does not require ground truth label might help improve co-training [4] when no ground truth label is available.

Our work also contributes to the literature of fair classification with imperfect information [16, 29, 10, 28, 21, 3]. For instance, [16] considers amending noisy sensitive attributes by appropriately re-scaling the fairness tolerance but is only restricted to class-conditional random noise. [28] proposes two noise-resistant fair classification approaches by constructing unbiased estimators with group-dependent label noise. While most of existing works would either require the noise rates to be balanced between groups or the knowledge of the noise rates to perform noise correction [28], our work requires neither.

2 Preliminaries

We will consider a binary classification problem with d-dimensional feature space $X \in \mathbb{R}^d$ and label space $Y \in \{-1, +1\}$. (X, Y) generate according to distribution $\mathcal{D}$. Instead of accessing the clean data, we consider the setting where the learners only have access to noisy labels $\tilde{Y}$. The generation of $\tilde{Y}$ follows the noise transition model: $e_+(x) := \mathbb{P}(\tilde{Y} = -1 | Y = +1, X = x)$, $e_-(x) := \mathbb{P}(\tilde{Y} = +1 | Y = -1, X = x)$. We further assume $\tilde{Y}$ generates conditionally independently for different X. For now, we don't restrict how $e_+(x), e_-(x)$ depend on x but to enable our analysis, later we will. Denote the noisy distribution for $X, \tilde{Y}$ as $\tilde{\mathcal{D}}$.

In the second constrained learning setting, we assume that for each instance (X, Y) we also observe a sensitive attribute $Z \in \{a, b\}$ that defines two protected groups (e.g., senior v.s. young people) that we will later enforce fairness constraints over.

Specifically, we consider two learning problems, one unconstrained and another fairness constrained:

Unconstrained learning problem In the unconstrained setting, denote the collected training data as $\tilde{D} := \{(x_n, \tilde{y}_n)\}_{n=1}^N$. The goal is to train a classifier $h \in \mathcal{H}$ that minimizes the empirical risk $\mathbb{E}_{\mathcal{D}}[1(h(X) \neq Y)]$ with only accessing $\tilde{D}$, where $1(\cdot)$ is the 0-1 loss function. We will focus on the standard class-dependent noise rate model where $\underline{e_+(x) \equiv e_+, e_-(x) \equiv e_-, \forall x}$. That is the noise rates depend only on the true label class [22], but not the specific instance x when conditional on the true label. We will operate in the challenging asymmetric error rates setting when $e_+ \neq e_-$.

Fairness constrained learning problem In the fairness constrained setting, denote the collected training data as $\tilde{D} := \{(x_n, \tilde{y}_n, z_n)\}_{n=1}^N$. The goal is to train a classifier $h \in \mathcal{H}$ that maximizes the accuracy while satisfying a certain fairness constraint, measured by a fairness metric $F_z(h)$:

$$\min_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}[1(h(X) \neq Y)] \;\; s.t. \; |F_a(h) - F_b(h)| \leq \delta, \tag{1}$$

where in above $\delta > 0$ is a tolerance parameter for fairness violations. Examplery $F_a(h)$ includes the Equal Opportunity measure $F_a(h) := \mathbb{P}(h(X) = +1 | Y = +1, Z = a)$ [12]. Some others include the demographic parity $F_a(h) := \mathbb{P}(h(X) = +1 | Z = a)$ [7].

We assume the error rates are group dependent [28]: $\underline{e_+(x) = e_-(x) = e_z, \; z \in \{a, b\}}$. Within each group, we can further assume that the noise rates are class-dependent but then we will simply apply our solution within each group and across group in order. We remove this additional layer of complication to let our presentation stay concise.

Notations We define the following quantities that we will repeatedly use. We will denote by $R_{\mathcal{D}}(h) := \mathbb{E}_{\mathcal{D}}[1(h(X) \neq Y)]$ the true generalization error of a classifier h incurred on the clean distribution $\mathcal{D}$. For a classification-calibrated loss function, we will denote its risk as $R_{\ell,\mathcal{D}}(h) := \mathbb{E}_{\mathcal{D}}[\ell(h(X), Y)]$. Denote the empirical ℓ-risk of a classifier h on dataset D as $\hat{R}_{\ell,D}(h) := \frac{1}{|D|} \sum_{(x,y) \in D} \ell(h(x), y)$.

Our solution is generically applicable to multi-class settings, which we will demonstrate in our experiments. But for a clear exposition of our core idea, we will focus on the binary case.

3 Equalizing error rates can improve model accuracy

In this section we focus on the unconstrained learning setting. We will show that while increasing one label class' noise rate raises the learning difficulty, at the same time, it improves robustness to possible mis-specification of noise rates (e_+, e_-). We will quantify such tradeoff.

Loss correction Once having the knowledge of e_+, e_-, there exist different ways to improve training robustness when using noisy labels. We consider demonstrating our idea using the classical loss correction framework [22, 24]: denote by $\tilde{e}_+, \tilde{e}_-$ the estimated version of e_+, e_-. In practice, this knowledge can be obtained by using existing techniques [23, 17, 32]. For a given calibrated loss ℓ, the loss correction approach defines the following loss:

$$\hat{\ell}(h(x_n), \tilde{y}_n) := (1 - \tilde{e}_{-sgn(\tilde{y}_n)}) \cdot \ell(h(x_n), \tilde{y}_n) - \tilde{e}_{sgn(\tilde{y}_n)} \cdot \ell(h(x_n), -\tilde{y}_n) \tag{2}$$

where in above sgn is the sign function. A classifier will be trained following empirical risk minimiziation (ERM) using $\hat{\ell}$: $\hat{h}_\ell^* := \arg\min_{h \in \mathcal{H}} \sum_{n=1}^N \hat{\ell}(h(x_n), \tilde{y}_n)$. Note that compared to the standard loss correction, we omit the $1 - \tilde{e}_- - \tilde{e}_+$ term from the denonminator, which won't affect the minimizer of the empirical risk minimization problem.

