Robust Learning from Noisily Labeled Long-Tailed Data via Fairness Regularizer

Anonymous submission

Abstract

Both long-tailed and noisily labeled data frequently appear in real-world applications and impose significant challenges for learning. Most prior works treat either problem in an isolated way and do not explicitly consider the coupling effects of the two. Our empirical observation reveals that such solutions fail to consistently improve the learning when the dataset is long-tailed with label noise. Moreover, with the presence of label noise, existing methods do not observe universal improvements across different sub-populations; in other words, some sub-populations enjoyed the benefits of improved accuracy at the cost of hurting others. Based on these observations, we introduce the **F**airness **R**egularizer (**FR**), inspired by regularizing the performance gap between any two sub-populations. We show that the introduced fairness regularizer improves the performances of sub-populations on the tail and the overall learning performance. Extensive experiments demonstrate the effectiveness of the proposed solution when complemented with certain existing popular robust or class-balanced methods.

1. Introduction

Biased and noisy training datasets are prevalent and impose challenges for learning (Salakhutdinov, Torralba, and Tenenbaum 2011; Zhu, Anguelov, and Ramanan 2014; Liu 2021). The biases and noise can happen both at the sampling and label collection stages: A dataset often contains numerous sub-populations and the size of these sub-populations tends to be long-tailed distributed (Salakhutdinov, Torralba, and Tenenbaum 2011; Zhu, Anguelov, and Ramanan 2014), where the tail sub-populations have an exponentially scaled probability of being under-sampled. Meanwhile, a dataset tends to suffer from noisy labels if collected from unverified sources (Wei et al. 2022c). Most prior works treat either population bias or label noise in an isolated way and do not explicitly consider the coupling effects of the two. In particular, existing works on learning with noisy labels mainly focus on a homogeneous treatment of the entire population, and the underlying clean data is often balanced (Natarajan et al. 2013; Liu and Tao 2015; Patrini et al. 2017; Liu and Guo 2020).

The main inquiry of our paper is to understand and mitigate the possible heterogeneous effects of label noise when considering the imbalanced distribution of sub-populations. We start by presenting strong evidence of disparate impacts of sub-populations with a synthetic long-tailed noisy CIFAR-100 dataset (Krizhevsky, Hinton et al. 2009) when using existing learning with noisy labels methods. Figure 2 illustrates the per-population (100 sub-populations in all, where we consider the class information as a natural separation of sub-populations) performance comparisons between applying the traditional Cross-Entropy (CE) loss and the recently proposed robust treatment to either noisy (i.e., Label Smoothing (LS) (Lukasik et al. 2020) and PeerLoss (PL) (Liu and Guo 2020)) or long-tailed data (Focal (Lin et al. 2017), Logit-adjustment (Menon et al. 2021)). There are three main takeaways:

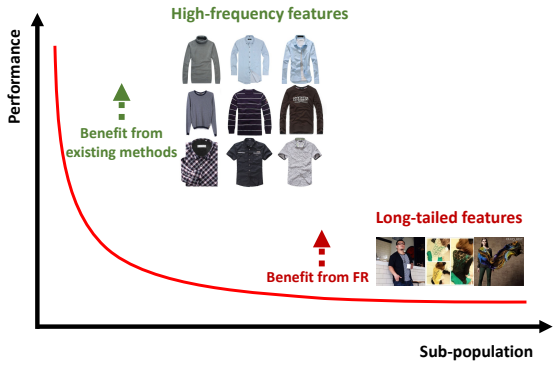

Figure 1: Overview: Different robust solutions incur varied impacts on noisily labeled long-tailed distributed sub-populations. We show adding **F**airness **R**egularizer (**FR**) between head and tail populations encourages the classifier to achieve relatively fair performances by reducing performance gaps among sub-populations, and improves the overall learning performance.

- The same robust treatment may have disparate impacts on different sub-populations, e.g., different sub-populations are improved differently by losses such as the Logit-adj loss (Menon et al. 2021).
- Different robust treatments have disparate impacts on the same part of data, e.g., LS (Lukasik et al. 2020) performs badly (almost 0 accuracies) on sub-populations with low CE accuracy (<50) and improves the others, while PL (Liu and Guo 2020) has a reversed effect that the high CE accuracy part (>50) is likely to be degraded.
- The prior works fail to address the coupling effects of population imbalance and noisy labels.

The above observations motivate us to explore how sub-

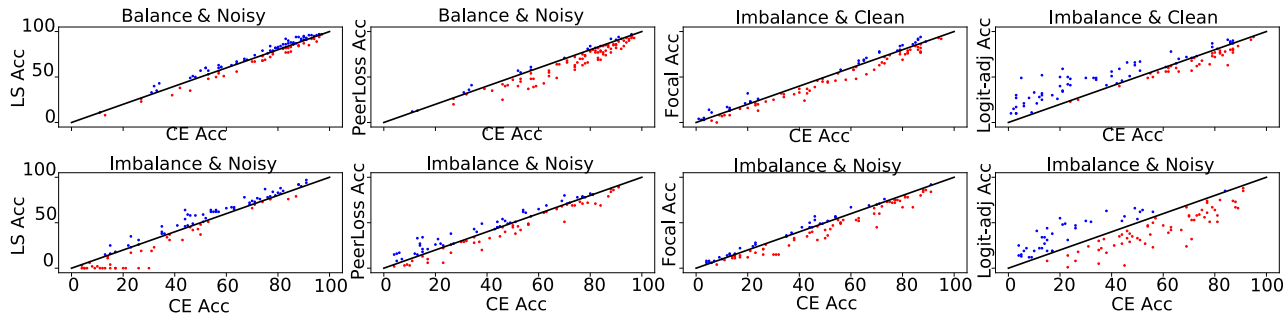

Figure 2: How each method improves per sub-population test accuracy w.r.t. CE loss on CIFAR-100 dataset. All methods are trained on 20 coarse classes in CIFAR-100. Each coarse class includes 5 different sub-populations (fine classes in CIFAR-100). For each sub-figure, x-axis indicates the CE accuracy. y-axis denotes the performance of robust/long-tail approaches. Each dot denotes the test accuracy pair $(\text{Acc}_{\text{CE}}, \text{Acc}_{\text{Method}})$ for each sub-population. The line $y = x$ means that CE performs the same as the robust treatment on a particular sub-population. The blue (red) dot above (below) the line shows the robust treatment has positive (negative) effect on this sub-population compared with CE. In sub-titles, "Balance" denotes the balanced training data (w.r.t. clean labels); "Imbalance" means the training dataset follows a long-tailed distribution where the ratio between max and min number of samples in the sub-populations is 100; "Clean": the labels of training samples are clean; "Noisy": 25.6% training samples are wrongly labeled. The test dataset is clean and balanced.

population data should be treated when learning from noisily labeled long-tailed data. This work formally investigates the influence of sub-populations when learning with long-tailed and noisily labeled data. The analysis inspires us to define a fairness regularizer for this learning task. Figure 1 overviews our work. Our contributions are primarily two-fold. We quantify the influence of sub-populations using a number of metrics and discover disparate impacts of long-tailed sub-populations when label noise presents (Section 3). Following the above observation, we propose the **F**airness **R**egularizer (**FR**), which encourages the learned classifier to reduce the performance gap between the head and tail sub-populations. As a result, **FR** not only improves the performances of tail populations but also improves overall learning performance. Extensive experiments on the CIFAR and Clothing1M datasets demonstrate the effectiveness of **FR** when complemented with certain robust or long-tailed solutions (Section 5). Contrary to most existing fairness-accuracy trade-offs observed in the literature (Hardt, Price, and Srebro 2016; Menon and Williamson 2018; Martinez, Bertran, and Sapiro 2019; Zhao and Gordon 2019; Ustun, Liu, and Parkes 2019; Islam, Pan, and Foulds 2021), we show that adding this fairness regularizer alleviates disparate impacts across populations of different sizes and improves the learning from noisily labeled long-tailed data.

