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

Yang Liu
Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
yangliu@ucsc.edu

Jialu Wang
Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
faldict@ucsc.edu

Abstract

In this paper, we answer the question of when inserting label noise (*less informative labels*) can instead return us *more* accurate and fair models. We are primarily inspired by three observations: 1) In contrast to reducing label noise rates, increasing the noise rates is easy to implement; 2) Increasing a certain class of instances' label noise to balance the noise rates (increasing-to-balancing) results in an easier learning problem; 3) Increasing-to-balancing improves fairness guarantees against label bias. In this paper, we first quantify the trade-offs introduced by increasing a certain group of instances' label noise rate w.r.t. the loss of label informativeness and the lowered learning difficulties. We analytically demonstrate when such an increase is beneficial, in terms of either improved generalization power or the fairness guarantees. Then we present a method to insert label noise properly 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 labels. 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 [4, 11]. This paper discusses the feasibility of the opposite strategy and shows cases when adding label noise, leading to a scenario with *less* informative labels, will result in *more* accurate and fair models. We are primarily motivated by three observations:

Observation I: Reducing label noise is hard, but increasing it is easy. While the literature has provided us with solutions to perform data cleaning, they often require a highly customized training process. Their lack of theoretical rigor has also posed challenges when performing evaluations. On the other hand, as we will see later, increasing label noise is easy — one can always do so by randomly flipping the current noisy labels to increase it further.

Observation II: 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 [25]. However, the learner often needs to identify the unknown noise rates by some estimation procedure [26, 27, 36]. We articulate that the mis-specification of noise rates will introduce additional learning errors, especially when the label noise is asymmetric (see Theorem 1). On the other

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

hand, although increasing the noise rate of the lower class one to balanced error rates reduces the informativeness of the training labels, it is regarded as an easier case to handle — when error rates are balanced, invoking loss correction procedures becomes unnecessary (see Lemma 2 & Theorem 3).

Observation III: Increasing-to-balancing improves fairness guarantees against label bias. When the label noise rates are group-dependent (data from young group v.s. senior group), it has been reported in the literature that imposing fairness constraints directly on the noisy labels would instead reinforce the unfairness [30]. Fixing the fairness constraints again requires the knowledge of the label noise rates. We will show that increasing and balancing noise rates allows us to directly impose fairness guarantees on the noisy training data without knowing the noise rates (Theorem 6).

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. The primary technical challenge we face is that 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. In response, 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 metrics that check the agreements of noisy labels among local neighbors. These two metrics can be easily estimated using noisy labels only and are shown to be sufficient to inform us of the class of labels with a higher noise rate. With this knowledge, we propose an algorithm (NOISE+) to gradually insert noise into the lower noise class of labels. Our contributions summarize as follows:

- Our paper prototypes the idea of pre-processing noisy training labels for both the constrained and unconstrained learning problems. We show the possibility of improving the training accuracy and fairness guarantees by increasing a certain class of instances' noise rates (increasing-to-balancing).
- 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 at https://github.com/UCSC-REAL/CanLessBeMore.

1.1 Related work

Our work is a contribution to the well-established literature of learning with noisy labels [11, 22, 25, 27–29, 33]. The classical solutions within the literature leveraged the knowledge of noise rates to perform loss correction [25, 27, 28], label correction [23, 27], loss reweighting [19], 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 [26], anchor point [5], and more recently clustering-based ones [36]. Nonetheless, the possibility of unintended harm due to mis-specified noise rates remains. More recent works have looked into robust loss functions that would not require the knowledge of noise rate, hoping to improve the resistance to the mis-specifications [14, 21, 32, 34, 35]. Our work can be viewed as a contribution to this specific line of literature.

Our work also contributes to the understanding of fairness implications when learning with noisy labels and other imperfect information [3, 7, 10, 17, 20, 24, 30, 31]. For instance, [17] considers amending noisy sensitive attributes by appropriately re-scaling the fairness tolerance but is only restricted to class-conditional random noise. [30] proposes two noise-resistant fair classification approaches by constructing unbiased estimators with group-dependent label noise. While most of the existing works would either require the noise rates to be balanced between groups or the knowledge of the noise rates to perform noise correction [30], 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 do not 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.

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 $e_+(x) \equiv e_+, e_-(x) \equiv e_-, \forall x$. We assume we have informative labels: $e_+, e_- < 0.5$. Hereby the noise rates depend only on the true label class [25], 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 $F_z(h)$:

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

wherein $\delta > 0$ is a tolerance parameter for fairness violations. Exemplary $F_a(h)$ includes the Equal Opportunity measure $F_a(h) := \mathbb{P}(h(X) = +1|Y = +1, Z = a)$ [12], the demographic parity $F_a(h) := \mathbb{P}(h(X) = +1|Z = a)$ [6], among others.

We assume the error rates are group dependent [30]: $e_+(x) = e_-(x) = e_z, \; z \in \{a, b\}$. Again we assume $e_a, e_b < 0.5$. 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 (between $+1/-1$ classes) and across the group. We remove this additional layer of complication to 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 explain later and demonstrate with experiments. For a clear exposition of our core idea, we 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 generalization errors due to less informative labels, at the same time, it helps remove the training's dependency on the noise rates (e_+, e_-), and therefore improves robustness to possible mis-specification of them. We will quantify such trade-off.

We would mainly demonstrate our idea by comparing to methods that explicitly use the knowledge of noise rates. For one reason, these approaches often have strong theoretical guarantees. The other reason is that the benefit of our approach of not requiring the noise rate would be clearer for this comparison. Particularly, we will use loss correction [25] as the running example. We are aware of the other recent works that do not require the noise rates. First of all, these approaches focus on improving model accuracy, but not the fairness constraints. Secondly, empirically we observe that these approaches can also benefit from our noise rate balancing procedure. We left the discussions to our experiments, where we will demonstrate using one of such losses, peer loss [21].

Loss correction. Once knowing e_+, e_-, there exist different ways to improve training robustness in the presence of noisy labels. We consider demonstrating our idea using the classical loss correction framework [25, 27]: denote by $\tilde{e}_+, \tilde{e}_-$ the estimated version of e_+, e_-. In practice, this knowledge

can be obtained by using existing techniques [19, 26, 36]. For a given calibrated loss ℓ, the loss correction approach defines the following loss:

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

where in above sgn is the sign function. A favorable property of $\tilde{\ell}$ is its unbiasedness (expectation over $\tilde{Y}|Y = y$) w.r.t. $\ell(h(x), y)$ when given the correct noise rate parameters e_+, e_-. For readers who need more background knowledge of loss correction, we reproduce details in Appendix A.1. A classifier will be trained following empirical risk minimization (ERM) using $\tilde{\ell}$: $h^*_{\tilde{\ell},\tilde{D}} := \arg\min_{h\in\mathcal{H}} \sum_{n=1}^{N} \tilde{\ell}(h(x_n), \tilde{y}_n)$. Compared to the standard loss correction, we omit the $1 - \tilde{e}_- - \tilde{e}_+$ term from the denominator, which will not affect the minimizer of our ERM problem. Denote $h^*_{\ell,\mathcal{D}} = \arg\min_{h\in\mathcal{H}} R_{\ell,\mathcal{D}}(h)$, the optimal classifier for $R_{\ell,\mathcal{D}}(h)$, and $\texttt{err}_{\texttt{M}} := \max\{|\tilde{e}_+ - e_+|, |\tilde{e}_- - e_-|\}$ the maximal mis-specification error. When ℓ is L-Lipschitz and bounded by $\bar{\ell}$, let $L_1 := 4L, L_2 := 2\bar{\ell} > 0$, adapting the proof from [25] we show:

Theorem 1. *For any $\delta > 0$, we have with probability at least $1 - \delta$:*

$$R_{\ell,\mathcal{D}}(h^*_{\tilde{\ell},\tilde{D}}) - R_{\ell,\mathcal{D}}(h^*_{\ell,\mathcal{D}}) \leq \underbrace{\frac{1}{1 - e_+ - e_-}}_{\textit{label informativeness}} \cdot L_1 \cdot \mathcal{R}(\mathcal{H}) + \underbrace{\frac{\texttt{err}_{\texttt{M}}}{1 - e_+ - e_-}}_{\textit{mis-specification}} \cdot L_2 + 2\sqrt{\frac{\log 1/\delta}{2N}}, \tag{3}$$

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

Clearly, e_+, e_- control the error due to label informativeness: a pair of e_+, e_- with higher sum $(e_+ + e_-)$ induce larger generalization error bound. On the other hand, the potential imperfect estimates of them introduce additional error due to the mis-specification.