Denote $h_{\ell,\mathcal{D}}^\diamond = \arg\min_{h \in \mathcal{H}} R_{\ell,\mathcal{D}}(h)$, the optimal classifier for $R_{\ell,\mathcal{D}}(h)$, and $\epsilon_M := \max\{|\tilde{e}_+ - e_+|, |\tilde{e}_- - e_-|\}$ the max model mis-specification error. Adapting the proof from [22] we show that:

Theorem 1. *When ℓ is L-Lipschitz and bounded by $\bar{\ell}$, let $L_1 := 4L, L_2 := 2\bar{\ell} > 0$, we have:*

$$R_{\ell,\mathcal{D}}(\hat{h}_\ell^*) - R_{\ell,\mathcal{D}}(h_{\ell,\mathcal{D}}^\diamond) \leq \underbrace{\frac{1}{1 - e_+ - e_-}}_{\textit{learning difficulty}} \cdot L_1 \cdot \mathcal{R}(\mathcal{H}) + \underbrace{\frac{\epsilon_M}{1 - e_+ - e_-}}_{\textit{Mis-specification}} \cdot L_2 + 2\sqrt{\frac{\log 1/\delta}{2N}}, \tag{3}$$

where in above $\mathcal{R}(\mathcal{H})$ denotes the Rademacher complexity of $\mathcal{H}$.

Clearly, e_+, e_- controls the learning difficulty: a pair of e_+, e_- with higher $e_+ + e_-$ induce larger generalization error bound. On the other hand, the potential imperfect estimates of them introduces additional error due to the mis-specification.

When equalizing error rate improves performance Now we first provide intuitions for why increasing noise rate to balance e_+ and e_- might be considered helpful. Without loss of generality, suppose that $e_+ > e_-$. When we increase e_- to match e_+ such that $e = e_+ = e_-$ (later we will explain how we do so), denote by $\hat{Y}$ the newly generated noisy label with e symmetric error rates for both groups. Denote correspondingly $\hat{y}_n$ the newly generated training labels. We first show that:

Lemma 2. *Minimizing $\mathbb{P}(h(X) \neq \hat{Y})$ is equivalent with minimizing $\mathbb{P}(h(X) \neq Y)$.*

The above lemma says that the minimizer of $\mathbb{P}(h(X) \neq \hat{Y})$ is equivalent to the optimal classifier for $\mathbb{P}(h(X) \neq Y)$ when the hypothesis space $\mathcal{H}$ covers the optimal h! Therefore when ℓ is classification-calibrated on the noisy distribution $\hat{\mathcal{D}}$ we know that the optimal classifier for $\mathbb{E}_{\tilde{\mathcal{D}}}[\ell(h(X), \tilde{Y})]$ will be the same as $\mathbb{E}_{\mathcal{D}}[\ell(h(X), Y)]$, i.e., when the error rates are symmetric, the training would not need the error rates information, and performing ERM on the noisy data:

$$\tilde{h}^*_\ell := \arg\min_{h \in \mathcal{H}} \sum_{n=1}^{N} \ell(h(x_n), \hat{y}_n) \tag{4}$$

suffices to find the optimal classifier for the clean distribution. The above argument is slightly trickier when considering a finite hypothesis space $\mathcal{H}$. Denote by $-h$ the classifier that always flips the prediction from h. Then we prove that

Theorem 3. *When 1) ℓ is symmetric s.t. $\ell(h(x), -y) = \ell(-h(x), y)$ and 2) $\hat{R}_{\ell,\tilde{D}}(-\tilde{h}^*_\ell) - \hat{R}_{\ell,\tilde{D}}(-h^\diamond_{\ell,\mathcal{D}}) \geq 0$, we have $R_{\ell,\mathcal{D}}(\tilde{h}^*_\ell) - R_{\ell,\mathcal{D}}(h^\diamond_{\ell,\mathcal{D}}) \leq \underbrace{\frac{1}{1-2e}}_{\textit{learning difficulty}} \cdot L_1 \cdot \mathcal{R}(\mathcal{H}) + 2\sqrt{\frac{\log 1/\delta}{2N}}$.*

The second condition above is simply stating that the opposite classifier of the empirically optimal one incurs high empirical loss, and particularly higher than the opposite one of $h^\diamond_{\ell,\mathcal{D}}$. Now comparing Theorem 1 and 3 we observe that equalizing noise rates increases the learning difficulty by

$$\left| \frac{1}{1-2e} - \frac{1}{1-e_+-e_-} \right| \cdot L_1 \cdot \mathcal{R}(\mathcal{H}) = \frac{|e_+ - e_-|}{|1-e_+-e_-| \cdot |1-2e_+|} \cdot L_1 \cdot \mathcal{R}(\mathcal{H}).$$

On the other hand, equalizing noise rates avoids the error due to mis-specifying e_+, e_- by $\frac{\epsilon_M}{1-e_+-e_-} \cdot L_2$. Therefore it is considered helpful to increase e_- to e_+ when

$$\epsilon_M \geq \frac{|e_+ - e_-| \cdot L_1 \cdot \mathcal{R}(\mathcal{H})}{|1-2e_+| \cdot L_2}$$

The above condition is likely to hold when the gap between noise rates $|e_+ - e_-|$ is sufficiently small, and when we don't have high confidence in the estimation of e_+, e_- (therefore a high ϵ_M).