1.1 Related Works

Learning with Noisy Labels Obtaining perfect annotations in supervised learning is a challenging task (Xiao et al. 2015; Luo et al. 2020; Wei et al. 2022c,d). Due to the restrictions of human recognition, noisy annotations impose challenges to performing robust training. A line of popular approaches of learning with label noise firstly estimates the noise transition matrix, and then proceeds to use this knowledge to perform loss or sample correction (Jiang et al. 2022; Natarajan et al. 2013; Liu and Tao 2015; Patrini et al. 2017; Zhu, Song, and Liu 2021; Li et al. 2022), i.e., the surrogate loss uses the transition matrix to define unbiased estimates of the true losses (Scott et al. 2013; Natarajan et al. 2013; Scott 2015; Menon et al. 2015). Noting that the estimation of the noise transition matrix is non-trivial (Zhu, Song, and Liu 2021; Zhu, Wang, and Liu 2022), another line of works aims to propose training methods without requiring knowing the noise rates, e.g., using robust loss functions (Kim et al. 2019; Liu and Guo 2020; Wei and Liu 2020; Wei et al. 2022b) training deep neural nets directly without the knowledge of noise rates (Han et al. 2018; Wei et al. 2020, 2022a; Qin, Wang, and Fu 2022), making use of the early stopping strategy (Liu et al. 2020; Xia et al. 2021a; Liu et al. 2022a,b; Huang et al. 2022), or designing a pipeline which dynamically selects/corrects and trains on "clean" samples with small loss (Cheng et al. 2021; Xia et al. 2021b, 2022b; Jiang et al. 2022; Zhang et al. 2022a). Recent works also explored the possibility of using open-set data to improve the closed-set robustness (Wei et al. 2021a; Xia et al. 2022a).

Learning with Long-Tailed Data The most relevant mainstream solution of learning with long-tailed clean data is the logit/loss adjustment approaches, which modify the loss values during the training procedure, for example, adjust the loss values w.r.t. the label frequency (Ren et al. 2020), sample influence (Park et al. 2021), or the distribution alignment between model prediction and a set of the balanced validation set (Wei et al. 2023), among many other solutions. More recently, based on the label frequencies, the logit adjustments over classic approaches (Menon et al. 2021) are proposed, either through a post-hoc modification w.r.t. a trained model or enforcement in the loss during training. Open-set data may also be used to improve complement long-tailed data (Wei et al. 2021a). Please refer to a comprehensive survey (Zhang et al. 2023) for more details.

Existing robust approaches targeted mainly the class or sub-population level balanced training data. More recently, the literature observed several approaches to address the issue of label-noise in the long-tailed tasks, through decoupled treatments for head classes and tail ones, i.e., detecting noisy labels and performing robust solutions to the head class, meanwhile adopting a self/semi-supervised learning manner to deal with the tail classes (Zhong et al. 2019; Wei et al.

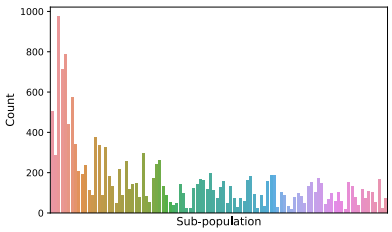

Figure 3: Count plot of a synthetic long-tailed CIFAR-100 train dataset: x-axis denotes the sub-population index; y-axis indicates the number of samples in each sub-population.

2021c; Karthik, Revaud, and Boris 2021). Beyond classes, it has been demonstrated that sub-populations with different noise rates cause disparate impacts (Liu 2021; Zhu, Luo, and Liu 2022) and need decoupled treatments (Zhu, Liu, and Liu 2021; Wang, Liu, and Levy 2021), which is more crucial for long-tailed sub-populations.

2. Preliminary

2.1 Sub-Populations of Features

In a K-class classification task, denote a set of data samples with clean labels as $S := \{(x_i, y_i)\}_{i=1}^n$, given by random variables (X, Y), which is assumed to be drawn from $\mathcal{D}$. In this work, we are interested in how sub-populations intervene with learning. Formally, we denote $G \in \{1, 2, ..., N\}$ as the random variable for the index of sub-population, and each sample (x_i, y_i) is further associated with a g_i. The set of sub-population k could then be denote as $\mathcal{G}_k := \{i : g_i = k\}$. We consider a long-tail scenario where the head population and the tail population differ significantly in their sizes, i.e., $\max_k |\mathcal{G}_k| \gg \min_{k'} |\mathcal{G}_{k'}|$.

Consider Figure 3 for an example of sub-population separations using the CIFAR-100 dataset (Krizhevsky, Hinton et al. 2009): images are grouped into 20 coarse classes, and each coarse class could be further categorized into 5 fine classes. For example, the coarse class "aquatic mammals" was further split into "beaver", "dolphin", "otter", "seal", and "whale". From Figure 3, we observe a strong imbalanced distribution of different sub-populations and a long-tailed pattern. In Section 5.1, we provide more details on long-tail data generation models for our synthetic experiments.

2.2 Our Task

In practice, obtaining "clean" labels from human annotators is both time-consuming and expensive. The obtained human-annotated labels usually consist of certain noise (Xiao et al. 2015; Lee et al. 2018; Jiang et al. 2020; Wei et al. 2022c). The flipping from clean to noisy labels is described by the noise transition matrix $T(X)$, with its element denoted by $T_{ij}(X) = \mathbb{P}(\widetilde{Y} = j|Y = i, X)$. We denote the obtained noisy training dataset as $\widetilde{S} := \{(x_i, \tilde{y}_i)\}_{i=1}^n$, given by random variables $(X, \widetilde{Y})$, which is assumed to be drawn from $\widetilde{\mathcal{D}}$.

Though we only have access to noisily labeled long-tailed data $\widetilde{S}$, our goal remains to obtain the optimal classifier with respect to a clean and balanced distribution $\mathcal{D}$: $\min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)\sim\mathcal{D}} [\ell(f(X), Y)]$, where f is the classifier chosen from the hypothesis space $\mathcal{F}$, and $\ell(\cdot)$ is a calibrated loss function (e.g., CE) Furthermore, we do not assume the knowledge of the sub-population information during training. We are interested in how sub-populations intervene with the learning performance and how we could improve by treating the sub-populations with special care.