3.1 When does equalizing error rate improve 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 $e_+ > e_-$. 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 classes. Note that $e < 0.5$. Denote by $\hat{y}_n$ the generated label for example n, $\hat{\mathcal{D}}$ the distribution of $(X, \hat{Y})$, and $\hat{D}$ the dataset $\{(x_n, \hat{y}_n)\}_{n=1}^N$. We first show:

Lemma 2. *When $e < 0.5$, 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 $\hat{\mathcal{D}}$, and when the hypothesis space is sufficiently large, we know that the optimal classifier for $\mathbb{E}_{\hat{\mathcal{D}}}[\ell(h(X), \hat{Y})]$ will be exactly the same as the one for $\mathbb{E}_{\mathcal{D}}[\ell(h(X), Y)]$. This further implies that when the error rates are symmetric, the training might not need the error rates information, and performing ERM on the noisy data:

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

suffices to find the optimal classifier for the clean distribution, when given a sufficient amount of data. 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. *Suppose 1) ℓ is symmetric s.t. $\ell(h(x), -y) = \ell(-h(x), y)$ and 2) $-h^*_{\ell,\hat{D}}, -h^*_{\ell,\mathcal{D}} \in \mathcal{H}$: $\hat{R}_{\ell,\hat{D}}(-h^*_{\ell,\hat{D}}) \geq \hat{R}_{\ell,\hat{D}}(-h^*_{\ell,\mathcal{D}})$. Then for any $\delta > 0$, we have with probability at least $1 - \delta$: $R_{\ell,\mathcal{D}}(h^*_{\ell,\hat{D}}) - R_{\ell,\mathcal{D}}(h^*_{\ell,\mathcal{D}}) \leq \frac{1}{1-2e} \cdot L_1 \cdot \mathcal{R}(\mathcal{H}) + 2\sqrt{\frac{\log 1/\delta}{2N}}$.*

The second condition above is simply stating that the opposite classifier of the empirically optimal one incurs a high empirical loss, and particularly higher than $-h^*_{\ell,\mathcal{D}}$. Intuitively, this condition is satisfied for binary classification — if one classifier performs the best on the empirical data, flipping its prediction often results in a very wrong one. Now comparing Theorem 1 and 3 we observe that equalizing noise rates increases the error due to loss of informative labels 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 $\frac{\texttt{err}_{\texttt{M}}}{1-e_+-e_-} \cdot L_2$. Therefore, it is considered helpful to increase e_- to e_+ when $\texttt{err}_{\texttt{M}} \geq \frac{|e_+-e_-|\cdot L_1 \cdot \mathcal{R}(\mathcal{H})}{|1-2e_+|\cdot L_2}$. The above condition is likely to hold when the gap between noise rates $|e_+ - e_-|$ is sufficiently small, and when we do not have high confidence in the estimation of e_+, e_- (therefore a high $\texttt{err}_{\texttt{M}}$).

4 Equalizing error rates improves fairness guarantee

In this section, we consider the fairness constrained learning problem and show that equalizing noise rates helps improve fairness guarantee when only accessing the noisy labels. There are generally two types of fairness constraints: those that do not depend on the label and those that 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 does not impose additional challenges when enforcing such constraints. For those that do, typically they are functions of true positive rates (TPR, also known as the equal opportunity measure) and false positive rates (FPR). It was shown in [30] that equalizing TPR and FPR on the noisy label leads to disparities when $e_a \neq e_b$. We will reproduce the same observations. We focus on the second type of constraints: 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 recent works on performing fairness constraint correction [30]. 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

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

Resampling the noisy data examples such that $\mathbb{P}(\tilde{Y} = +1|Z = z) = \mathbb{P}(\tilde{Y} = -1|Z = z) = 0.5, z \in \{a, b\}$, define $C_{z,1} := 0.5 \cdot e_z, C_{z,2} := 0.5 \cdot (1 - e_z)$, we derive the following relationship:

Lemma 4. *$TPR_z(h)$, $FPR_z(h)$ relate to $\widetilde{TPR}_z(h)$, $\widetilde{FPR}_z(h)$ as follows:*

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

The above lemma first implies that 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. Denote by $\tilde{e}_a, \tilde{e}_b$ (both < 0.5) the estimated copies of e_a, e_b that we have access to. Suppose we suffer from the following mis-specifications: $\texttt{err}_{\texttt{M}} := \max\{|\tilde{e}_a - e_a|, |\tilde{e}_b - e_b|\}$. Denote the corrected TPR and FPR 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 $\tilde{C}_{z,1} := 0.5 \cdot \tilde{e}_z, \tilde{C}_{z,2} := 0.5 \cdot (1 - \tilde{e}_z)$, and:

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

Theorem 5 establishes possible fairness violation due to $\texttt{err}_{\texttt{M}}$, noise rates mis-specification:

Theorem 5. *Equalizing $TPR^c_z(h)$ & $FPR^c_z(h)$ for a, b leads to following possible fairness violation:*

$$|TPR_a(h) - TPR_b(h)| \leq \texttt{err}_{\texttt{M}} \cdot \left(\frac{\widetilde{TPR}_a(h)}{(2e_a - 1)(2\tilde{e}_a - 1)} + \frac{\widetilde{TPR}_b(h)}{(2e_b - 1)(2\tilde{e}_b - 1)} \right)$$
$$|FPR_a(h) - FPR_b(h)| \leq \texttt{err}_{\texttt{M}} \cdot \left(\frac{\widetilde{FPR}_a(h)}{(2e_a - 1)(2\tilde{e}_a - 1)} + \frac{\widetilde{FPR}_b(h)}{(2e_b - 1)(2\tilde{e}_b - 1)} \right)$$

But as a consequence of Lemma 4, we immediately know that

Theorem 6. *When $e_a = e_b$, equalizing $\widetilde{TPR}$ and $\widetilde{FPR}$ 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. Consider the case $e_a > e_b$. Now suppose we are able to 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, and correspondingly $\hat{y}_n$ the newly generated training labels. Then performing fairness constrained ERM directly on $\hat{D} := \{(x_n, \hat{y}_n, z_n)\}_n$:

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

will help us equalize TPR and FPR between two groups when the number of training data is sufficiently large. In above, $\hat{F}_a(h), \hat{F}_b(h)$ are the empirically computed fairness measures using $\hat{D}$.

5 Identifying cleaner class without using ground truth label

With the aforementioned benefits of balancing the noise rates, the remaining technical question is to determine how does e_+ or e_a compare to e_- or e_b? Of course, when we have access to ground truth labels, we will 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 for ground truth labels. We prove its sufficiency in identifying the noisier class of labels. We demonstrate our 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.

5.1 Noisy label agreements

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 [36] and has been shown to be a requirement that is mild to satisfy. Now we define the following two quantities that are central to the development of our idea. For an arbitrary instance X_1 with noisy label $\tilde{Y}_1$, denote the noisy labels for two nearest neighbor instances of X_1 as $\tilde{Y}_2, \tilde{Y}_3$. Define the following agreement measures:

Definition 8 (2-NN Agreements)**.** *Let $\tilde{Y}_1$ denote the noisy label for a randomly selected instance X_1. $\tilde{Y}_2, \tilde{Y}_3$ are the noisy labels of X_1's 2-NN instances (measured by $||x - x'||_2$).*

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

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

$\text{PA}_{\mathcal{D}}$ 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 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 the order between e_+, e_- ($sgn(e_+ - e_-)$):

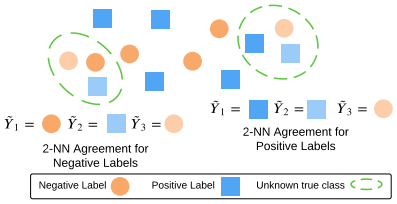

Figure 1: **2-NN Agreements.** "Transparent" ones are the solid instance's 2-NN instances.

Theorem 9. *When $e_+, e_- < 0.5$ and D satisfies 2-NN clusterability, $\textit{PA}_{\mathcal{D}^\diamond}, \textit{NA}_{\mathcal{D}^\diamond}$ relate to e_+, e_- as follows: $\textit{PA}_{\mathcal{D}^\diamond} - \textit{NA}_{\mathcal{D}^\diamond} = 2(0.5 - e_+)(0.5 - e_-)(e_- - e_+)$. Then if $\textit{PA}_{\mathcal{D}^\diamond} > \textit{NA}_{\mathcal{D}^\diamond}$, we know that $e_+ < e_-$; otherwise $e_+ > e_-$. If $\textit{PA}_{\mathcal{D}^\diamond} = \textit{NA}_{\mathcal{D}^\diamond}$, then $e_+ = e_-$.*

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 $\frac{\#\{n:\, \tilde{y}_n = \tilde{y}_{n_1} = \tilde{y}_{n_2} = +1\}}{\#\{n:\, \tilde{y}_n = +1\}}$. $\#(\cdot)$ counts the number of samples that satisfy the condition.

We show pseudocode for an implementation of estimating PA in Figure A1 in Appendix A.11.