4 Equalizing error rates improves fairness guarantee

Now we consider the fairness constrained learning problem and in this section we will show that equalizing noise rates helps improve fairness guarantee when only accessing the noisy labels.

There are generally two types of fairness constraints: ones that do not depend on the label and the ones do. For the ones that do not, for example, demographic parity $\mathbb{P}(h(X) = +1|Z = a) = \mathbb{P}(h(X) = +1|Z = b)$, the existence of noisy labels do not impose additional challenges when enforcing such constraints. For the ones that do, typically they are functions of the true positive rates (TPR, also known as the equal opportunity measure) and false positive rates (FPR). It was shown in a recent paper [28] that, equalizing TPR and FPR on the noisy label leads to disparities when $e_a \neq e_b$. Later in the section, we will reprorduce the same observations. We will focus on the second type of constraints, and in particular equalizing TPR and FPR using the noisy labels.

Having noted that equalizing TPR and FPR on the noisy labels leads to unintended consequences, the literature has observed some recent works on performing fairness constraint correction [28]. Similar to the loss correction approaches, such correction methods would again require the knowledge of e_a, e_b, which is subject to mis-specification error. Define

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

We first derive the following relationship:

Lemma 4. *Define $C_{z,1} := 0.5 \cdot e_z, C_{z,2} := 0.5 \cdot (1 - e_z)$. The TPR defined on the clean data relates to the noisy ones as follows:*

$$TPR_z(h) = \frac{C_{z,1}\widetilde{TPR}_z(h) - C_{z,2} \cdot \widetilde{FPR}_z(h)}{e_z - 0.5},\ \ FPR_z(h) = \frac{C_{z,1}\widetilde{FPR}_z(h) - C_{z,2}TPR_z(h)}{e_z - 0.5} \tag{5}$$

The above lemma first shows since two groups a, b might have different $C_{z,1}, C_{z,2}$, equalizing $\widetilde{\text{TPR}}, \widetilde{\text{FPR}}$ using the noisy labels naively is insufficient to guarantee equalization of TPR, FPR.

Unfairness due to model error On the other hand, Lemma 4 points out a way to perform constraint correction using the knowledge of e_a, e_b. Suppose we suffer from the following model specifications: $|\tilde{e}_a - e_a| \geq \epsilon, |\tilde{e}_b - e_b| \geq \epsilon$. Denote the corrected TPR and FRP using the noisy $\widetilde{\text{TPR}}$ and $\widetilde{\text{FPR}}$ as well as $\tilde{e}_a, \tilde{e}_b$ as $\text{TPR}^c_z(h)$, $\text{FPR}^c_z(h)$: Define $\tilde{C}_{z,1} := 0.5 \cdot \tilde{e}_z, \tilde{C}_{z,2} := 0.5 \cdot (1 - \tilde{e}_z)$. Define:

$$\text{TPR}^c_z(h) = \frac{\tilde{C}_{z,1}\widetilde{\text{TPR}}_z(h) - \tilde{C}_{z,2} \cdot \widetilde{\text{FPR}}_z(h)}{\tilde{e}_z - 0.5},\ \ \text{FPR}^c_z(h) = \frac{\tilde{C}_{z,1}\widetilde{\text{FPR}}_z(h) - \tilde{C}_{z,2}\text{TPR}_z(h)}{\tilde{e}_z - 0.5} \tag{6}$$

Then we prove that

Lemma 5. $|TPR^c_z(h) - TPR_z(h)| \geq \frac{\epsilon \cdot \widetilde{TPR}_z(h)}{(2e_z - 1)(2\tilde{e}_a - 1)}$.

Lemma 5 therefore established a positive TPR ineuality as long as $\widetilde{\text{TPR}}_z(h) > 0$. But as a consequence of the Lemma 4, we also immediately know that

Theorem 6. *When $e_a = e_b$, equalizing $\widetilde{TPR}$ and $\widetilde{FPR}$ using noisy labels suffices to equalizing the true TPR and FPR.*

This is simply because equalizing e_a and e_b also equalizes both $C_{z,1}$ and $C_{z,2}$. Theorem 6 helps us establish a strong fairness guarantee: as long as we know the training data has balanced label noise rates across different groups, enforcing the constraints directly on the noisy data suffices to guarantee the equality between true TPR and FPR.

Our above observation points out increasing one group's label noise to match another helps establish the fairness guarantees more easily. Again suppose that $e_a > e_b$. When we increase e_b to match e_a such that $e = e_a = e_b$, denote by $\hat{Y}$ the newly generated noisy label with e symmetric error rates for both groups. Denote correspondingly $\hat{y}_n$ the newly generated training labels. Then performing fairness constrained ERM directly on $\hat{D} := \{(x_n, \hat{y}_n)\}_n$ will help us approximately equalize TPR and FPR between two groups when the number of training data is sufficiently large.

$$\tilde{h}^*_\ell := \arg\min_{h \in \mathcal{H}} \sum_{n=1}^{N} \ell(h(x_n), \hat{y}_n)\ \ s.t.\ \ |\tilde{F}_a(h) - \tilde{F}_b(h)| \leq \delta, \tag{7}$$

where $\tilde{F}_a(h), \tilde{F}_b(h)$ are the empirically computed fairness measures using $\hat{D}$.