3. Disparate Influences of Sub-Populations

In this section, we empirically illustrate the disparate influence of sub-populations when learning with noisily labeled data. Inspired by the literature on using the influence function to capture the impact of training samples, we define influence metrics at the sub-population level and perform a multi-faceted evaluation of how imbalanced sub-populations affect the learning performance. We take the long-tail populations for illustration and defer the results of head populations to Appendix C.4.

Influences: In the literature of explainable deep learning, the notions of influence can be different, e.g., the influences of features on an individual sample prediction (Ribeiro, Singh, and Guestrin 2016; Sundararajan, Taly, and Yan 2017; Lundberg and Lee 2017; Feldman and Zhang 2020), the influences of features on the loss/accuracy of the model (Owen and Prieur 2017; Owen 2014), the influences of training samples on the loss/accuracy of the model (Jia et al. 2019). In this section, we focus on the influence of a sub-population on both the sub-population level and the individual sample level.

We now empirically demonstrate the role of sub-populations when measuring the test accuracy, and the prediction of model confidence on test samples. For the synthetic long-tailed noisy training dataset, we first flip clean labels of the class-balanced CIFAR-10 dataset to any other classes, and there exist 20% wrong labels in all. We then adopt the class-imbalanced (Cui et al. 2019) CIFAR-10 dataset to select a long-tailed distributed amount of samples for each class (by referring to clean labels). As for the separation of sub-populations, we adopt the k-means clustering to categorize the extracted features of each feature given by the Image-Net pre-trained model. Since sub-population information sometimes may not be available for training use, understanding the influences of such division of sub-populations is beneficial. More details can be found at Sec 5.1.

We explore the influences of tail sub-populations on performances of cross-entropy (ce) loss, the forward loss correction (fw) (Patrini et al. 2017), label smoothing (ls) (Lukasik et al. 2020), and the peer loss (pl) (Liu and Guo 2020). There are 17 sub-populations (train) with less than 50 instances considered as the tail section. We illustrate observations on several randomly selected tail sub-populations. Results of more sub-populations are deferred to Appendix C.4.

3.1 Influences on Sub-Population (Test Accuracy)

We start with the influence of sub-populations in the test set. We adopt the (population-level) test accuracy changes when removing all samples in the sub-population $\mathcal{G}_i$ during the training procedure to capture the influences of a sub-population on each sub-population at the test set:

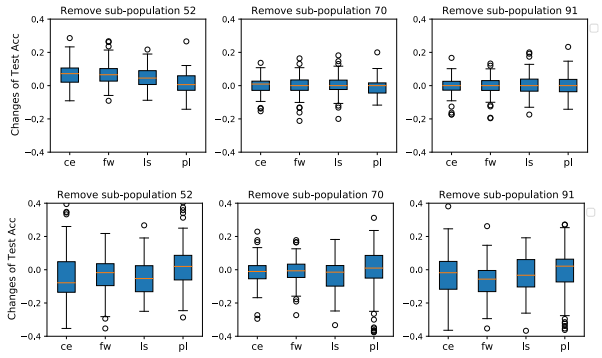

Figure 4: Box plot of the population-level test accuracy changes when removing all samples of a selected long-tailed sub-population during the training w.r.t. 4 methods. (Left: trained on clean labels; Right: trained on noisy labels.)

$$\mathrm{Acc_p}(\mathcal{A}, \widetilde{S}, i, j) = \mathbb{P}_{\substack{f \leftarrow \mathcal{A}(\widetilde{S}) \\ (X', Y', G=j)}} (f(X') = Y') - \mathbb{P}_{\substack{f \leftarrow \mathcal{A}(\widetilde{S}^{\setminus i}) \\ (X', Y', G=j)}} (f(X') = Y'),$$

where in the above two quantities, $f \leftarrow \mathcal{A}(\widetilde{S})$ indicates that the classifier f is trained from the whole noisy training dataset $\widetilde{S}$ via Algorithm $\mathcal{A}$, $f \leftarrow \mathcal{A}(\widetilde{S}^{\setminus i})$ means f is trained on $\widetilde{S}$ without samples in the sub-population $\mathcal{G}_i$. $(X', Y', G = j)$ denotes the test data distribution given that the samples are from the j-th sub-population.

In Figure 4, the x-axis denotes the loss function for training, and the y-axis visualized the distribution of $\{\mathrm{Acc_p}(\mathcal{A}, S, i, j)\}_{j \in [100]}$ (left) and $\{\mathrm{Acc_p}(\mathcal{A}, \widetilde{S}, i, j)\}_{j \in [100]}$ (right) for several randomly selected long-tail sub-populations ($i = 52, 70, 91$, results of more populations can be found in Appendix C.1) under each robust method, where "S" refers to the clean training samples and "$\widetilde{S}$" denotes the noisy version. The blue zone shows the 25-th percentile (Q_1) and 75-th percentile (Q_3) accuracy changes, and the orange line indicates the median value. Accuracy changes that are drawn as circles are viewed as outliers. Note that all sub-figures (distributions) have the same amount of samples, it is clear to observe the left three figures have lower variance than the right ones, indicating that:

Observation 0.1. The tail sub-populations in the noisy training tend to have a more significant influence on the test accuracy than that in clean training.

3.2 Influences on Samples (Prediction Confidence)

Note that grouping testing samples into classes/sub-populations for analysis may ignore some individual behavior changes, we next consider the influence of sub-populations on the individual test samples. Instead of insisting on the accuracy measure, we adopt the model prediction confidence as a proxy, to see how each test sample was influenced. And we introduce $\mathrm{Infl}(\mathcal{A}, \widetilde{S}, i, j)$ to quantify the influence of a sub-population on a specific test sample:

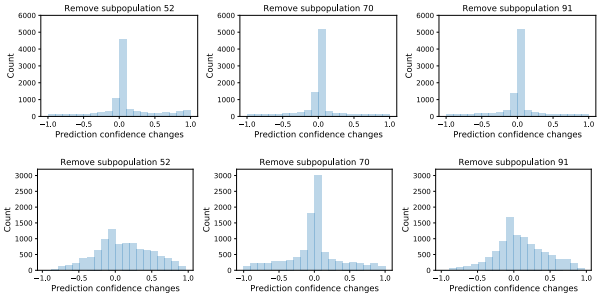

Figure 5: Distribution plot w.r.t. the changes of model confidence on the test data samples using CE loss (Left: trained on clean labels; Right: trained on noisy labels). See Appendix C.1 for more details.

$$\mathrm{Infl}(\mathcal{A}, \widetilde{S}, i, j) = \mathbb{P}_{f \leftarrow \mathcal{A}(\widetilde{S})}(f(x'_j) = y'_j) - \mathbb{P}_{f \leftarrow \mathcal{A}(\widetilde{S}^{\setminus i})}(f(x'_j) = y'_j).$$

As shown in Figure 5, we visualize $\mathrm{Infl}(\mathcal{A}, S, i, j)$ (left) and $\mathrm{Infl}(\mathcal{A}, \widetilde{S}, i, j)$ (right), where $j \in [10000]$ means 10K test samples. For example, $\mathrm{Infl}(\mathcal{A}, \widetilde{S}, i, j) = -1$ means the model prediction confidence on test sample x'_j changed from 0 to 1. With the presence of label noise, we observe:

Observation 0.2. Compared with clean training, removing certain tail sub-populations in the noisy training leads to significant changes/influences on the model prediction confidence of more test samples.