5.2 Our solution: NOISE+

We show that randomly flipping $\tilde{Y} = y$ with the smaller $e_{sgn(y)}$ by a small probability 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 ones with $\tilde{Y} = -1$). We show numpy-like pseudocode for the flipping function in Figure 3, and the implementation for exclusively flipping $\tilde{Y} = -1$ labels is symmetric. Denote by $\hat{Y}$ as the flipped version of $\tilde{Y}$: $\mathbb{P}(\hat{Y} = -1 | \tilde{Y} = +1) = \epsilon$, and $\hat{e}_+ := \mathbb{P}(\hat{Y} = -1 | Y = +1), \hat{e}_- := \mathbb{P}(\hat{Y} = +1 | Y = -1)$. We have:

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

Since $1 - e_+ + e_- > 0$, when ϵ is small, the above derivation shows the effectiveness in reducing the noise rate gap $e_- - e_+$ 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 s.t. $\hat{e}_- - \hat{e}_+ = 0$. 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 $f(\epsilon) := 0.5 \cdot (\text{PA}_{\mathcal{D}^\diamond} - \text{NA}_{\mathcal{D}^\diamond})$. Easy to derive the three solutions for $f(\epsilon) = 0$ (setting each of the terms to 0): $\epsilon_1 = 1 - \frac{0.5}{e_-} < 0$, $\epsilon_2 = \frac{e_- - e_+}{1 - e_+ + e_-}$, $\epsilon_3 = \frac{0.5 - e_+}{1 - e_+}$: note that $\epsilon_2 = \frac{e_- - e_+}{1 - e_+ + e_-} < \frac{(e_- - e_+) + (1 - e_+ - e_-)}{(1 - e_+ + e_-) + (1 - e_+ - e_-)} = \frac{1 - 2e_+}{2(1 - e_+)} = \epsilon_3$, and ϵ_3 will lead to an uninformative state where $\hat{e}_+ = 0.5$. Therefore ϵ_2 is our target root.

The monotonicity of $f(\epsilon)$ from 0 to ϵ_2 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 the synthetic datasets induced by them are $\mathcal{D}_l$ and $\mathcal{D}_r$, and the gaps of the counted agreements are $C_l = \text{PA}_{\mathcal{D}_l} - \text{NA}_{\mathcal{D}_l}$ and $C_r = \text{PA}_{\mathcal{D}_r} - \text{NA}_{\mathcal{D}_r}$, 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 $-\gamma \le C_{\text{mid}} \le \gamma$, we return the labels flipped by ϵ_{mid}. Otherwise, we update the values of ϵ_l and ϵ_r according to the sign of C_{mid}: if $C_{\text{mid}} < 0$, we set $\epsilon_r \leftarrow \epsilon_{\text{mid}}, \mathcal{D}_r \leftarrow \mathcal{D}_{\text{mid}}$ (reducing ϵ_r); otherwise $\epsilon_l \leftarrow \epsilon_{\text{mid}}, \mathcal{D}_l \leftarrow \mathcal{D}_{\text{mid}}$ (increasing ϵ_l). 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. Note that Algorithm 1 assumes $e_+ < e_-$, and the implementation is symmetric for $e_+ > e_-$.

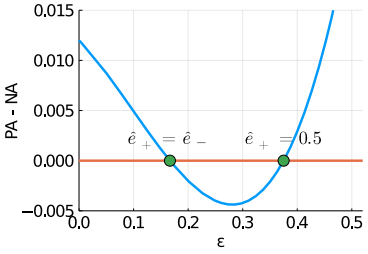

Figure 2: **Agreement gap PA−NA varies for different ϵ.** There are only two positive roots for PA − NA = 0. The less one results in $\hat{e}_+ = \hat{e}_-$.

Selection of loss function. Our algorithm is particularly suitable for losses that do not require the knowledge of noise rates. Standard cross entropy (CE) would certainly be applicable. Other robust loss functions are great options too. We note a recently proposed loss, peer loss function [21], does not require the specification of noise rates:

$$\ell_{\text{peer}}(h(x_n), \tilde{y}_n) := \ell\left(h(x_n), \tilde{y}_n\right) - \alpha \cdot \ell\left(h(x_{p_1}), \tilde{y}_{p_2}\right).$$

Here, $\alpha > 0$ is a hyper-parameter to control the balance of the instances for each label, p_1 and p_2 are uniformly and randomly selected peer samples. Peer loss is also promoted in [30] for the fairness constrained setting. We believe it is particularly suitable for our NOISE+. We will evaluate CE and peer loss in the experiment section.

Fairness constrained learning. Our above derivations and observations can be similarly applied to the fairness constrained setting. 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.9 without repeating them here.

Training. Once we have the output and noise-incremented 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 in Eqn. (7) directly too.

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 details to Appendix A.10. In the next section, we will demonstrate the effectiveness of our results with the CIFAR-10 dataset.

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

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 3: **Pseudocode for** Flip. Flip takes the dataset and a small probability ϵ as input, and only flips positive examples with probability ϵ.

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 [15] face attribute dataset, and CIFAR 10 [16] dataset. We defer more dataset details to Appendix B.1.

6.1 Unconstrained learning

Setup. For the 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 [25] with mis-specified noise rates (**Mis. SL**) and estimated noise rates (**Est. SL**) respectively, vanilla cross entropy (**CE**), and peer loss functions [21] (**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 [26]. We note that the training of peer loss functions does not 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 its much smaller data size. We repeat all the methods 5 runs with different random seeds and report the mean and standard deviation.

Table 1: **Binary classification accuracy of compared methods on 3 datasets across different levels of noise rates.** Mis. SL: surrogate loss [25] with misspecified parameters. Est. SL: surrogate loss [25] with estimated parameters. CE: vanilla cross entropy. Peer: peer loss function [21]. 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 blue.

			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
	0.3	0.4	52.65 ± 0.53	72.67 ± 0.26	71.55 ± 0.88	73.49 ± 0.18	72.54 ± 1.84	74.27 ± 0.20
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

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 across different levels of noise gap for multi-class classification.**

		LESS NOISE				MORE NOISE	
Dataset	noise gap	Mis. SL	Est. SL	CE	Peer	CE	Peer
MNIST	0.1	89.59	89.69	86.66	88.12	86.81	89.19
	0.2	88.10	88.61	84.53	87.21	85.97	89.12
	0.3	84.97	86.88	85.24	86.35	81.89	88.75
CIFAR-10	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: **Accuracy of compared methods on Clothing1M dataset.**

Method	Test Accuracy
CE	68.94%
Loss Correction [27]	69.84%
Co-Teaching [11]	70.15%
CE + NOISE+	**70.37%**

Multi-class extension. We test NOISE+ in the multi-class setting by balancing the noise rates class by class. We evaluate the compared methods on MNIST [18] and CIFAR-10 [16] with more sophisticated noise transition matrices. Considering that both datasets have 10 classes, we adopt the following procedure to generate the 10×10 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. An MLP model is trained from scratch on MNIST dataset, while a pre-trained vision transformer [8] is used to extract visual features on CIFAR-10 dataset. As shown in Table 2, we observe that when the noise gap is 0.1 and 0.2, CE and Peer with NOISE+ outperforms the pure CE and Peer, respectively. When the noise gap is 0.3, balancing cannot compensate for the performance drop due to increased noise.

Experiments on Clothing1M. Clothing1M is a large-scale dataset collected from online shopping websites, comprising of 1 million training images of clothes with realistic noisy and 10,000 test data

Table 4: **Constrained learning results with group-dependent label noise.** We only report the results for $e_a = 0.2$ and $e_b = 0.4$. LR: naïve logistic regression without noise correction. GPR: group-weighted peer loss [30]. Peer: peer loss [21]. 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

with clean labels. For a fair comparison, we adopted the same setting as described in [27] and trained a ResNet-50 [13] classifier. Table 3 compares our methods with some other baselines, including loss correction [27] and Co-Teaching [11]. Without a careful tuning of training parameters, vanilla CE with NOISE+reached 70.37% test accuracy. In comparison, standard CE achieves 68.94%, loss correction achieves 69.84%, and Co-Teaching is 70.15%.

6.2 Constrained learning

Setup. We conduct experiments with equal odds constraints [12] on Adult, Compas, and Fairface datasets. 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) [30] with and without our increasing-to-balancing program.[1] We note that GPL degrades to peer loss when noise rates are homogeneous across protected groups [30], 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 4 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 profitable on Compas dataset because of uninformative features (only 10-dimensional features), and more robust on Fairface dataset with rich image information.

7 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 several 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.

Our noise rate balancing solution is primarily a data pre-processing procedure and is compatible with most existing learning with noisy label solutions. Our experiment result using peer loss is one such example. Empirically we do observe that many other solutions may also enjoy the benefits of balancing the noise rates. We believe this data pre-processing technique has the potential to find applications in other learning tasks when balancing label noise or bias is desired.

[1] We did not repeat testing surrogate loss as it is reported in [30] that 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.

Acknowledgments

This work is partially supported by the National Science Foundation (NSF) under grant IIS-2007951 and the NSF FAI program in collaboration with Amazon under grant IIS-2040800. The authors would like to thank Jiaheng Wei and Zhaowei Zhu for their prompt discussion and help on the experimental setup for the Clothing1M dataset.

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] 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.

[5] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1789–1799. PMLR, 2020.

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

[7] Christos Dimitrakakis, Yang Liu, David C Parkes, and Goran Radanovic. Bayesian fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 509–516, 2019.

[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. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2325–2336. PMLR, 2020.

[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] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[15] 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.

[16] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[17] 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.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[19] 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.

[20] Yang Liu. Understanding instance-level label noise: Disparate impacts and treatments. In *International Conference on Machine Learning*, pages 6725–6735. PMLR, 2021.

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

[22] 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.

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

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] Jiaheng Wei and Yang Liu. When optimizing f-divergence is robust with label noise. In *International Conference on Learning Representations*, 2021.

[33] 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.

[34] 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.

[35] Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10113–10123, 2021.