5 Identifying cleaner class without using ground truth label

With the above demonstrated benefits of balancing the error rates, the remaining technical question is how do we know if $e_+ > e_-$ or not? how does e_a compare to e_b? so that we will be able to add label noise to the right group of data. Of course, when we have access to some ground truth labels, we will be able to have a rough estimate and carry on to insert label noise (e.g., by further randomly flipping

the noisy labels, which we will discuss at the end of this section). In this section, we describe another approach without the need of ground truth label. We prove its sufficiency in identifying the noisier class of labels. We demonstrate our main idea using the unconstrained learning setting for detecting $sgn(e_+ - e_-)$, but the idea easily generalizes to the fairness setting to detect the order of e_a and e_b. We first present the following definition:

Definition 7 (Clusterability)**.** *We say the dataset D satisfies 2-NN clusterability if each instance x shares the same true label class with its two nearest neighbors measured by $||x - x'||_2$.*

For the rest of the section, we will assume that D satisfies 2-NN clusterability. 2-NN was similarly introduced in a recent work [32] and has been shown to be a requirement that is mild to satisfy. Now we define the following two quantities that are central for the development of our idea. For an arbitrary noisy label $\tilde{Y}_1$, denote the ones for its two nearest neighbor instances' as $\tilde{Y}_2, \tilde{Y}_3$. Define the following agreement measures:

Definition 8 (2-NN Agreements)**.**

$$\textit{Positive Agreements} \quad PA_{\mathcal{D}} := \mathbb{P}(\tilde{Y}_2 = \tilde{Y}_3 = +1 | \tilde{Y}_1 = +1) \tag{8}$$

$$\textit{Negative Agreements} \quad NA_{\mathcal{D}} := \mathbb{P}(\tilde{Y}_2 = \tilde{Y}_3 = -1 | \tilde{Y}_1 = -1) \tag{9}$$

Intuitively speaking, $\text{PA}_{\mathcal{D}}$ computes the likelihood of the neighbor data points "agreeing" on a positive label. $\text{NA}_{\mathcal{D}}$ computes the one for the negative label. Now we will first sub-sample the noisy distribution and compute $\text{PA}_{\mathcal{D}}, \text{NA}_{\mathcal{D}}$: **Step 1** Sample $\tilde{Y}$ such that $\mathbb{P}(\tilde{Y} = +1) = \mathbb{P}(\tilde{Y} = -1) = 0.5$. Denote this resampled and label balanced distribution as $\mathcal{D}^\diamond$. **Step 2** Compute $\text{NA}_{\mathcal{D}^\diamond}, \text{PA}_{\mathcal{D}^\diamond}$. Next we prove that knowing $\text{NA}_{\mathcal{D}^\diamond}, \text{PA}_{\mathcal{D}^\diamond}$ suffices to inform $sgn(e_+ - e_-)$:

Theorem 9. *When $e_+, e_- < 0.5$ and D satisfies 2-NN clusterability, $PA_{\mathcal{D}^\diamond}, NA_{\mathcal{D}^\diamond}$ relate to e_+, e_- in the following way: $PA_{\mathcal{D}^\diamond} - NA_{\mathcal{D}^\diamond} = 2(0.5 - e_+)(0.5 - e_-)(e_- - e_+)$. Then if $PA_{\mathcal{D}^\diamond} > NA_{\mathcal{D}^\diamond}$, we know that $e_+ < e_-$; otherwise $e_+ > e_-$.*

Estimating $\text{PA}_{\mathcal{D}^\diamond}, \text{NA}_{\mathcal{D}^\diamond}$ using $\mathcal{D}^\diamond$ Note that both $\text{PA}_{\mathcal{D}^\diamond}$ and $\text{NA}_{\mathcal{D}^\diamond}$, $\mathbb{P}(\tilde{Y}_2 = \tilde{Y}_3 = +1 | \tilde{Y}_1 = +1)$ and $\mathbb{P}(\tilde{Y}_2 = \tilde{Y}_3 = -1 | \tilde{Y}_1 = -1)$ can be counted from the noisy data without accessing the ground truth. For instance for $\text{PA}_{\mathcal{D}^\diamond}$:

- **Step 1** Find all x_n in $\mathcal{D}^\diamond$ with $\tilde{y}_n = +1$ and their 2-NN n_1, n_2.
- **Step 2** Count $\frac{\#\{n:\, \tilde{y}_n = \tilde{y}_{n_1} = \tilde{y}_{n_2} = +1\}}{\#\{n:\, \tilde{y}_n = +1\}}$.

We show numpy-like pseudocode for an implementation of estimating PA in Figure 3 in Appendix A.10.