To conclude, we have shown that given certain robust methods, significant disparate impacts on sub-populations are observed, when learning from long-tailed data with noisy labels. Such impacts also differ when complemented with different robust solutions, i.e., robust loss functions implicitly incur disparate impacts on the populations/samples. Recall in Figure 2, we revealed that existing robust treatments may result in unfair performances among sub-populations, when learning from noisily labeled long-tailed data. All these observations motivate us to explore ways that will reduce the gaps between the head and the tail populations.

4. Fairness Regularizer (FR)

In this section, we propose to assign fairness constraints to the learning objective. Leveraged into its Lagrangian form, such fairness constraints could be viewed as fairness regularizers that explicitly encourage the classifier to achieve fair performances among sub-populations. We name our solution the **F**airness **R**egularizer (**FR**), which encourages the learned classifier to achieve fair performance across sub-populations.

4.1 Fairness Constraints

Note that when learning with robust methods, the classifier tends to result in fitting on certain sub-populations more easily. We propose to constrain the classifier's performance on sub-populations:

$$\min_{f:\mathrm{domain}(X) \to [K]} \mathbb{E}_{(X,\widetilde{Y}) \sim \widetilde{\mathcal{D}}}[\ell(f(X), \widetilde{Y})],$$
$$\text{s.t. Constraint w.r.t. } \mathbb{P}(f(X) = \widetilde{Y} \mid G = i), \qquad (1)$$

where ℓ is a generic loss function that could be any robust losses and the ultimate goal of the classifier f is categorizing the feature X into a specific class within $[K]$. Since we do not wish certain sub-populations to fall much behind others, i.e., in terms of accuracy, we constrain the performance gap between any two sub-populations by adopting the following constraint for Eqn. (1), specifically, for any sub-population $i \in [N]$, we require its performance to have a bounded distance from the average performance. Define $\mathsf{Dist}_i := \left|\mathbb{P}(f(X) = \widetilde{Y} \mid G = i) - \mathbb{P}(f(X) = \widetilde{Y})\right|$ as the distance (absolute performance gap), then the optimization problem is formulated as:

$$\min_{f:X\to[K]} \mathbb{E}_{(X,\widetilde{Y})\sim\widetilde{\mathcal{D}}}[\ell(f(X),\widetilde{Y})], \quad \text{s.t. } \mathsf{Dist}_i \le \delta, \forall i \in [N], \tag{2}$$

where $\delta \ge 0$ is a constant. Setting $\delta = 0$ implies that the classifier should achieve fair performances among all sub-populations, in order to satisfy the constraints.

4.2 Using Fairness Constraints as a Regularizer

In practice, forcing sub-populations to achieve absolutely fair or equalized performances (i.e., accuracy) may produce side effects. For example, one trivial solution to achieve $\delta = 0$ is simply reducing the performance of all the other sub-populations to be aligned with the worst sub-population, leading to poor overall performance. Even though we can fine-tune δ to set an appropriate tolerance of the gap, the sub-population with the worst performance may still violate the constraint. Noting our goal is to improve the overall performance on clean and balanced test data, it is arguably a better strategy to not over-addressing the worst sub-population.

To balance the trade-off between mitigating the disparate impacts among sub-populations and the possible negative effect due to constraining, rather than strictly solving the constrained optimization problem in Eqn. (2), we use the constraint as a regularizer by adopting the Lagrangian form:

$$\min_{f:X\to[K]} \mathcal{L}_\lambda(f) := \mathbb{E}_{(X,\widetilde{Y})\sim\widetilde{\mathcal{D}}}[\ell(f(X),\widetilde{Y})] + \underbrace{\boxed{\sum_{i=1}^{N} \lambda_i \mathsf{Dist}_i}}_{\to \textbf{FR}},$$

where $\lambda_i \ge 0$. Different from the traditional dual ascent of Lagrange multipliers (Boyd et al. 2011), we fix λ_i during our training. Intuitively, applying dual ascent is likely to result in a large λ_i on the worst sub-population, inducing possible negative effects as we discussed above. Therefore, in such a minimization task, the accuracy/performance gaps between sub-populations are encouraged to be small and do not have to be exactly lower than any threshold. To clarify, we do not require strict fair performance among sub-populations, instead, we wish to improve the worst group performance at the minimum cost of the better group. Hence, we did explore the usage of other fairness constraints since all these definitions/constraints will serve with the same purpose – avoiding the performance gap among sub-populations from being overly large, given noisy labeled long-tailed data.

Implementation Denote by $\boldsymbol{f}_x[\tilde{y}]$ the model's prediction probability on the noisy label $\tilde{y}$ given input x. Noting the probability in Dist_i is non-differentiable w.r.t f, we apply the following empirical relaxation (Wang, Wang, and Liu 2022):

$$\mathsf{Dist}_i := \left| \frac{\sum_{k=1}^{N} \boldsymbol{f}_{x_k}[\tilde{y}_k] \cdot 1(g_k = i)}{\sum_{k=1}^{N} 1(g_k = i)} - \frac{\sum_{k=1}^{N} \boldsymbol{f}_{x_k}[\tilde{y}_k]}{N} \right|, \tag{3}$$

where $1(g_k = i) = 1$ when $g_k = i$ and 0 otherwise. For simplicity, we set all λ_i to a constant λ.

To demonstrate why **FR** helps with improving the learning from noisily labeled long-tailed data, we will provide extensive experiment studies in the next section. We also adopted a binary Gaussian example and provide Observation 0.3. Detailed discussions are deferred to Appendix A.3.

Observation 0.3. Theoretically, we show the connection between error probability under the noisy data distribution and under the clean data distribution. Then, we provide insights on how **FR** mitigates the incurred bias term brought by the noisy data distribution.

5. Experiments

In this section, we verify the effectiveness of **FR** on the synthetic long-tailed noisy CIFAR datasets (Krizhevsky, Hinton et al. 2009) and a real-world large-scale noisily labeled long-tailed dataset Clothing1M (Xiao et al. 2015).

5.1 Experiment Designs on Synthetic CIFAR

We empirically test the performance of **FR** on CIFAR-10, and CIFAR-100 (Krizhevsky, Hinton et al. 2009).

Generation of Synthetic Long-Tailed Data with Noisy Labels Note that the class information could be viewed as a special case of sub-populations, in this subsection, we treat classes as the natural separation of sub-populations and consider the class-imbalance experiment setting with noisy labels. For the balanced K-class classification task with n samples per class, the synthetic long-tail setting assumes that $k-$th class has only $n/(r^{\frac{k-1}{K-1}})$ samples by referring to the ground-truth labels (Cui et al. 2019). We adopt two label-noise transition models below.

Model 1 (Imb): The entries of the noise transition matrix are given by $T_{i,j} := \mathbb{P}(\widetilde{Y} = j | Y = i, X = x)$: $T_{i,j}$ returns $1-\rho$ if $i = j$; otherwise, $\frac{\mathbb{P}(Y=j)\cdot\rho}{1-\mathbb{P}(Y=i)}$. ρ is viewed as the overall error/noise rate. The Imb noise model (Wei et al. 2021c) assumes that samples are more likely to be mislabeled as frequent ones in real-world situations.