[36] Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. *Proceedings of the 38th International Conference on Machine Learning*, 2021.

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

Yang Liu
Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
yangliu@ucsc.edu

Jialu Wang
Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
faldict@ucsc.edu

This Appendix is organized as follows:

- Section A includes omitted proofs for theoretical conclusions in the main paper, as well as the extension to fairness constrained setting (A.9) and multi-class classification (A.10).
- Section B presents more experimental details and results.

A Omitted Proofs

A.1 Proof for Theorem 1

Proof Let $\ell^\diamond$ denote the noise-corrected loss with respect to true noise parameters e_+, e_-:

$$\ell^\diamond(h(x_n), \tilde{y}_n) := (1 - e_{-sgn(\tilde{y}_n)}) \cdot \ell(h(x_n), \tilde{y}_n) - e_{sgn(\tilde{y}_n)} \cdot \ell(h(x_n), -\tilde{y}_n) \tag{A1}$$

It was established in [25] the unbiasedness of $\ell^\diamond$:

Lemma 11 (Unbiasedness of $\ell^\diamond$, [25]). $\frac{1}{1-e_+-e_-} \cdot \mathbb{E}_{\tilde{Y}|Y=y}[\ell^\diamond(h(x), \tilde{Y})] = \ell(h(x), y)$.

A direct consequence of this lemma, via repeatedly applying to each (X, Y), is its unbiasedness on the population level:

$$\frac{1}{1-e_+-e_-} \cdot \mathbb{E}_{\tilde{D}|D}[\hat{R}_{\ell^\diamond,\tilde{D}}(h)] = \hat{R}_{\ell,D}(h), \quad \frac{1}{1-e_+-e_-} \cdot R_{\ell^\diamond,\tilde{\mathcal{D}}}(h) = R_{\ell,\mathcal{D}}(h)$$

The following fact holds by subtracting $\ell^\diamond$ from $\tilde{\ell}$:

$$\tilde{\ell}(h(x_n), \tilde{y}_n) = \ell^\diamond(h(x_n), \tilde{y}_n) + (e_{-\tilde{y}_n} - \tilde{e}_{-\tilde{y}_n}) \cdot \ell(h(x_n), \tilde{y}_n) + (\tilde{e}_{-\tilde{y}_n} - e_{\tilde{y}_n}) \cdot \ell(h(x_n), -\tilde{y}_n)$$

Using the triangle inequality of $|\cdot|$ we establish that

$$|\tilde{\ell}(h(x_n), \tilde{y}_n) - \ell^\diamond(h(x_n), \tilde{y}_n)| \leq \max\{|\tilde{e}_+ - e_+|, |\tilde{e}_- - e_-|\} \cdot \bar{\ell}. \tag{A2}$$

This further helps us bound the differences in the empirical risks:

$$|\hat{R}_{\tilde{\ell},\tilde{D}}(h) - \hat{R}_{\ell^\diamond,\tilde{D}}(h)| \leq \max\{|\tilde{e}_+ - e_+|, |\tilde{e}_- - e_-|\} \cdot \bar{\ell} \tag{A3}$$

Therefore

$$\begin{aligned}
\hat{R}_{\ell^\diamond,\tilde{D}}(h^*_{\tilde{\ell},\tilde{D}}) &\leq \hat{R}_{\tilde{\ell},\tilde{D}}(h^*_{\tilde{\ell},\tilde{D}}) + \max\{|\tilde{e}_+ - e_+|, |\tilde{e}_- - e_-|\} \cdot \bar{\ell} \\
&\leq \hat{R}_{\tilde{\ell},\tilde{D}}(h^*_{\ell,\mathcal{D}}) + \max\{|\tilde{e}_+ - e_+|, |\tilde{e}_- - e_-|\} \cdot \bar{\ell} && \text{(Opt. of } h^*_{\ell,\tilde{D}}\text{)} \\
&\leq \hat{R}_{\ell^\diamond,\tilde{D}}(h^*_{\ell,\mathcal{D}}) + 2\max\{|\tilde{e}_+ - e_+|, |\tilde{e}_- - e_-|\} \cdot \bar{\ell} && \text{(A4)}
\end{aligned}$$

Calling the results in [25], [Rademacher bound for max risk deviation, Proof of Lemma 2 therein], we know that for any $\delta > 0$, with probability at least $1-\delta$:

$$\sup_{h\in\mathcal{H}} \frac{1}{1-e_+-e_-}\left|R_{\ell^\diamond,\tilde{\mathcal{D}}}(h) - \hat{R}_{\ell^\diamond,\tilde{D}}(h)\right| \leq \frac{2L}{1-e_+-e_-}\cdot\mathcal{R}(\mathcal{H}) + \sqrt{\frac{\log 1/\delta}{2N}} \tag{A5}$$

The above knowledge further helps us establish that

$$\begin{aligned}
& R_{\ell,\mathcal{D}}(h^*_{\tilde{\ell},\tilde{D}}) - R_{\ell,\mathcal{D}}(h^*_{\ell,\mathcal{D}}) \\
=\ & \frac{1}{1-e_+-e_-}(R_{\ell^\diamond,\tilde{\mathcal{D}}}(h^*_{\tilde{\ell},\tilde{D}}) - R_{\ell^\diamond,\tilde{\mathcal{D}}}(h^*_{\ell,\mathcal{D}})) && \text{(Unbiasedness of } \ell^\diamond \text{ on } \tilde{\mathcal{D}}) \\
=\ & \frac{1}{1-e_+-e_-}(R_{\ell^\diamond,\tilde{\mathcal{D}}}(h^*_{\tilde{\ell},\tilde{D}}) - \hat{R}_{\ell^\diamond,\tilde{D}}(h^*_{\tilde{\ell},\tilde{D}})) && \text{(Rademacher bound)} \\
& + \frac{1}{1-e_+-e_-}(\hat{R}_{\ell^\diamond,\tilde{D}}(h^*_{\tilde{\ell},\tilde{D}}) - \hat{R}_{\ell^\diamond,\tilde{D}}(h^*_{\ell,\mathcal{D}})) && \text{(Eqn. (A4))} \\
& + \frac{1}{1-e_+-e_-}(\hat{R}_{\ell^\diamond,\tilde{D}}(h^*_{\ell,\mathcal{D}}) - R_{\ell^\diamond,\tilde{\mathcal{D}}}(h^*_{\ell,\mathcal{D}})) && \text{(Rademacher bound)} \\
\leq\ & \frac{4L}{1-e_+-e_-}\cdot\mathcal{R}(\mathcal{H}) + 2\sqrt{\frac{\log 1/\delta}{2N}} + 2\frac{\max\{|\tilde{e}_+-e_+|, |\tilde{e}_--e_-|\}}{1-e_+-e_-}\cdot\bar{\ell} \\
\leq\ & \frac{4L}{1-e_+-e_-}\cdot\mathcal{R}(\mathcal{H}) + 2\sqrt{\frac{\log 1/\delta}{2N}} + 2\frac{\texttt{err}_{\texttt{M}}}{1-e_+-e_-}\cdot\bar{\ell}\,.
\end{aligned}$$

We complete the proof. ■

A.2 Proof for Lemma 2

Proof Expanding $\mathbb{P}(h(X)\neq\hat{Y})$ using the law of total probability we have

$$\begin{aligned}
\mathbb{P}(h(X)\neq\hat{Y}) =\ & \mathbb{P}(h(X)\neq\hat{Y}, \hat{Y}\neq Y) + \mathbb{P}(h(X)\neq\hat{Y}, \hat{Y}= Y) \\
=\ & \mathbb{P}(h(X)\neq\hat{Y}\mid\hat{Y}\neq Y)\cdot\mathbb{P}(\hat{Y}\neq Y) + \mathbb{P}(h(X)\neq\hat{Y}\mid\hat{Y}= Y)\cdot\mathbb{P}(\hat{Y}= Y).
\end{aligned}$$

In binary classification, $h(X)\neq\tilde{Y}, \tilde{Y}\neq Y$ implies $h(X)=Y$, such that

$$\mathbb{P}(h(X)\neq\tilde{Y}\mid\tilde{Y}\neq Y) = \mathbb{P}(h(X)= Y\mid\tilde{Y}\neq Y).$$

Due to the independence of $\tilde{Y}$ and X given Y,

$$\mathbb{P}(h(X)= Y\mid\tilde{Y}\neq Y) = \frac{\mathbb{P}(h(X)= Y, \tilde{Y}\neq Y)}{\mathbb{P}(\tilde{Y}\neq Y)} = \frac{\mathbb{P}(h(X)= Y)\mathbb{P}(\tilde{Y}\neq Y)}{\mathbb{P}(\tilde{Y}\neq Y)} = \mathbb{P}(h(X)= Y)$$

Similarly, we have

$$\mathbb{P}(h(X)\neq\hat{Y}\mid\hat{Y}= Y) = \mathbb{P}(h(X)\neq\tilde{Y}).$$

Combining all above we finished the proof when $e<0.5$ by having:

$$\begin{aligned}
\mathbb{P}(h(X)\neq\hat{Y}) =\ & \mathbb{P}(h(X)= Y)\cdot e + \mathbb{P}(h(X)\neq Y)\cdot(1-e) \\
=\ & (1-2e)\cdot\mathbb{P}(h(X)\neq Y) + e
\end{aligned}$$