5.1 Our solution: NOISE+

Before presenting the complete solution, we first show that randomly flipping $\tilde{Y} = y$ with the smaller $e_{sgn(y)}$ by a small constant ϵ monotonically decreases the gap between noise rates $|e_+ - e_-|$. Without the loss of generality suppose $e_+ < e_-$, and we will only flip the $\tilde{Y} = +1$ labels (but not flipping the noisy labels with $\tilde{Y} = -1$). We show numpy-like pseudocode for the flipping function in Figure 2, and the implementation for exclusively flipping $\tilde{Y} = -1$ labels is symmetric. Denote by $\hat{Y}$ as the randomly flipped version of $\tilde{Y}$ by ϵ, and denote $\hat{e}_+ := \mathbb{P}(\hat{Y} = -1 | Y = +1), \hat{e}_- := \mathbb{P}(\hat{Y} = +1 | Y = -1)$. we prove:

Proposition 10. $\hat{e}_+ = (1 - e_+)\epsilon + e_+, \;\; \hat{e}_- = (1 - \epsilon)e_-$. *Further, the new gap between the noise rates of the flipped label $\hat{Y}$ is a monotonic function of ϵ: $\hat{e}_+ - \hat{e}_- = (1 - e_+ + e_-)\epsilon + e_+ - e_-$.*

Since $1 - e_+ + e_- > 0$, when ϵ is small, the above derivation shows the effectiveness in reducing the noise rate gap by randomly flipping the noisy labels that correspond to the class with lower noise rate. The only remaining question is how to find the optimal ϵ without knowing the true e_+ and e_-. Calling Theorem 9, we know $\text{PA}_{\mathcal{D}^\diamond} - \text{NA}_{\mathcal{D}^\diamond} = 2(0.5 - \hat{e}_+)(0.5 - \hat{e}_-)(\hat{e}_- - \hat{e}_+)$. Denote by $f(\epsilon) := 0.5 \cdot (\text{PA}_{\mathcal{D}^\diamond} - \text{NA}_{\mathcal{D}^\diamond})$. There are three solutions for $f(\epsilon) = 0$: $\epsilon_1 = 1 - \frac{0.5}{e_-} < 0$, $\epsilon_2 = \frac{e_- - e_+}{1 - e_+ + e_-}$, $\epsilon_3 = \frac{0.5 - e_+}{1 - e_+}$. We should notice that $\epsilon_2 = \frac{e_- - e_+}{1 - e_+ + e_-} < \frac{(e_- - e_+) + (1 - e_+ - e_-)}{(1 - e_+ + e_-) + (1 - e_+ - e_-)} = \frac{1 - 2e_+}{2(1 - e_+)} = \epsilon_3$.

The monotonicity of $f(\epsilon)$ from 0 to ϵ_2 (our targeted root) motivates us to look for a proper ϵ by a binary search procedure. Suppose we have two different flipping parameters $\epsilon_l < \epsilon_r$. Initially their gaps between the counted agreements are C_l and C_r satisfying $C_l > 0 > C_r$. In each iteration, we try a new flip parameter $\epsilon_{\text{mid}} = (\epsilon_l + \epsilon_r)/2$, and check whether the new gap C_{mid} is bounded by a threshold γ. If $-\gamma \leq C_{\text{mid}} \leq \gamma$, we return the labels flipped by ϵ_{mid}. Otherwise, we update the values of ϵ_l and ϵ_r according to C_{mid}. We summarize NOISE+ in Algorithm 1. Here, we initialize $\epsilon_r = 0.3$ which empirically succeeds in varied noise settings. In case $\epsilon_r = 0.3$ fails, we can grid search a variety of ϵ values (e.g., 0.1, 0.2) satisfying $C_r < 0$ as the initial ϵ_r to proceed our binary search algorithm.

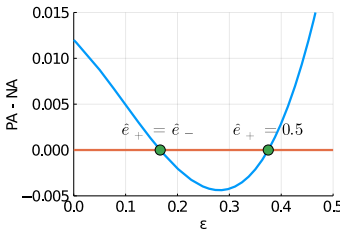

Figure 1: Agreement gap PA$-$NA varies for different ϵ. There are only two positive roots for PA $-$ NA $= 0$. The less one results in $\hat{\epsilon}_+ = \hat{\epsilon}_-$.

Algorithm 1 NOISE+: A binary search algorithm for balancing noise rates.