Model 2 (Sym): The generation of the symmetric noisy dataset is adopted from (Kim et al. 2019), where it assumed that $T_{i,j} = \frac{\rho}{K-1}, \forall i \neq j$, indicating that any other classes $i \neq j$ has the same chance of being flipped to class j. The diagonal entry $T_{i,i}$ (chance of a correct label) becomes $1 - \rho$.

For both noise settings, we test **FR** with noise rates $\rho \in \{0.2, 0.5\}$, meaning the proportion of wrong labels in the long-tailed training set is 0.2 or 0.5.

Table 1: Performance comparisons on synthetic long-tailed noisy CIFAR datasets, best-achieved averaged accuracy on a class-balanced test data are reported. Results in **bold**: **FR** improves the performance of the baseline methods, respectively.

Noise type: Imbalance Noise												
Noise Ratio	CIFAR-10 ($\rho = 0.2$)			CIFAR-10 ($\rho = 0.5$)			CIFAR-100 ($\rho = 0.2$)			CIFAR-100 ($\rho = 0.5$)		
Imbalance Ratio	$r = 10$	$r = 50$	$r = 100$	$r = 10$	$r = 50$	$r = 100$	$r = 10$	$r = 50$	$r = 100$	$r = 10$	$r = 50$	$r = 100$
CE	79.75	65.98	60.03	65.38	47.51	37.44	46.02	31.44	26.98	29.58	16.93	13.87
CE + FR (KNN)	**80.46**	**69.00**	**61.64**	**65.87**	46.69	**39.97**	**46.18**	31.03	**27.60**	**30.25**	**16.79**	**15.19**
CE + FR (G2)	**80.44**	**67.29**	**65.12**	**68.62**	**49.43**	**39.69**	**46.38**	**32.32**	**28.53**	**32.35**	**19.03**	**15.93**
LS (Lukasik et al. 2020)	82.52	69.08	59.07	67.73	36.17	32.92	47.80	33.66	26.36	34.02	17.28	14.10
LS + FR (KNN)	**82.78**	**70.06**	**59.27**	**68.99**	**36.55**	**36.63**	**48.27**	33.01	**27.60**	32.01	17.14	14.07
LS + FR (G2)	82.02	**70.24**	**60.33**	**70.50**	**44.11**	**35.49**	47.30	**33.86**	**29.67**	**34.51**	**17.84**	**16.68**
NLS (Wei et al. 2021b)	79.91	65.98	58.82	64.74	41.01	34.16	46.05	31.48	27.09	29.86	16.84	13.87
NLS + FR (KNN)	**80.17**	**68.61**	**62.88**	**68.65**	**47.42**	**36.79**	45.72	**32.25**	27.01	28.85	**17.23**	**14.18**
NLS + FR (G2)	**80.36**	**68.25**	**63.50**	**69.70**	**49.01**	**38.26**	43.15	**33.78**	**28.69**	**32.30**	**19.62**	**15.64**
Focal (Lin et al. 2017)	76.24	64.16	57.68	62.40	40.25	34.56	43.63	29.10	24.88	26.93	14.45	12.57
Focal + FR (KNN)	**77.54**	62.97	57.24	61.47	**42.28**	**37.04**	42.44	28.90	**25.14**	**28.34**	**16.02**	**13.27**
Focal + FR (G2)	**78.56**	**66.07**	56.55	**64.10**	**43.61**	**38.15**	**45.63**	**31.87**	**27.58**	**29.80**	**17.67**	**15.30**
PL (Liu and Guo 2020)	78.43	55.61	54.20	47.71	31.96	30.13	45.32	33.05	29.91	28.01	20.25	16.65
PL + FR (KNN)	**79.50**	**65.37**	53.36	**51.82**	**35.68**	**30.16**	44.89	**33.12**	28.63	27.66	19.79	**17.72**
PL + FR (G2)	**78.79**	**66.16**	**54.39**	**50.72**	**33.22**	28.01	44.78	**33.35**	29.51	**29.82**	20.15	**16.81**
Logit-adj (Menon et al. 2021)	82.09	73.23	68.18	68.30	51.51	42.17	47.28	33.11	29.47	30.92	17.97	14.68
Logit-adj + FR (G2)	**82.92**	**75.67**	**72.20**	**73.72**	**55.09**	40.85	41.21	**35.39**	28.84	27.57	**18.93**	**15.44**
Logit-adj + FR (KNN)	**82.48**	**73.65**	**68.48**	**70.89**	49.23	**42.93**	**47.66**	**33.18**	**29.50**	**31.85**	17.59	**15.25**
Noise type: Symmetric Noise												
Noise Ratio	CIFAR-10 ($\rho = 0.2$)			CIFAR-10 ($\rho = 0.5$)			CIFAR-100 ($\rho = 0.2$)			CIFAR-100 ($\rho = 0.5$)		
Imbalance Ratio	$r = 10$	$r = 50$	$r = 100$	$r = 10$	$r = 50$	$r = 100$	$r = 10$	$r = 50$	$r = 100$	$r = 10$	$r = 50$	$r = 100$
CE	80.70	65.04	61.80	70.48	51.53	36.44	46.02	30.93	26.98	29.93	16.70	4.76
CE + FR (KNN)	**81.19**	**69.95**	63.97	**71.75**	**52.93**	**45.63**	**46.33**	30.82	**27.19**	**31.12**	**17.68**	**15.39**
CE + FR (G2)	**81.64**	**70.84**	**65.14**	**71.44**	**56.50**	**46.33**	**47.70**	**34.34**	**30.78**	**31.58**	**21.70**	**19.10**
LS (Lukasik et al. 2020)	83.23	71.69	65.69	72.85	50.59	30.98	47.90	33.81	29.95	26.56	21.74	19.39
LS + FR (KNN)	**83.28**	70.64	60.91	**73.92**	**53.01**	**43.48**	**49.05**	33.40	**30.05**	**34.86**	20.73	19.10
LS + FR (G2)	82.22	70.85	62.43	**74.59**	**54.15**	**44.77**	**48.16**	**34.08**	**30.69**	**36.40**	**22.06**	**20.10**
NLS (Wei et al. 2021b)	80.79	66.22	61.47	70.11	50.57	36.55	46.11	31.14	27.32	30.51	17.16	5.18
NLS + FR (KNN)	**81.08**	**69.29**	**63.58**	**70.27**	**54.86**	36.50	**48.20**	**35.03**	**28.29**	28.87	**19.10**	**6.65**
NLS + FR (G2)	**81.37**	**70.60**	**64.73**	**71.30**	**56.24**	**37.29**	**47.67**	**34.32**	**30.75**	29.62	**22.17**	**8.04**
Focal (Lin et al. 2017)	77.77	61.54	56.02	67.20	43.12	38.20	35.93	23.23	21.84	27.31	16.18	14.71
Focal + FR (KNN)	**78.03**	**64.57**	**56.77**	**67.87**	41.89	36.34	**42.79**	**30.17**	**25.08**	**28.22**	**16.37**	14.50
Focal + FR (G2)	**78.83**	**65.56**	**60.35**	**68.21**	**47.09**	**41.74**	**46.33**	**32.56**	**27.77**	**29.70**	**16.47**	**15.29**
PL (Liu and Guo 2020)	79.73	66.82	42.12	55.52	33.18	33.06	44.60	32.91	28.69	27.38	18.52	17.25
PL + FR (KNN)	79.42	64.91	**58.80**	53.86	**38.41**	32.71	**45.60**	32.32	28.34	**27.63**	**18.86**	16.48
PL + FR (G2)	79.37	66.71	**58.98**	**55.68**	**38.08**	**33.52**	**46.83**	**33.17**	**29.67**	**28.12**	**19.48**	**17.62**
Logit-adj (Menon et al. 2021)	80.50	62.42	50.28	60.38	32.45	27.32	46.50	29.24	23.80	28.79	12.65	9.22
Logit-adj + FR (KNN)	**80.66**	62.07	**51.04**	**62.32**	31.23	22.41	**47.22**	**29.34**	**24.70**	**29.95**	12.44	**9.28**
Logit-adj + FR (G2)	**81.82**	**62.62**	**52.35**	**63.34**	31.14	21.93	**48.13**	**30.18**	**24.06**	**29.35**	12.37	**9.26**