■

A.3 Proof for Theorem 3

Proof Again let $\ell^\diamond$ denote the noise-corrected loss with respect to true noise parameters e_+, e_-:

$$\ell^\diamond(h(x_n), \tilde{y}_n) := (1-e_{-sgn(\tilde{y}_n)})\cdot\ell(h(x_n),\tilde{y}_n) - e_{sgn(\tilde{y}_n)}\cdot\ell(h(x_n),-\tilde{y}_n) \tag{A6}$$

First notice the following when ℓ is a symmetric loss:

$$\begin{aligned}
& R_{\ell,\mathcal{D}}(h^*_{\ell,\hat{D}}) \\
=\ & \frac{1}{1-2e}\cdot R_{\ell^\diamond,\hat{\mathcal{D}}}(h^*_{\ell,\hat{D}}) && \text{(Unbiasedness of } \ell^\diamond \text{ on } \hat{\mathcal{D}} \text{ using symmetric } e) \\
=\ & \frac{1-e}{1-2e}\cdot R_{\ell,\hat{\mathcal{D}}}(h^*_{\ell,\hat{D}}) - \frac{e}{1-2e}\cdot R_{\ell,\hat{\mathcal{D}}}(-h^*_{\ell,\hat{D}}) && \text{(A7)}
\end{aligned}$$

The last equality uses the definition of $\ell^\diamond$, and is due to ℓ being symmetric. Then we show that

$$\begin{aligned}
&\frac{1}{1-2e}\cdot\left(R_{\ell,\hat{\mathcal{D}}}(h^*_{\ell,\hat{D}})-R_{\ell,\hat{\mathcal{D}}}(h^*_{\ell,\mathcal{D}})\right)\\
=\;&\frac{1}{1-2e}\left(R_{\ell,\hat{\mathcal{D}}}(h^*_{\ell,\hat{D}})-\hat{R}_{\ell,\hat{D}}(h^*_{\ell,\hat{D}})\right) && \text{(Rademacher bound)}\\
&+\frac{1}{1-2e}\left(\hat{R}_{\ell,\hat{D}}(h^*_{\ell,\hat{D}})-\hat{R}_{\ell,\hat{D}}(h^*_{\ell,\mathcal{D}})\right) && (\leq 0 \text{ Optimality of } h^*_{\ell,\hat{D}} \text{ on } \ell,\hat{D})\\
&+\frac{1}{1-2e}\left(\hat{R}_{\ell,\hat{D}}(h^*_{\ell,\mathcal{D}})-R_{\ell,\hat{\mathcal{D}}}(h^*_{\ell,\mathcal{D}})\right) && \text{(Rademacher bound)}\\
\leq\;&\frac{4L}{1-2e}\mathcal{R}(\mathcal{H})+2\sqrt{\frac{\log 1/\delta}{2N}}
\end{aligned}$$

The inequality is due to the Rademacher bound we invoked as in Eqn. (A5) as well as the optimality of $h^*_{\ell,\hat{D}}$ on $\ell,\hat{D}$. That is we have proved with probability at least $1-\delta$ that

$$\frac{1}{1-2e}\cdot R_{\ell,\hat{\mathcal{D}}}(h^*_{\ell,\hat{D}})\leq\frac{1}{1-2e}\cdot R_{\ell,\hat{\mathcal{D}}}(h^*_{\ell,\mathcal{D}})+\frac{4L}{1-2e}\mathcal{R}(\mathcal{H})+2\sqrt{\frac{\log 1/\delta}{2N}} \tag{A8}$$

Repeating the same analysis and using the assumed condition that $\hat{R}_{\ell,\hat{D}}(-h^*_{\ell,\hat{D}})-\hat{R}_{\ell,\hat{D}}(-h^*_{\ell,\mathcal{D}})\geq 0$ we have

$$\frac{1}{1-2e}\cdot R_{\ell,\hat{\mathcal{D}}}(-h^*_{\ell,\hat{D}})\geq\frac{1}{1-2e}\cdot R_{\ell,\hat{\mathcal{D}}}(-h^*_{\ell,\mathcal{D}})-\frac{4L}{1-2e}\mathcal{R}(\mathcal{H})-2\sqrt{\frac{\log 1/\delta}{2N}} \tag{A9}$$

Combining above (Eqn. (A8) and (A9)),we have with probability at least $1-\delta$ (that both of the above bounds will happen simultaneously)

$$\begin{aligned}
R_{\ell,\mathcal{D}}(h^*_{\ell,\hat{D}})=\;&\frac{1-e}{1-2e}\cdot R_{\ell,\hat{\mathcal{D}}}(h^*_{\ell,\hat{D}})-\frac{e}{1-2e}\cdot R_{\ell,\hat{\mathcal{D}}}(-h^*_{\ell,\hat{D}})\\
\leq\;&\frac{1-e}{1-2e}\cdot R_{\ell,\hat{\mathcal{D}}}(h^*_{\ell,\mathcal{D}})-\frac{e}{1-2e}\cdot R_{\ell,\hat{\mathcal{D}}}(-h^*_{\ell,\mathcal{D}})+\frac{4L}{1-2e}\mathcal{R}(\mathcal{H})+2\sqrt{\frac{\log 1/\delta}{2N}}\\
=\;&R_{\ell,\mathcal{D}}(h^*_{\ell,\mathcal{D}})+\frac{4L}{1-2e}\mathcal{R}(\mathcal{H})+2\sqrt{\frac{\log 1/\delta}{2N}}.
\end{aligned}$$

The inequality uses Eqn. (A8) and (A9). Again the last equality is reusing Eqn. (A7). This completes the proof. ■

A.4 Proof for Lemma 4

Proof

$$\mathbb{P}(h(X)=+1|\tilde{Y}=+1,Z=a)=\frac{\mathbb{P}(h(X)=+1,\tilde{Y}=+1|Z=a)}{\mathbb{P}(\tilde{Y}=+1|Z=a)} \tag{A10}$$

Again we do the trick of sampling $\mathbb{P}(\tilde{Y}=+1|Z=a)$ to be 0.5, which allows us to focus on the numerator.

$$\begin{aligned}
&\mathbb{P}(h(X)=+1,\tilde{Y}=+1|Z=a)\\
=\;&\mathbb{P}(h(X)=+1,\tilde{Y}=+1,Y=+1|Z=a)\\
&+\mathbb{P}(h(X)=+1,\tilde{Y}=+1,Y=-1|Z=a)\\
=\;&\mathbb{P}(h(X)=+1,\tilde{Y}=+1|Y=+1,Z=a)\cdot\mathbb{P}(Y=+1|Z=a)\\
&+\mathbb{P}(h(X)=+1,\tilde{Y}=+1|Y=-1,Z=a)\cdot\mathbb{P}(Y=-1|Z=a)\\
=\;&\mathbb{P}(h(X)=+1|Y=+1,Z=a)\cdot(1-e_a)\cdot\mathbb{P}(Y=+1|Z=a)\\
&+\mathbb{P}(h(X)=+1|Y=-1,Z=a)\cdot e_a\cdot\mathbb{P}(Y=-1|Z=a)\\
&\qquad\qquad\text{(Independence of } X \text{ and } \tilde{Y} \text{ given } Y)
\end{aligned}$$

That is

$$0.5 \cdot \widetilde{\text{TPR}}_a(h) = \text{TPR}_a(h) \cdot (1 - e_a) \cdot \mathbb{P}(Y = +1|Z = a) + \text{FPR}_a(h) \cdot e_a \cdot \mathbb{P}(Y = -1|Z = a) \tag{A11}$$

Similarly for FPR we have

$$\mathbb{P}(h(X) = +1|\tilde{Y} = -1, Z = a) = \frac{\mathbb{P}(h(X) = +1, \tilde{Y} = -1|Z = a)}{\mathbb{P}(\tilde{Y} = -1|Z = a)} \tag{A12}$$

Following similar steps as above, the numerator further derives as

$$\begin{aligned} & \mathbb{P}(h(X) = +1, \tilde{Y} = +1|Z = a) \\ = \ & \mathbb{P}(h(X) = +1|Y = -1, Z = a) \cdot (1 - e_a) \cdot \mathbb{P}(Y = +1|Z = a) \\ & + \mathbb{P}(h(X) = +1|Y = +1, Z = a) \cdot e_a \cdot \mathbb{P}(Y = -1|Z = a) \end{aligned}$$

That is

$$0.5 \cdot \widetilde{\text{FPR}}_a(h) = \text{FPR}_a(h) \cdot (1 - e_a) \cdot \mathbb{P}(Y = +1|Z = a) + \text{TPR}_a(h) \cdot e_a \cdot \mathbb{P}(Y = -1|Z = a) \tag{A13}$$