Ensure: the positive class has a higher noise rate than the negative class
Require: $\gamma > 0, \epsilon_l = 0, \epsilon_r = 0.3$
  Resample a balanced set $\mathcal{D}^\diamond$ from $\tilde{D}$;
  Initialize $\mathcal{D}_l = \mathcal{D}^\diamond, \mathcal{D}_r = \mathsf{Flip}(\mathcal{D}^\diamond, \epsilon_r)$;
  Estimate $\text{PA}_{\mathcal{D}_l}, \text{PA}_{\mathcal{D}_r}, \text{NA}_{\mathcal{D}_l}, \text{NA}_{\mathcal{D}_r}$.
  while $\text{PA}_{\mathcal{D}_l} - \text{NA}_{\mathcal{D}_l} > \gamma$ and $\text{PA}_{\mathcal{D}_r} - \text{NA}_{\mathcal{D}_r} < -\gamma$ **do**
    $\epsilon_{\text{mid}} \leftarrow (\epsilon_l + \epsilon_r)/2, \mathcal{D}_{\text{mid}} \leftarrow \mathsf{Flip}(\mathcal{D}^\diamond, \epsilon_{\text{mid}})$;
    Estimate $\text{PA}_{\mathcal{D}_{\text{mid}}}$ and $\text{NA}_{\mathcal{D}_{\text{mid}}}$;
    if $\text{PA}_{\mathcal{D}_{\text{mid}}} < \text{NA}_{\mathcal{D}_{\text{mid}}} - \gamma$ **then**
      /* ϵ_{mid} is at the right of the root */
      $\epsilon_r \leftarrow \epsilon_{\text{mid}}, \mathcal{D}_r \leftarrow \mathcal{D}_{\text{mid}}$;
      $\text{PA}_{\mathcal{D}_r} \leftarrow \text{PA}_{\mathcal{D}_{\text{mid}}}, \text{NA}_{\mathcal{D}_r} \leftarrow \text{NA}_{\mathcal{D}_{\text{mid}}}$;
    else if $\text{PA}_{\mathcal{D}_{\text{mid}}} > \text{NA}_{\mathcal{D}_{\text{mid}}} + \gamma$ **then**
      /* ϵ_{mid} is at the left of the root */
      $\epsilon_l \leftarrow \epsilon_{\text{mid}}, \mathcal{D}_l \leftarrow \mathcal{D}_{\text{mid}}$;
      $\text{PA}_{\mathcal{D}_l} \leftarrow \text{PA}_{\mathcal{D}_{\text{mid}}}, \text{NA}_{\mathcal{D}_l} \leftarrow \text{NA}_{\mathcal{D}_{\text{mid}}}$;
    else
      return $\hat{D} = \mathsf{Flip}(\tilde{D}, (\epsilon_l + \epsilon_r)/2)$;
    end if
  end while
  return unsuccessful;

```
def flip(dataset, epsilon):
    # unpack the dataset
    (X, y) = dataset
    # filter positives
    y_plus = y[y > 0]
    # flip the positive
        labels w.p. epsilon
    is_flipped = numpy.
        random.binomial(1,
        epsilon, y_plus.size
        )
    y[y > 0] = is_flipped *
        (-y_plus) + (1 -
        is_flipped) * y_plus
    # pack and return the
        dataset
    return (X, y)
```

Figure 2: Numpy-like pseudocode for an implementation of Flip function. Flip takes the dataset and a small probability ϵ as input, and only flips positive examples with probability ϵ. The code for exclusively flipping negative examples is symmetric.

Which baseline loss to use ℓ Our algorithm is particularly suitable for losses that do not require the knowledge of noise rates. Standard CE loss would certainly be applicable. Other robust losses are great options. We note a recently proposed loss, peer loss [18], does not require the specification of noise rates (as promoted in [28] for the fairness constrained setting too). We believe it is particularly suitable for our NOISE+. We will test it in the experiment section.

Fairness constrained learning Our above derivations and observations can be similarly applied to the fairness constrained setting. To do so, the only difference lies in that now $\text{PA}_{\mathcal{D}^\diamond}, \text{NA}_{\mathcal{D}^\diamond}$ are corresponding to the agreements within each of the two groups a, b instead of the two label class group $\tilde{Y} = +1$ and $\tilde{Y} = -1$. We provide details in the Appendix A.8 without repeating them here.

Extension to multi-class As explained at the beginning, our algorithm can largely extend to the multi-class/group setting. The primary requirement of the extension is to extend the definition of $\text{PA}_{\mathcal{D}^\diamond}, \text{NA}_{\mathcal{D}^\diamond}$ to each label class/group. We defer the details to Appendix A.9.

Training Once we have the output and noise-added dataset $\hat{D} := \{(x_n, \hat{y}_n)\}_n$ from Algorithm 1, we will use it directly in our ERM framework for the unconstrained case as in Eqn. (4) and will use it in the fairness constrained ERM directly too Eqn. (7).

Table 1: Binary classification accuracy of compared methods on 3 datasets across different levels of noise rates. Mis. SL: surrogate loss [22] with misspecified parameters. Est. SL: surrogate loss [22] with estimated parameters. CE: vanilla cross entropy. Peer: peer loss function [18]. All methods are trained with one-layer perceptron with the same hyper-parameters. For each noise setting, we average across 5 runs and report the mean and standard deviation. We find that the increasing-to-balancing can boost the vanilla cross entropy on all the noise settings. We highlight any boost of increasing-to-balancing for vanilla CE in green, and the best accuracy in orange.