Separation of $\mathcal{G}_i$ We consider two kinds of sub-population separation methods for **FR**.

- **Separation with Clustering Methods:** $\forall x \in X$, the representation of feature x is given by the representation extractor $\phi(x)$, where $\phi(\cdot) : X \to \mathbb{R}^d$ maps the feature x to a d-dimensional representation vector. Given a distance metric DM (i.e., the Euclidean distance), the distance between two extracted representations $\phi(x_1), \phi(x_2)$ is $\text{DM}(\phi(x_1), \phi(x_2))$. The sub-population could be separated through clustering algorithms such as k-means ($k = N$ here). Admittedly, obtaining a good representation extractor is non-trivial, we want to highlight that the separation of sub-populations is not highly demanding on the quality of the representation extractor, and the focus is to perform fairness regularizations on varied features.
- **Separation Directly via Pre-Trained Models:** In this case, $\forall x \in X$, we adopt an (Image-Net) pre-trained model for the separation, i.e., such a feature extractor $\phi(\cdot)$ maps each x into a $d = N$-dimensional representation vector so that all features are automatically categorized into N sub-populations.

5.2 Experiment Results on CIFAR

In Table 1, we empirically show how **FR** helps with improving the classifier's performance when complemented with several methods in robust losses as well as approaches in class-imbalanced learning, under synthetic class-imbalanced CIFAR datasets with noisy labels, including Cross-Entropy loss (CE), Label Smoothing (LS) (Lukasik et al. 2020), Negative Label Smoothing (NLS) (Wei et al. 2021b), Focal Loss (Lin et al. 2017), PeerLoss (PL) (Liu and Guo 2020), and Logit-adjustment (Logit-adj) (Menon et al. 2021). We fix the same training samples and labels for all methods. More details are available in Appendix C.2.

For **FR**, we adopted the fixed λ for all sub-populations. We consider two types of sub-population separation methods: (i) KNN clustering, which splits the extracted features into K clusters, with K being the number of classes; (ii) Generate the separation by referring to the direct prediction made by a (Image-Net) pre-trained model. In our experiments, this method separates features into a head and a tail sub-population, and the ratio w.r.t. the amount of samples between two sub-populations is ≈ 5.

Results In Table 1, we provide the baseline performance as well as the corresponding performances when **FR** is introduced. **FR** (KNN) denotes the scenario where we adopt the KNN clustering for sub-population separation, and the number of sub-populations is the same as the number of classes. We did not consider the noisy (class) labels as the sub-population index due to the fact that the noisy labels may contain the wrong ones. Empirically, we observe that **FR** (KNN) consistently improves the baseline methods on the class-imbalanced CIFAR-10 dataset, under the Imb and Sym noise. However, **FR** (KNN) could not improve significantly on the class-imbalanced CIFAR-100 dataset. One reason is that, in the batch update, the number of samples

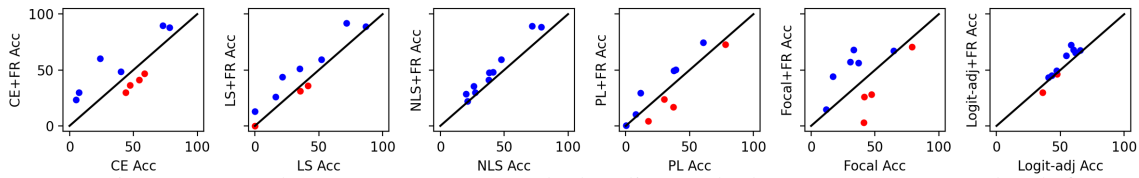

Figure 6: How **FR** improves per class test accuracy w.r.t. the baseline method on CIFAR-10. In each sub-figure, the x-axis indicates the accuracy of a baseline. y-axis denotes the performance of baseline when **FR** is introduced. Each dot denotes the test accuracy pair $(\text{Acc}_{\text{Method}}, \text{Acc}_{\text{Method+FR}})$ for each sub-population. The black line $y = x$ stands for the case that **FR** has no effects on a particular sub-population. The blue (red) dot above (below) the line shows the robust treatment has positive (negative) effect on this sub-population compared with CE.

Table 2: Paired student t-test results w.r.t. the effectiveness of **FR**. Rows marked with "$\surd$" mean **FR** improve the performance of the baseline methods significantly (p-value satisfies that $p < 0.1$ and the statistics is positive); "–" indicates there exist no significant differences after adopting **FR**.

		CIFAR-10			CIFAR-100		
Method	**FR Type**	**statistics**	**p-value**	**Better**	**statistics**	**p-value**	**Better**
CE	FR (KNN)	2.962	0.013	$\surd$	1.489	0.165	–
CE	FR (G2)	4.313	0.001	$\surd$	3.083	0.010	$\surd$
LS	FR (KNN)	1.214	0.250	–	0.748	0.470	–
LS	FR (G2)	1.851	0.091	$\surd$	1.926	0.080	$\surd$
NLS	FR (KNN)	4.235	0.001	$\surd$	1.692	0.119	–
NLS	FR (G2)	4.909	<0.000	$\surd$	3.237	0.008	$\surd$
PL	FR (KNN)	1.859	0.090	$\surd$	-0.620	0.548	–
PL	FR (G2)	1.847	0.092	$\surd$	2.345	0.039	$\surd$
Focal	FR (KNN)	0.886	0.395	–	2.218	0.049	$\surd$
Focal	FR (G2)	5.249	<0.000	$\surd$	4.105	0.002	$\surd$
Logit-adj	FR (KNN)	1.171	0.266	–	-0.419	0.684	–
Logit-adj	FR (G2)	0.255	0.803	–	2.410	0.035	$\surd$

in each sub-population is too small (the average number is $128/100 = 1.28$), resulting in large variance in calculating **FR** as Eqn. (3). As an alternative, we report the performance of **FR** (G2) as well, where samples are categorized into 2 sub-populations by the (Image-Net) pre-trained model. Surprisingly, **FR** (G2) improves the performance of 6 baselines in most settings, as highlighted in Table 1. Constraining the classifier to have relative fairness performances is beneficial when learning with noisy and long-tailed data.