When $\mathbb{P}(\tilde{Y} = +1|Z = a) = \mathbb{P}(\tilde{Y} = +1|Z = b) = 0.5$, we will also have

$$0.5 = \mathbb{P}(\tilde{Y} = +1|Z = a) = \mathbb{P}(Y = +1|Z = a)(1 - e_a) + \mathbb{P}(Y = -1|Z = a)e_a \tag{A14}$$

which returns us that $\mathbb{P}(Y = +1|Z = a) = \frac{0.5 - e_a}{1 - 2e_a} := p = 0.5$. Using this knowledge and solving the linear equations defined by Eqn. (A17) and (A13) we have

$$\text{TPR}_a(h) = \frac{C_{a,1} \cdot \widetilde{\text{TPR}}_a(h) - C_{a,2} \cdot \widetilde{\text{FPR}}_a(h)}{e_a - 0.5} \tag{A15}$$

$$\text{FPR}_a(h) = \frac{C_{a,1} \cdot \widetilde{\text{FPR}}_a(h) - C_{a,2} \cdot \widetilde{\text{TPR}}_a(h)}{e_a - 0.5} \tag{A16}$$

■

A.5 Proof for Theorem 5

Proof Combining Eqn. (5) and (6) we have

$$\begin{aligned} & |\text{TPR}_z(h) - \text{TPR}_z^c(h)| \\ = & \left| \frac{0.5 \cdot e_z \cdot \widetilde{\text{TPR}}_z(h) - 0.5(1 - e_z) \cdot \widetilde{\text{FPR}}_z(h)}{e_z - 0.5} - \frac{0.5 \cdot \widetilde{e}_z \cdot \widetilde{\text{TPR}}_z(h) - 0.5(1 - \widetilde{e}_z) \cdot \widetilde{\text{FPR}}_z(h)}{\tilde{e}_z - 0.5} \right| \\ = & \frac{|\tilde{e}_z - e_z| \cdot \widetilde{\text{TPR}}_z(h)}{(2e_z - 1)(2\tilde{e}_z - 1)} \\ \leq & \frac{\texttt{err}_{\texttt{M}} \cdot \widetilde{\text{TPR}}_z(h)}{(2e_z - 1)(2\tilde{e}_z - 1)} \ . \end{aligned} \tag{A17}$$

Then equalizing TPR that $\text{TPR}_a^c(h) = \text{TPR}_b^c(h)$ returns us

$$\begin{aligned} & |\text{TPR}_a(h) - \text{TPR}_b(h)| \\ = & |\text{TPR}_a(h) - \text{TPR}_a^c(h) + \text{TPR}_b^c(h) - \text{TPR}_b(h)| \\ \leq & |\text{TPR}_a(h) - \text{TPR}_a^c(h)| + |\text{TPR}_b^c(h) - \text{TPR}_b(h)| \\ \leq & \frac{\texttt{err}_{\texttt{M}} \cdot \widetilde{\text{TPR}}_a(h)}{(2e_a - 1)(2\tilde{e}_a - 1)} + \frac{\texttt{err}_{\texttt{M}} \cdot \widetilde{\text{TPR}}_b(h)}{(2e_b - 1)(2\tilde{e}_b - 1)}, \end{aligned}$$

where the last inequality is an application of Eqn. (A17).

Similarly

$$
\begin{aligned}
&|\text{FPR}_z(h) - \text{FPR}^c_z(h)| \\
=& \left| \frac{0.5 \cdot e_z \cdot \widetilde{\text{FPR}}_z(h) - 0.5(1-e_z) \cdot \widetilde{\text{TPR}}_z(h)}{e_z - 0.5} - \frac{0.5 \cdot \tilde{e}_z \cdot \widetilde{\text{FPR}}_z(h) - 0.5(1-\tilde{e}_z) \cdot \widetilde{\text{TPR}}_z(h)}{\tilde{e}_z - 0.5} \right| \\
=& \frac{|\tilde{e}_z - e_z| \cdot \widetilde{\text{FPR}}_z(h)}{(2e_z - 1)(2\tilde{e}_z - 1)} \\
\leq& \frac{\texttt{err}_{\texttt{M}} \cdot \widetilde{\text{FPR}}_z(h)}{(2e_z - 1)(2\tilde{e}_z - 1)}.
\end{aligned}
$$

Then equalizing FPR that $\text{FPR}^c_a(h) = \text{FPR}^c_b(h)$ we have

$$
\begin{aligned}
&|\text{FPR}_a(h) - \text{FPR}_b(h)| \\
=&|\text{FPR}_a(h) - \text{FPR}^c_a(h) + \text{FPR}^c_b(h) - \text{FPR}_b(h)| \\
\leq&|\text{FPR}_a(h) - \text{FPR}^c_a(h)| + |\text{FPR}^c_b(h) - \text{FPR}_b(h)| \\
\leq&\frac{\texttt{err}_{\texttt{M}} \cdot \widetilde{\text{FPR}}_a(h)}{(2e_a - 1)(2\tilde{e}_a - 1)} + \frac{\texttt{err}_{\texttt{M}} \cdot \widetilde{\text{FPR}}_b(h)}{(2e_b - 1)(2\tilde{e}_b - 1)}.
\end{aligned}
$$

■

A.6 Proof for Theorem 6

Proof Easy to show that when $e_a = e_b$, $C_{a,1} = C_{b,1}$ and $C_{a,2} = C_{b,2}$. Therefore, from Eqn. (5) we know equalizing

$$
\widetilde{\text{TPR}}_a(h) = \widetilde{\text{TPR}}_b(h), \;\; \widetilde{\text{FPR}}_a(h) = \widetilde{\text{FPR}}_b(h) \tag{A18}
$$

will also return us

$$
\text{TPR}_a(h) = \text{TPR}_b(h), \;\; \text{FPR}_a(h) = \text{FPR}_b(h) \tag{A19}
$$

■

A.7 Proof for Theorem 9

Proof We start with deriving $\text{PA}_{\mathcal{D}^\diamond}$.

$$
\text{PA}_{\mathcal{D}^\diamond} = \mathbb{P}(\tilde{Y}_2 = \tilde{Y}_3 = +1 | \tilde{Y}_1 = +1) = \frac{\mathbb{P}(\tilde{Y}_1 = \tilde{Y}_2 = \tilde{Y}_3 = +1)}{\mathbb{P}(\tilde{Y}_1 = +1)}
$$

Due to the sampling step, we have $\mathbb{P}(\tilde{Y}_1 = +1) = 0.5$ - this allows us to focus on the denominator:

$$
\begin{aligned}
\mathbb{P}(\tilde{Y}_1 = \tilde{Y}_2 = \tilde{Y}_3 = +1) &\overset{(1)}{=} \mathbb{P}(Y = +1) \prod_{i=1}^{3} \mathbb{P}(\tilde{Y}_i = +1 | Y = +1) + \mathbb{P}(Y = -1) \prod_{i=1}^{3} \mathbb{P}(\tilde{Y}_i = +1 | Y = -1) \\
&\overset{(2)}{=} \mathbb{P}(Y = +1) \cdot (1 - e_+)^3 + \mathbb{P}(Y = -1) \cdot e_-^3
\end{aligned}
$$

where in above, (1) uses the 2-NN clusterability of D, and (2) uses the conditional independence between the noisy labels. Similarly for $\text{NA}_{\mathcal{D}^\diamond}$ we have:

$$
\mathbb{P}(\tilde{Y}_2 = \tilde{Y}_3 = -1 | \tilde{Y}_1 = -1) = \frac{\mathbb{P}(\tilde{Y}_1 = \tilde{Y}_2 = \tilde{Y}_3 = -1)}{\mathbb{P}(\tilde{Y}_1 = -1)}
$$

Again we have that $\mathbb{P}(\tilde{Y}_1 = -1) = 0.5$, and the numerator derives as

$$
\begin{aligned}
\mathbb{P}(\tilde{Y}_1 = \tilde{Y}_2 = \tilde{Y}_3 = -1) &= \mathbb{P}(Y = +1) \prod_{i=1}^{3} \mathbb{P}(\tilde{Y}_i = -1 | Y = +1) + \mathbb{P}(Y = -1) \prod_{i=1}^{3} \mathbb{P}(\tilde{Y}_i = -1 | Y = -1) \\
&= \mathbb{P}(Y = +1) \cdot e_+^3 + \mathbb{P}(Y = -1) \cdot (1 - e_-)^3
\end{aligned}
$$

Taking the difference (and normalize by 0.5) we have

$$\begin{aligned} & 0.5 \cdot (\text{PA}_{\mathcal{D}^\diamond} - \text{NA}_{\mathcal{D}^\diamond}) \\ = \; & \mathbb{P}(\tilde{Y}_2 = \tilde{Y}_3 = +1 | \tilde{Y}_1 = +1) - \mathbb{P}(\tilde{Y}_2 = \tilde{Y}_3 = -1 | \tilde{Y}_1 = -1) \\ = \; & \mathbb{P}(Y = +1) \left((1 - e_+)^3 - e_+^3 \right) + \mathbb{P}(Y = -1) \left(e_-^3 - (1 - e_-)^3 \right) \end{aligned} \tag{A20}$$

Notice two facts: first we can derive that

$$(1 - e_+)^3 - e_+^3 = (1 - 2e_+)(e_+^2 - e_+ + 1), \;\; e_-^3 - (1 - e_-)^3 = -(1 - 2e_-)(e_-^2 - e_- + 1)$$

Second, we will use the following fact:

$$0.5 = \mathbb{P}(\tilde{Y} = +1) = \mathbb{P}(Y = +1)(1 - e_+) + \mathbb{P}(Y = -1) e_- \tag{A21}$$

from which we solve that $\mathbb{P}(Y = +1) = \frac{0.5 - e_-}{1 - e_+ - e_-}$. Symmetrically, $\mathbb{P}(Y = -1) = \frac{0.5 - e_+}{1 - e_+ - e_-}$.