			BASELINES (LESS NOISE)				NOISE+ (MORE NOISE)	
Dataset	e_-	e_+	Mis. SL	Est. SL	CE	Peer	CE	Peer
Adult	0.0	0.1	72.79 ± 0.34	72.64 ± 0.38	72.63 ± 0.29	72.77 ± 0.32	73.62 ± 0.37	73.86 ± 0.41
$n = 48,842$	0.0	0.2	72.27 ± 0.39	72.13 ± 0.37	71.26 ± 0.38	71.95 ± 0.34	72.73 ± 0.71	73.52 ± 0.58
$d = 28$	0.1	0.2	73.02 ± 0.50	72.68 ± 0.16	72.31 ± 0.25	72.88 ± 0.14	71.92 ± 1.98	73.81 ± 0.40
	0.1	0.3	72.44 ± 0.47	72.15 ± 0.23	69.06 ± 2.01	72.26 ± 0.43	69.53 ± 4.90	73.34 ± 1.27
	0.2	0.3	72.81 ± 0.51	72.43 ± 0.14	71.44 ± 0.93	72.78 ± 0.28	71.55 ± 2.04	73.75 ± 0.26
	0.2	0.4	72.06 ± 0.19	71.97 ± 0.41	63.49 ± 1.58	71.97 ± 0.37	65.99 ± 7.99	71.43 ± 2.26
Compas	0.0	0.1	66.36 ± 1.05	66.04 ± 1.14	66.16 ± 1.13	68.06 ± 0.70	67.14 ± 0.92	68.22 ± 0.68
$n = 7,168$	0.0	0.2	66.84 ± 0.69	66.06 ± 0.81	65.38 ± 1.40	68.03 ± 0.77	66.51 ± 1.90	68.40 ± 0.78
$d = 10$	0.1	0.2	66.41 ± 0.43	65.69 ± 0.57	65.91 ± 0.97	67.49 ± 0.40	66.54 ± 0.21	67.80 ± 0.44
	0.1	0.3	65.91 ± 0.42	65.22 ± 0.63	61.24 ± 0.70	67.36 ± 0.79	65.76 ± 2.09	68.05 ± 0.56
	0.2	0.3	65.06 ± 0.72	65.86 ± 1.69	65.06 ± 1.48	68.02 ± 0.94	66.46 ± 1.27	68.04 ± 1.11
	0.2	0.4	64.82 ± 0.52	65.47 ± 0.46	59.68 ± 2.49	67.37 ± 0.54	63.85 ± 3.31	68.39 ± 0.56
Fairface	0.0	0.1	87.64 ± 0.03	87.75 ± 0.03	87.41 ± 0.11	87.58 ± 0.15	88.23 ± 0.07	88.49 ± 0.12
$n = 108,501$	0.0	0.2	85.22 ± 0.06	85.83 ± 0.08	85.08 ± 0.16	85.18 ± 0.16	88.55 ± 0.03	88.67 ± 0.03
$d = 50$	0.1	0.2	87.67 ± 0.07	87.56 ± 0.04	87.21 ± 0.08	87.28 ± 0.05	88.45 ± 0.06	88.65 ± 0.07
	0.1	0.3	72.03 ± 0.13	85.68 ± 0.07	83.20 ± 0.12	84.58 ± 0.09	87.81 ± 0.14	88.50 ± 0.12
	0.2	0.3	74.18 ± 0.20	87.34 ± 0.14	86.47 ± 0.09	87.00 ± 0.11	88.46 ± 0.08	88.58 ± 0.10
	0.2	0.4	58.30 ± 0.23	85.48 ± 0.09	78.33 ± 0.63	84.05 ± 0.13	81.90 ± 0.58	87.69 ± 0.15

6 Experiments

In order to verify the power of our increasing-to-balancing method, we conduct extensive experiments on both unconstrained learning and constrained learning settings. The datasets include: the UCI Adult Income dataset [9], the Compas recidivism dataset [2], Fairface [14] face attribute dataset, and CIFAR 10 [15] dataset. We defer more dataset details to Appendix B.1. Code for reproducing experimental results is available in the supplementary material.

6.1 Unconstrained learning

Setup. For unconstrained learning, we implement a one-layer perceptron for binary classification on Adult, Compas, and Fairface datasets. We flip the true labels on training set according to a set of asymmetric noise rates, train the models on the corrupted labels, and evaluate the accuracy on test set with clean labels. The baseline methods we compare include: surrogate loss [22] with misspecified noise rates (**Mis. SL**) and estimated noise rates (**Est. SL**) respectively, vanilla cross entropy (**CE**), and peer loss functions [18] (**Peer**). For Mis. SL, we randomly generate the noise parameters but ensure that the trace is equal to that of the true noise transition matrix. For Est. SL, we estimate the noise rates by adopting the confident learning procedure [23]. We note that the training of peer loss functions doesn't require the knowledge of noise rates, and follow their instructions to tune the hyper-parameters through cross-validation. For the same corrupted training labels, we deploy our increasing-to-balancing procedure, and evaluate the performance of cross entropy and peer loss functions again. We set the threshold γ used in Algorithm 1 as 0.1% on Adult and Fairface datasets, and loose it to 1% on Compas due to much smaller data size. We repeat all the methods 5 runs with different random seeds and report the mean and standard deviation.

Results. We show the experimental results in Table 1. We observe that our increasing-to-balancing method significantly improves the accuracy of the vanilla cross entropy on the majority of noise settings. It is notable that the cross entropy with increasing-to-balancing has a comparable performance with Est. SL, when the noise is relatively small. When the noise rates are large (e.g. 0.2 and 0.4), we observe a significant degradation for cross entropy, but increasing-to-balancing still boosts the accuracy. Moreover, we find that peer loss after increasing-to-balancing is more robust to noise, and dominantly achieve the highest accuracy.

Table 2: Accuracy of compared methods on CIFAR-10 dataset across different levels of noise gap. All methods are trained on image features extracted by a pre-trained vision transformer [8]. When noise gap is less than 0.3, CE with increasing-to-balancing reaches a higher accuracy than CE at a lower noise (highlighted in green). When noise gap is 0.3, balancing cannot compensate for the loss of increasing noise (highlighted in red).

	LESS NOISE				MORE NOISE	
noise gap	Mis. SL	Est. SL	CE	Peer	CE	Peer
0.1	70.90	85.76	88.03	89.66	88.69	89.90
0.2	80.51	86.34	88.43	89.36	89.01	90.08
0.3	81.30	90.61	89.78	90.24	87.98	89.92

Table 3: Constrained learning results with group-dependent label noise where $e_a = 0.2$ and $e_b = 0.4$. LR: naïve logistic regression without noise correction. GPR: group-weighted peer loss [28]. Peer: peer loss [18]. We highlight numbers of low fairness violation.