We further adopt the CIFAR-10 dataset ($\rho = 0.5, r = 50$) and visualize how **FR** influences the per-class accuracy by referring to the performance of each baseline. Each blue point in Figure 6 indicates the scenario where **FR** improves the test accuracy of a class over the corresponding baseline. Points in the lower left corner (where tail populations are usually located) further illustrate that **FR** consistently improves the performance of tail sub-populations.

Hypothesis Testing w.r.t. FR We adopt paired student t-test to verify the conclusion that **FR** helps with improving the test accuracy. In Table 2, positive statistics indicate that the **FR** generally improves the performance (test accuracy) of the baseline method. The p-value that is smaller than 0.1 means there exist significant differences between the two accuracy lists. In such scenarios, we should reject the null hypothesis and adopt the alternative hypothesis. Table 2 shows that **FR** (G2) brings significant performance improvements in most settings (5/6 in CIFAR-10 and 6/6 in CIFAR-100), indicating the effectiveness of our method. Besides, **FR** (KNN) shows significant performance improvements only in several settings (but there are still improvements in most cases), which can be explained by our previous discussion that a large number of sub-populations may make the learning unstable. More details appear in Appendix C.2.

Table 3: Performance comparisons on real-world imbalanced noisily labeled dataset (Clothing1M), best and last-epoch achieved test accuracy are reported. Results in **bold** mean **FR** improves the performance of the baseline methods, respectively. Performances of **FR** with different λs are provided.

Method	λ	0.0	0.1	0.2	0.4	0.6	0.8	1.0	2.0
CE	Best	72.68	72.44	**72.93**	**72.74**	**73.10**	**72.80**	**72.99**	72.45
	Last	72.22	71.99	**72.25**	**72.24**	**72.51**	**72.53**	**72.58**	72.20
LS	Best	72.55	**72.71**	**72.69**	72.34	72.41	72.44	**72.70**	**72.56**
	Last	72.03	**72.11**	**72.14**	**72.12**	**72.12**	**72.06**	**72.33**	**72.24**
NLS	Best	74.46	**74.48**	**74.47**	**74.49**	**74.48**	**74.49**	**74.49**	**74.50**
	Last	74.00	73.99	73.97	73.98	73.98	73.97	73.97	73.97
PL	Best	73.00	**73.27**	**73.13**	**73.15**	**73.13**	**73.22**	**73.08**	**73.02**
	Last	72.73	**72.91**	**72.87**	72.69	**72.76**	**73.12**	72.71	72.69
Focal	Best	72.71	72.60	**72.71**	72.60	**72.92**	72.66	**72.91**	72.46
	Last	72.16	**72.21**	72.04	**72.18**	**72.30**	**72.36**	**72.51**	**72.46**
Logit-adj	Best	72.43	**72.52**	**72.48**	71.88	72.22	**72.45**	**72.67**	72.06
	Last	72.22	72.15	72.14	71.58	71.83	71.94	**72.23**	71.92

5.3 Experiment Results on Clothing1M

Clothing1M is a large-scale feature-dependent human-level noisy clothes dataset. We adopt the same baselines as reported in CIFAR experiments for Clothing1M. More detailed descriptions are given in Appendix C.3.

We try implementing **FR** with different λ chosen from the set $\{0.0, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 2.0\}$, where $\lambda = 0.0$ indicates the training of baseline methods without **FR**. In Table 3, the default setting of **FR** ($\lambda = 1.0$) consistently reaches competitive performances by comparing to other λs, except for the experiments w.r.t. NLS. Besides, we observe that most positive λs that are close to $\lambda = 1.0$ tend to have better performances than those close to $\lambda = 0.0$, indicating the effectiveness as well as hyper-parameter in-sensitiveness of the introduced fairness regularizer.

Conclusions In this paper, we qualitatively and quantitatively analyzed the influence of sub-populations under various metrics, where we observed disparate impacts incurred by sub-populations, especially when the label noise presents. What is more, our experiment results also reveal that existing robust solutions improve the performance of certain sub-populations at the cost of hurting others, hence leading to unfair performances among sub-populations. We then propose **F**airness **R**egularizer (**FR**), which encourages the learned classifier to achieve fair performances across sub-populations. Extensive experiment results demonstrate the effectiveness of **FR**, indicating that fairness constraints improve the learning from noisily labeled long-tailed data. One limitation is that our proposed method has only been tested on image classification tasks. The performance on other tasks needs more exploration. We defer more detailed discussions to the beginning of the Appendix.

References

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J.; et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1): 1–122.

Cheng, H.; Zhu, Z.; Li, X.; Gong, Y.; Sun, X.; and Liu, Y. 2021. Learning with Instance-Dependent Label Noise: A Sample Sieve Approach. In *International Conference on Learning Representations*.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Feldman, V.; and Zhang, C. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33: 2881–2891.

Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, 8527–8537.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Huang, H.; Kang, H.; Liu, S.; Salvado, O.; Rakotoarivelo, T.; Wang, D.; and Liu, T. 2022. PADDLES: Phase-Amplitude Spectrum Disentangled Early Stopping for Learning with Noisy Labels. *arXiv preprint arXiv:2212.03462*.

Islam, R.; Pan, S.; and Foulds, J. R. 2021. Can we obtain fairness for free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 586–596.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Hynes, N.; Gürel, N. M.; Li, B.; Zhang, C.; Song, D.; and Spanos, C. J. 2019. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1167–1176. PMLR.

Jiang, L.; Huang, D.; Liu, M.; and Yang, W. 2020. Beyond synthetic noise: Deep learning on controlled noisy labels. In *International Conference on Machine Learning*, 4804–4815. PMLR.

Jiang, Z.; Zhou, K.; Liu, Z.; Li, L.; Chen, R.; Choi, S.-H.; and Hu, X. 2022. An information fusion approach to learning with instance-dependent label noise. In *International Conference on Learning Representations*.

Karthik, S.; Revaud, J.; and Boris, C. 2021. Learning From Long-Tailed Data With Noisy Labels. *arXiv preprint arXiv:2108.11096*.

Kim, Y.; Yim, J.; Yun, J.; and Kim, J. 2019. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 101–110.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lee, K.-H.; He, X.; Zhang, L.; and Yang, L. 2018. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5447–5456.

Li, S.; Xia, X.; Zhang, H.; Zhan, Y.; Ge, S.; and Liu, T. 2022. Estimating noise transition matrix with label correlations for noisy multi-label learning. In *Advances in Neural Information Processing Systems*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, S.; Liu, K.; Zhu, W.; Shen, Y.; and Fernandez-Granda, C. 2022a. Adaptive early-learning correction for segmentation from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2606–2616.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33: 20331–20342.

Liu, S.; Zhu, Z.; Qu, Q.; and You, C. 2022b. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*, 14153–14172. PMLR.

Liu, T.; and Tao, D. 2015. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3): 447–461.

Liu, Y. 2021. Understanding Instance-Level Label Noise: Disparate Impacts and Treatments. In *International Conference on Machine Learning*, 6725–6735. PMLR.

Liu, Y.; and Guo, H. 2020. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, 6226–6236. PMLR.