Return the above two facts back into Eqn. (A20), we have

$$\begin{aligned} & \mathbb{P}(Y = +1)((1 - e_+)^3 - e_+^3) + \mathbb{P}(Y = -1)(e_-^3 - (1 - e_-)^3) \\ = \; & 2 \cdot \frac{(0.5 - e_+)(0.5 - e_-)}{1 - e_+ - e_-} \left((e_+^2 - e_+ + 1) - (e_-^2 - e_- + 1) \right) \\ = \; & 2 \cdot (0.5 - e_+) \cdot (0.5 - e_-) \cdot (e_- - e_+) \end{aligned}$$

completing the proof when $e_+, e_- < 0.5$. ■

A.8 Proof for Proposition 10

Proof Expanding $\mathbb{P}(\hat{Y} = -1 | Y = +1)$ using the law of total probability we have

$$\begin{aligned} \hat{e}_+ &= \mathbb{P}(\hat{Y} = -1 | Y = +1) \\ &= \mathbb{P}(\hat{Y} = -1, \tilde{Y} = +1 | Y = +1) + \mathbb{P}(\hat{Y} = -1, \tilde{Y} = -1 | Y = +1) \\ &= \mathbb{P}(\hat{Y} = -1 | \tilde{Y} = +1, Y = +1) \cdot \mathbb{P}(\tilde{Y} = +1 | Y = +1) \\ &\quad + \mathbb{P}(\hat{Y} = -1 | \tilde{Y} = -1, Y = +1) \cdot \mathbb{P}(\tilde{Y} = -1 | Y = +1) \\ &= \epsilon \cdot (1 - e_+) + 1 \cdot e_+ \qquad \text{(Independence between } \hat{Y} \text{ and } Y \text{ given } \tilde{Y}) \\ &= (1 - e_+) \cdot \epsilon + e_+ \end{aligned}$$

Similarly,

$$\begin{aligned} \hat{e}_- &= \mathbb{P}(\hat{Y} = +1 | Y = -1) \\ &= \mathbb{P}(\hat{Y} = +1, \tilde{Y} = +1 | Y = -1) + \mathbb{P}(\hat{Y} = +1, \tilde{Y} = -1 | Y = -1) \\ &= \mathbb{P}(\hat{Y} = +1 | \tilde{Y} = +1, Y = -1) \cdot \mathbb{P}(\tilde{Y} = +1 | Y = -1) \\ &\quad + \mathbb{P}(\hat{Y} = +1 | \tilde{Y} = -1, Y = -1) \cdot \mathbb{P}(\tilde{Y} = -1 | Y = -1) \\ &= (1 - \epsilon) \cdot e_-. \end{aligned}$$

The last equality is again due to the independence between $\hat{Y}$ and Y given $\tilde{Y}$, as well as the fact that we do not flip the $\tilde{Y} = -1$ labels so $\mathbb{P}(\hat{Y} = +1 | \tilde{Y} = -1, Y = -1) = 0$. Taking the difference we finish the proof.

■

A.9 Balancing noise for fairness constrained case

Define

$$\text{PA}_{\mathcal{D}^\diamond, a} = \mathbb{P}_{Z=a}(\tilde{Y}_2 = \tilde{Y}_3 = +1 | \tilde{Y}_1 = +1) \tag{A22}$$

$$\text{PA}_{\mathcal{D}^\diamond, b} = \mathbb{P}_{Z=b}(\tilde{Y}_2 = \tilde{Y}_3 = +1 | \tilde{Y}_1 = +1) \tag{A23}$$

We now claim that $sgn(\text{PA}_{\mathcal{D}^\diamond,a} - \text{PA}_{\mathcal{D}^\diamond,b}) = -sgn(e_a - e_b)$. We start with deriving $\text{PA}_{\mathcal{D}^\diamond,a}$.

$$\text{PA}_{\mathcal{D}^\diamond,a} = \mathbb{P}_{Z=a}(\tilde{Y}_2 = \tilde{Y}_3 = +1 | \tilde{Y}_1 = +1) = \frac{\mathbb{P}_{Z=a}(\tilde{Y}_1 = \tilde{Y}_2 = \tilde{Y}_3 = +1)}{\mathbb{P}_{Z=a}(\tilde{Y}_1 = +1)}$$

Due to the sampling step, we have $\mathbb{P}_{Z=a}(\tilde{Y}_1 = +1) = 0.5$ - this allows us to focus on the denominator:

$$\begin{aligned}
\mathbb{P}_{Z=a}(\tilde{Y}_1 = \tilde{Y}_2 = \tilde{Y}_3 = +1) &\overset{(1)}{=} \mathbb{P}_{Z=a}(Y = +1) \prod_{i=1}^{3} \mathbb{P}_{Z=a}(\tilde{Y}_i = +1 | Y = +1) \\
&\quad + \mathbb{P}_{Z=a}(Y = -1) \prod_{i=1}^{3} \mathbb{P}_{Z=a}(\tilde{Y}_i = +1 | Y = -1) \\
&\overset{(2)}{=} \mathbb{P}_{Z=a}(Y = +1) \cdot (1 - e_a)^3 + \mathbb{P}_{Z=a}(Y = -1) \cdot e_a^3
\end{aligned}$$

where in above, (1) uses the 2-NN clusterability of D, and (2) uses the conditional independence between the noisy labels. Similarly for $\text{PA}_{\mathcal{D}^\diamond,b}$ we have:

$$\text{PA}_{\mathcal{D}^\diamond,b} = \frac{\mathbb{P}_{Z=b}(Y = +1) \cdot (1 - e_b)^3 + \mathbb{P}_{Z=b}(Y = -1) \cdot e_b^3}{0.5} \tag{A24}$$

Firstly, we will use the following fact for $z \in \{a, b\}$:

$$\begin{aligned}
0.5 =& \mathbb{P}_{Z=z}(\tilde{Y} = +1) \\
=& \mathbb{P}_{Z=z}(\tilde{Y} = +1 | Y = +1) \cdot \mathbb{P}_{Z=z}(Y = +1) + \mathbb{P}_{Z=z}(\tilde{Y} = +1 | Y = -1) \cdot \mathbb{P}_{Z=z}(Y = -1) \\
=& \mathbb{P}_{Z=z}(Y = +1) \cdot (1 - e_z) + \mathbb{P}_{Z=z}(Y = -1) \cdot e_z
\end{aligned}$$

from which we solve that $\mathbb{P}_{Z=z}(Y = +1) = \frac{0.5 - e_z}{1 - 2e_z} = 0.5$. Therefore

$$\begin{aligned}
\text{PA}_{\mathcal{D}^\diamond,a} - \text{PA}_{\mathcal{D}^\diamond,b} &= (1 - e_a)^3 - (1 - e_b)^3 + e_a^3 - e_b^3 \\
&= (e_b - e_a) \left((1 - e_a)^2 + (1 - e_b)^2 + (1 - e_a)(1 - e_b) - e_a^2 - e_b^2 - e_a e_b \right) \\
&= (e_b - e_a) \left(1 - 2e_a + 1 - 2e_b + 1 - e_a - e_b \right)
\end{aligned} \tag{A25}$$

Note that $1 - 2e_a + 1 - 2e_b + 1 - e_a - e_b > 0$ when $e_a, e_b < 0.5$. This implies that we can use the 2-NN positive agreements $\text{PA}_{\mathcal{D}^\diamond,a} - \text{PA}_{\mathcal{D}^\diamond,b}$ across groups to compare e_a with e_b.

A.10 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. Consider a multi-class classification problem with K label classes, and the noise rates follow a uniform diagonal model:

$$\mathbb{P}(\tilde{Y} = k | Y = k) = 1 - e_k, \;\; \mathbb{P}(\tilde{Y} = k' | Y = k) = \frac{e_k}{K - 1}, \; \forall k' \neq k. \tag{A26}$$

Define $\text{KA}_{\mathcal{D}^\diamond,k} := \mathbb{P}(\tilde{Y}_2 = \tilde{Y}_3 = k | \tilde{Y}_1 = k), \; k = 1, 2, ..., K$. Similarly we can show that for any pair of k_1, k_2: $sgn(\text{KA}_{\mathcal{D}^\diamond,k_1} - \text{KA}_{\mathcal{D}^\diamond,k_2}) = -sgn(e_{k_1} - e_{k_2})$, wherein above e_{k_1}, e_{k_2} are the error rates of label class k_1, k_2. With the above, we can compute $\text{KA}_{\mathcal{D}^\diamond,k}$, rank them, and start inserting noise to the classes that are determined to have a lower error rate to match the highest one.