		LESS NOISE		MORE NOISE	
Dataset	Metrics	LR	GPL	LR	Peer
Adult	*accuracy*	72.73	71.2	71.88	73.02
	fairness	6.48	2.95	3.16	1.67
Compas	*accuracy*	62.60	66.03	66.22	64.15
	fairness	2.87	7.55	6.07	3.63
Fairface	*accuracy*	86.97	87.47	88.19	87.93
	fairness	5.87	4.70	1.38	0.25

Multi-class extension We text NOISE+ in multi-class setting by balancing the noise rates class by class. We evaluate the compared methods on CIFAR-10 [15] with more sophisticated noise transition matrix. We adopt the following procedure to generate the noise transition matrix: (1) manually set the diagonal elements at least 0.4; (2) permute the diagonal elements to increase the randomness; (3) fill out the non-diagonal elements randomly and ensure the sum of each column is 1. We observe that when noise gap is 0.1 and 0.2, CE and Peer with NOISE+ outperforms the pure CE and Peer, respectively. When noise gap is 0.3, balancing can't compensate for the loss of increasing noise.

6.2 Constrained learning

Setup. We conduct experiments with equal odds constraints [12] on Adult, Compas, and Fairface datasets. Especially we add heterogeneous noise, i.e., uneven noise rates for different group memberships but symmetric for different classes, into training labels but keep test labels clean. We stress-test the performance of the naïve logistic regression (LR) and group-weighted peer loss (GPL) [28] [1], when our increasing-to-balancing program is executed and not. We note that GPL degrades to peer loss when noise rates are homogeneous across protected groups [28], and directly apply peer loss after balancing. For a fair comparison, we use logistic regression as the base classifier for all the methods. We use the reductions approach [1] to solve the constrained optimization.

Results. We report the accuracy and fairness violation for $e_a = 0.2$ and $e_b = 0.4$ in Table 6 and defer more noise settings to Appendix B.4. We make the following observations: (1) All the methods retain a comparable accuracy. We conjecture that this benefits from the symmetric noise rates between positive and negative classes. (2) After balancing, both naïve LR and Peer significantly mitigate the fairness violations on Adult and Fairface datasets as a result of equalized noise rates across groups. (3) Peer loss generally has a lower fairness violation than naïve LR. This implies its capability of recovering unbiased classifiers. (4) We find out that the performance of LR with increasing-to-balancing is less proficient on Compas dataset because of uninformative features (only 10-dimensional features), and more robust on Fairface dataset with rich image information.

Concluding Remarks This paper elaborates the possibility of improving the training model accuracy and fairness guarantees by increasing a certain class of instances' noise rates. Our idea is based on a couple of observations that increasing-to-balance label noise rates can yield an easier learning problem that is robust to mis-specified noise rate parameters. The above robustness helps us improve generalization power as well as fairness guarantees. To deploy our idea, we propose a detection algorithm to identify the class of instances with a higher label noise rate, without using any ground truth label information. We believe this technique and our dataset pre-processing solution NOISE+ have the potential to find applications in other learning tasks when balancing label bias is desired.

Our work has limitations. Primarily, our method is built for class-conditional/dependent noise model. How our method generalizes to other human-level noise rate models, such as instance-dependent noise [6, 5], is of great interests to future investigations.

[1] We did not repeat testing surrogate loss as it is reported in [28] GPL appears to be a better and more robust solution. Also we focus on correcting fairness constraints in this section, both SL and GPL would use the same constraint corrections.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018.

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

[3] Pranjal Awasthi, Alex Beutel, Matthaeus Kleindessner, Jamie H. Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[4] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[5] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2021.

[6] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance-and label-dependent label noise. ICML, 2020.

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

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[9] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[10] Riccardo Fogliato, Alexandra Chouldechova, and Max G'Sell. Fairness evaluation in presence of biased noisy labels. volume 108 of *Proceedings of Machine Learning Research*, pages 2325–2336, Online, 26–28 Aug 2020. PMLR.

[11] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537, 2018.

[12] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

[13] Simon Jenni and Paolo Favaro. Deep bilevel learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–633, 2018.

[14] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[15] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[16] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. Noise-tolerant fair classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

[18] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. *ICML*, 2020.

[19] Nan Lu, Gang Niu, Aditya K Menon, and Masashi Sugiyama. On the minimal supervision for training any binary classifier from only unlabeled data. *arXiv preprint arXiv:1808.10585*, 2018.

[20] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? *arXiv preprint arXiv:2003.02819*, 2020.

[21] A. Mehrotra and L. E. Celis. Mitigating bias in set selection with noisy protected attributes. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[22] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

[23] Curtis G Northcutt, Lu Jiang, and Isaac L Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 2021.

[24] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

[25] Clayton Scott, Gilles Blanchard, Gregory Handy, Sara Pozzi, and Marek Flaska. Classification with asymmetric label noise: Consistency and maximal denoising. In *COLT*, pages 489–511, 2013.

[26] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5552–5560, 2018.

[27] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 526–536, New York, NY, USA, 2021. Association for Computing Machinery.

[28] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 526–536, 2021.

[29] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya R. Gupta, and Michael I. Jordan. Robust optimization for fairness with noisy protected groups. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[30] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *arXiv preprint arXiv:2006.07805*, 2020.

[31] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[32] Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. *arXiv preprint arXiv:2102.05291*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] One goal of our proposed method aims to improve fairness guarantees when training with noisy labels. The development of the paper is fully aware of any fairness concerns and 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. We are not aware of any negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] We have the full details in the Appendix (supplementary material).
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have included the details and codes in our supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]