Lukasik, M.; Bhojanapalli, S.; Menon, A.; and Kumar, S. 2020. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, 6448–6458. PMLR.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777.

Luo, T.; Li, X.; Wang, H.; and Liu, Y. 2020. Research Replication Prediction Using Weakly Supervised Learning. In *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*.

Martinez, N.; Bertran, M.; and Sapiro, G. 2019. Fairness with minimal harm: A pareto-optimal approach for healthcare. *arXiv preprint arXiv:1911.06935*.

Menon, A.; Van Rooyen, B.; Ong, C. S.; and Williamson, B. 2015. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, 125–134.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*.

Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, 107–118. PMLR.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. In *Advances in neural information processing systems*, 1196–1204.

Owen, A. B. 2014. Sobol'indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1): 245–251.

Owen, A. B.; and Prieur, C. 2017. On Shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1): 986–1002.

Park, S.; Lim, J.; Jeon, Y.; and Choi, J. Y. 2021. Influence-balanced loss for imbalanced visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 735–744.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1944–1952.

Qin, C.; Wang, Y.; and Fu, Y. 2022. Robust Semi-supervised Domain Adaptation against Noisy Labels. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4409–4413.

Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33: 4175–4186.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Salakhutdinov, R.; Torralba, A.; and Tenenbaum, J. 2011. Learning to share visual appearance for multiclass object detection. In *CVPR 2011*, 1481–1488. IEEE.

Scott, C. 2015. A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels. In *AISTATS*.

Scott, C.; Blanchard, G.; Handy, G.; Pozzi, S.; and Flaska, M. 2013. Classification with Asymmetric Label Noise: Consistency and Maximal Denoising. In *COLT*, 489–511.

Song, H.; Kim, M.; and Lee, J.-G. 2019. SELFIE: Refurbishing Unclean Samples for Robust Deep Learning. In *International Conference on Machine Learning*, 5907–5915.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 3319–3328. PMLR.

Ustun, B.; Liu, Y.; and Parkes, D. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, 6373–6382. PMLR.

Wang, J.; Liu, Y.; and Levy, C. 2021. Fair Classification with Group-Dependent Label Noise. FAccT, 526–536. New York, NY, USA.

Wang, J.; Wang, X. E.; and Liu, Y. 2022. Understanding instance-level impact of fairness constraints. In *International Conference on Machine Learning*, 23114–23130. PMLR.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13726–13735.

Wei, H.; Tao, L.; Xie, R.; and An, B. 2021a. Open-set label noise can improve robustness against inherent label noise. *Advances in Neural Information Processing Systems*, 34.

Wei, H.; Xie, R.; Cheng, H.; Feng, L.; An, B.; and Li, Y. 2022a. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, 23631–23644. PMLR.

Wei, H.; Zhuang, H.; Xie, R.; Feng, L.; Niu, G.; An, B.; and Li, Y. 2022b. Logit Clipping for Robust Learning against Label Noise. *arXiv preprint arXiv:2212.04055*.

Wei, J.; Liu, H.; Liu, T.; Niu, G.; and Liu, Y. 2021b. Understanding Generalized Label Smoothing when Learning with Noisy Labels. *arXiv preprint arXiv:2106.04149*.

Wei, J.; and Liu, Y. 2020. When Optimizing f-Divergence is Robust with Label Noise. In *International Conference on Learning Representations*.

Wei, J.; Narasimhan, H.; Amid, E.; Chu, W.-S.; Liu, Y.; and Kumar, A. 2023. Distributionally Robust Post-hoc Classifiers under Prior Shifts. In *The Eleventh International Conference on Learning Representations*.

Wei, J.; Zhu, Z.; Cheng, H.; Liu, T.; Niu, G.; and Liu, Y. 2022c. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. In *International Conference on Learning Representations*.

Wei, J.; Zhu, Z.; Luo, T.; Amid, E.; Kumar, A.; and Liu, Y. 2022d. To aggregate or not? learning with separate noisy labels. *arXiv preprint arXiv:2206.07181*.

Wei, T.; Shi, J.-X.; Li, Y.-F.; and Zhang, M.-L. 2022e. Prototypical classifier for robust class-imbalanced learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 44–57. Springer.

Wei, T.; Shi, J.-X.; Tu, W.-W.; and Li, Y.-F. 2021c. Robust long-tailed learning under label noise. *arXiv preprint arXiv:2108.11569*.

Xia, X.; Han, B.; Wang, N.; Deng, J.; Li, J.; Mao, Y.; and Liu, T. 2022a. Extended<? TeX T?>: Learning with Mixed Closed-set and Open-set Noisy Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xia, X.; Liu, T.; Han, B.; Gong, C.; Wang, N.; Ge, Z.; and Chang, Y. 2021a. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*.

Xia, X.; Liu, T.; Han, B.; Gong, M.; Yu, J.; Niu, G.; and Sugiyama, M. 2021b. Instance correction for learning with open-set noisy labels. *arXiv preprint arXiv:2106.00455*.

Xia, X.; Liu, T.; Han, B.; Gong, M.; Yu, J.; Niu, G.; and Sugiyama, M. 2022b. Sample Selection with Uncertainty of Losses for Learning with Noisy Labels. In *International Conference on Learning Representations*.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2691–2699.

Zhang, J.; Xu, X.; Han, B.; Liu, T.; Cui, L.; Niu, G.; and Sugiyama, M. 2022a. NoiLin: Improving adversarial training and correcting stereotype of noisy labels.

Zhang, M.; Yuan, C.; Yao, J.; and Huang, W. 2022b. Learning with noisily-labeled class-imbalanced data. *arXiv preprint arXiv:2211.10955*.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2023. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, H.; and Gordon, G. 2019. Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 32.

Zhong, Y.; Deng, W.; Wang, M.; Hu, J.; Peng, J.; Tao, X.; and Huang, Y. 2019. Unequal-training for deep face recognition with long-tailed noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7812–7821.

Zhu, X.; Anguelov, D.; and Ramanan, D. 2014. Capturing long-tail distributions of object subcategories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 915–922.

Zhu, Z.; Liu, T.; and Liu, Y. 2021. A Second-Order Approach to Learning with Instance-Dependent Label Noise. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, Z.; Luo, T.; and Liu, Y. 2022. The Rich Get Richer: Disparate Impact of Semi-Supervised Learning. In *International Conference on Learning Representations*.

Zhu, Z.; Song, Y.; and Liu, Y. 2021. Clusterability as an alternative to anchor points when learning with noisy labels. In *International Conference on Machine Learning*, 12912–12923. PMLR.

Zhu, Z.; Wang, J.; and Liu, Y. 2022. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In *International Conference on Machine Learning*, 27633–27653. PMLR.

Reproducibility Checklist

Instructions for Authors:
This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) Partial

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) Yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) Yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) Yes

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Yes

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Yes

2.4. Proofs of all novel claims are included (yes/partial/no) Yes

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Yes

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Yes

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Yes

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Yes

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) Yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) Yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) Yes

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) Yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) Yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) Yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) Yes

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) Yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) Yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no, we will open-source our code once the paper got accepted.

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) no, we will open-source our code once the paper got accepted.

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) Yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) Yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) Yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) Yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) Yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) Yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) Yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) partial

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) Yes