A.11 Pseudocodes

B Additional Experiment Details and Results

We provide more details on the experimental setup as well as further results.

```
import numpy as np
from sklearn.neighbors import NearestNeighbors
def estimate_PA(X, y):
    nbrs = NearestNei.(n_neighbors=3, algorithm='ball_tree').fit(X)
    _, indices = nbrs.kneighbors(X)
    return np.mean(np.array([np.all(y[i] == y[indices[i]]) for i in np
        .where(y > 0)[0]]))
```

Figure A1: **Numpy-like pseudocode for an implementation of estimating PA.** Our implementation utilizes scikit-learn's Nearest Neighbors module. The code for esimating NA is similar.

B.1 Datasets

We evaluate our methods on five datasets:

- Adult, the UCI Adult Income dataset [9]. The task is to predict whether an individual's income exceeds 50K. The dataset consists of 48,842 examples and 28 features. We select female and male as two protected groups in constrained learning. We resample the dataset to ensure that both the classes and groups are balanced.
- Compas, the COMPAS recidivism dataset for crime statistics with 7,168 instances and 10 features [2]. We select race as the protected attribute in constrained learning.
- Fairface, the face attribute dataset containing 108,501 images with balanced race and gender groups [15]. We use a pre-trained vision transformer (ViT/B-32) model [8] to extract image representations, and project them into 50-dimensional feature vectors. For both unconstrained and constrained learning, we take gender attribute as labels for binary classification. For constrained learning, we categorize race into White and Non-White groups.
- MNIST [18], consisting of 50,000 training images and 10,000 test images in 10 classes. We train a MLP model from scratch on the MNIST dataset.
- CIFAR-10 [16], consisting of 50,000 training images and 10,000 test images in 10 classes. We evaluate unconstrained multi-class classification on CIFAR-10 dataset. Similar to Fairface, we use a pre-trained vision transformer to extract 512-dimensional feature vectors.

For Adult, Compas, and German datasets, we perform random train/test splits in a ratio of 80 to 20. For Fairface, MNIST, and CIFAR-10, we follow their original splits.

B.2 Computing infrastructure

For all the experiments, we use a GPU cluster with 4 2080 Ti GPUs for training and evaluation.

B.3 Noise transition matrix for CIFAR-10

We adopt the following procedure to generate the noise transition matrix:

1. Manually set the diagonal elements at least 0.4. We ensure that the difference between the maximal elements and 0.4 is equal to the noise gap.
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 show one sample noise transition matrix generated by our procedure with noise gap 0.2 as follows:

$$\begin{bmatrix}
\mathbf{0.4} & 0.087 & 0.013 & 0.032 & 0.032 & 0.068 & 0.050 & 0.178 & 0.001 & 0.118 \\
0.043 & \mathbf{0.4} & 0.002 & 0.016 & 0.049 & 0.113 & 0.060 & 0.024 & 0.224 & 0.017 \\
0.181 & 0.111 & \mathbf{0.4} & 0.147 & 0.033 & 0.005 & 0.026 & 0.040 & 0.110 & 0.076 \\
0.051 & 0.001 & 0.060 & \mathbf{0.6} & 0.032 & 0.047 & 0.149 & 0.145 & 0.022 & 0.059 \\
0.001 & 0.167 & 0.119 & 0.032 & \mathbf{0.6} & 0.092 & 0.051 & 0.018 & 0.037 & 0.129 \\
0.097 & 0.007 & 0.001 & 0.059 & 0.016 & \mathbf{0.4} & 0.019 & 0.014 & 0.084 & 0.001 \\
0.018 & 0.023 & 0.277 & 0.041 & 0.034 & 0.014 & \mathbf{0.4} & 0.028 & 0.041 & 0.062 \\
0.149 & 0.096 & 0.081 & 0.019 & 0.041 & 0.015 & 0.143 & \mathbf{0.4} & 0.061 & 0.110 \\
0.031 & 0.066 & 0.022 & 0.007 & 0.133 & 0.080 & 0.049 & 0.113 & \mathbf{0.4} & 0.025 \\
0.029 & 0.040 & 0.023 & 0.043 & 0.027 & 0.162 & 0.048 & 0.036 & 0.018 & \mathbf{0.4}
\end{bmatrix}$$

B.4 Additional results

Table B1: **Binary classification accuracy of compared methods on 3 datasets across different levels of noise rates.** Mis. SL: surrogate loss [25] with misspecified parameters. Est. SL: surrogate loss [25] with estimated parameters. CE: vanilla cross entropy. Peer: peer loss function [21]. 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.

			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
$d = 28$	0.0	0.3	67.93 ± 0.52	71.58 ± 0.28	66.86 ± 0.47	71.33 ± 0.30	73.30 ± 0.27	73.74 ± 0.15
	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
	0.1	0.4	54.87 ± 0.85	71.48 ± 0.50	63.60 ± 1.04	71.44 ± 0.72	72.43 ± 1.90	73.56 ± 0.89
	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
$d = 10$	0.0	0.3	58.06 ± 0.32	62.69 ± 1.20	53.04 ± 3.69	66.41 ± 1.19	59.02 ± 7.78	65.93 ± 0.56
	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
	0.1	0.4	51.60 ± 0.12	63.34 ± 1.12	57.65 ± 3.90	66.47 ± 1.34	55.83 ± 6.43	67.04 ± 0.75
	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
$d = 50$	0.0	0.3	81.51 ± 0.09	83.36 ± 0.04	79.62 ± 0.12	81.37 ± 0.35	87.44 ± 0.15	88.25 ± 0.06
	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
	0.1	0.4	59.30 ± 0.11	83.10 ± 0.08	74.56 ± 0.53	80.51 ± 0.30	80.83 ± 2.24	87.10 ± 0.39
	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

Table B2: **Accuracy of compared methods across different levels of noise gap for multi-class classification.** Mis. SL: surrogate loss [25] with misspecified parameters. Est. SL: surrogate loss [25] with estimated parameters. CE: vanilla cross entropy. Peer: peer loss function [21]. When noise gap is less than 0.2, cross entropy with increasing-to-balancing reaches a higher accuracy than cross entropy at a lower noise. When noise gap is 0.3, balancing cannot compensate for the loss of increasing noise.

		BASELINES (LESS NOISE)				NOISE+ (MORE NOISE)	
Dataset	noise gap	Mis. SL	Est. SL	CE	Peer	CE	Peer
MNIST	0.1	89.59 ± 0.01	89.69 ± 0.07	86.66 ± 0.54	88.12 ± 0.01	86.81 ± 0.62	89.19 ± 0.05
	0.2	88.10 ± 0.10	88.61 ± 0.16	84.53 ± 1.60	87.21 ± 0.53	85.97 ± 0.69	89.12 ± 0.24
	0.3	84.97 ± 0.11	86.88 ± 0.17	85.24 ± 1.05	86.35 ± 0.33	81.89 ± 1.54	88.75 ± 0.19
CIFAR-10	0.1	70.90 ± 2.66	85.76 ± 1.44	88.03 ± 1.07	89.66 ± 1.18	88.69 ± 0.82	89.90 ± 0.52
	0.2	80.51 ± 4.51	86.34 ± 2.30	88.43 ± 1.29	89.36 ± 0.56	89.01 ± 1.27	90.08 ± 1.26
	0.3	81.30 ± 2.31	90.61 ± 0.52	89.78 ± 1.16	90.24 ± 1.05	87.98 ± 1.29	89.92 ± 0.92

Table B3: **Constrained learning results with group-dependent label noise.** LR: naïve logistic regression without noise correction. GPR: group-weighted peer loss [30]. Peer: peer loss [21].

				LESS NOISE		MORE NOISE	
Dataset	e_a	e_b	Metrics	LR	GPL	LR	Peer
Adult	0.1	0.3	*accuracy*	72.57	71.92	71.07	73.21
			fairness	2.37	3.39	1.83	1.95
	0.2	0.3	*accuracy*	72.4	72.92	73.07	71.8
			fairness	6.67	3.36	4.21	0.93
	0.2	0.4	*accuracy*	72.73	71.2	71.88	73.02
			fairness	6.48	2.95	3.16	1.67
	0.3	0.4	*accuracy*	73.15	73.74	71.36	72.74
			fairness	5.29	4.11	5.49	1.88
Compas	0.1	0.3	*accuracy*	63.88	63.73	64.56	64.33
			fairness	7.17	6.58	7.35	1.89
	0.2	0.3	*accuracy*	63.73	63.28	64.26	67.8
			fairness	10.52	4.47	7.10	2.76
	0.2	0.4	*accuracy*	62.60	66.03	66.22	64.15
			fairness	2.87	7.55	6.07	3.63
	0.3	0.4	*accuracy*	61.93	62.08	61.63	62.68
			fairness	17.97	3.06	7.70	3.74
Fairface	0.2	0.4	*accuracy*	86.97	87.47	88.19	87.93
			fairness	5.87	4.70	1.38	0.25
	0.1	0.3	*accuracy*	88.23	88.23	88.58	88.60
			fairness	5.53	4.93	2.11	2.17
	0.0	0.2	*accuracy*	88.61	88.53	88.90	88.85
			fairness	4.05	3.75	2.64	2.20
	0.0	0.1	*accuracy*	89.08	88.84	89.00	89.05
			fairness	3.99	3.92	2.97	2.91
	0.2	0.3	*accuracy*	88.63	88.78	88.80	88.83
			fairness	3.50	3.14	2.19	1.33