Robust Logit Adjustment for Learning with Long-Tailed Noisy Data

MingCai Chen[1,2*], Yuntao Du[3], Wenyu Jiang[4], Baoming Zhang[4], Shuai Feng[4], Yi Xin[4], Chongjun Wang[4*]

[1]Nanjing University of Posts and Telecommunications
[2]The State Key Laboratory of Tibetan Intelligence
[3]State Key Laboratory of General Artificial Intelligence, BIGA
[4]State Key Laboratory for Novel Software Technology at Nanjing University, Nanjing University
chenmc@njupt.edu.cn, duyuntao@bigai.ai, {lygjwy,zhangbm,shuaifeng,xinyi}@smail.nju.edu.cn, chjwang@nju.edu.cn

Abstract

Learning with noisy labels (LNL) methods have enabled the deployment of machine learning systems with imperfectly labeled data. However, these methods often struggle to identify noise in the presence of long-tailed (LT) class distributions, where the memorization effect becomes class-dependent. Conversely, LT methods are suboptimal under label noise, as it hinders access to accurate label frequency statistics. This study aims to address the long-tailed noisy data by bridging the methodological gap between LNL and LT approaches. We propose a direct solution, termed *Robust Logit Adjustment*, which estimates ground-truth labels through label refurbishment, thereby mitigating the impact of label noise. Simultaneously, our method incorporates the distribution of training-time corrected target labels into the LT method logit adjustment, providing class-rebalanced supervision. Extensive experiments on both synthetic and real-world long-tailed noisy datasets demonstrate the superior performance of our method.

Introduction

The training of deep models heavily relies on high-quality labeled datasets. However, manual data annotation is time-consuming, labor-intensive, error-prone, and subject to domain knowledge or privacy restrictions (Roh, Heo, and Whang 2021; Van Engelen and Hoos 2020; Algan and Ulusoy 2021). Moreover, billion-level data has been utilized in training recent large-scale vision and multi-modal models (Oquab et al. 2024; Li et al. 2022). Manually annotating such massive volumes of data is nearly impossible. In the absence of high-quality and large-scale annotated datasets, weakly supervised data, e.g., noisily labeled data, is readily available. For example, Li et al. (2022, 2023) constructed a dataset containing millions of samples using noisy text-image pairs from web pages and successfully trained the large-scale multi-modal vision-language model BLIP. Therefore, effectively leveraging noisy data becomes crucial when acquiring high-quality annotated data is challenging.

Recently, significant research has been conducted in the field of learning with noisy labels (LNL) (Han et al. 2020; Song et al. 2020). One category of methods that stands out is those that leverage the memorization effect (Arpit et al. 2017; Li, Soltanolkotabi, and Oymak 2020) to identify noise. The memorization effect refers to the phenomenon where deep networks learn from correctly labeled samples with common patterns faster than mislabeled samples with unique mapping relationships. Consequently, the model's fitting status on individual samples can serve as an indicator to detect mislabeled ones. For example, DivideMix (Li, Socher, and Hoi 2020) removes labels from samples with higher loss values and only retains possible clean supervision signals. Although such LNL methods have demonstrated superior performance on class-balanced datasets, they have not been thoroughly evaluated on more realistic data, which often exhibit a long-tailed (LT) distribution characterized by a few classes containing the majority of samples. When the trained models biased toward majority classes, LNL methods relying on the memorization effect would fail. As depicted in Fig. 1 (a), the model tends to exhibit higher loss values on both noisy data and clean tail classes. Consequently, LNL methods that remove the given labels for data with higher loss values would have poor generalization ability, especially on tail classes. On the other hand, directly adding LT techniques on top of LNL methods is suboptimal, as most of them depend on the true label frequencies. For example, re-weighting methods (Cui et al. 2019; Park et al. 2021) balance the proportion of different classes in the training objective by assigning different weights to the losses of head and tail classes, while re-sampling (Kang et al. 2020; Chawla et al. 2002) methods ensure that each class has an equal probability of being sampled. However, as shown in Fig. 1 (b), there is a discrepancy between the frequencies of observed labels and ground-truth labels under the coexistence of realistic label noise and class imbalance. The incorrect label frequency hinders LT techniques in achieving class balance. In conclusion, previous LNL and LT methods both fail to function properly under the disturbance of long-tailed noisy data.

In this work, we propose that the label correction framework can bridge the LNL and LT problems: Firstly, it allows the model to effectively mitigate the disturbance of label noise by learning from more accurate supervision signals. Secondly, the corrected labels provide a more reliable basis for label frequency statistics, thereby enabling the proper functioning of class re-balancing techniques. To ac-

*Corresponding author.

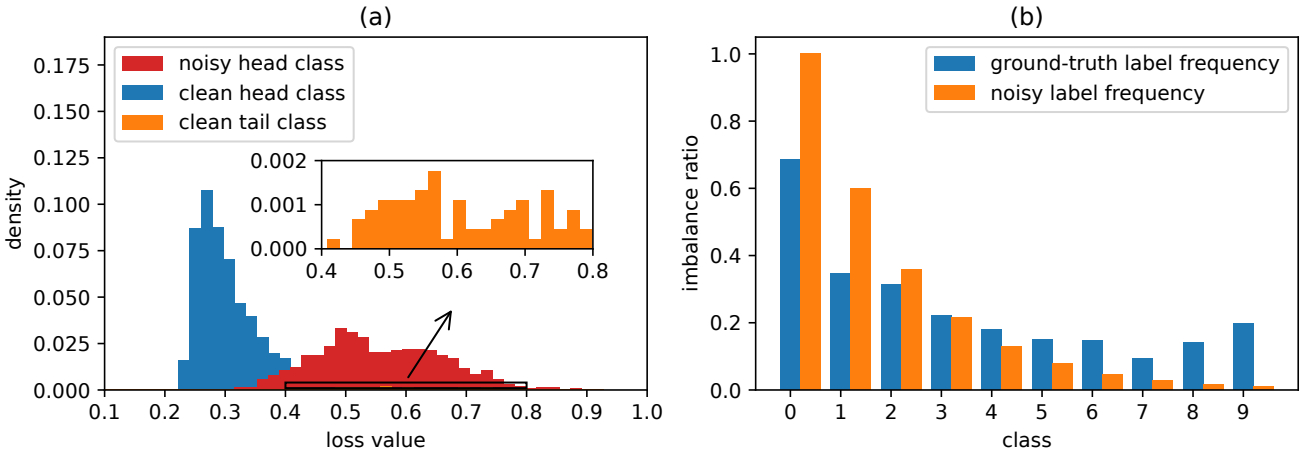

Figure 1: (a) Trained on the synthetic noisy long-tailed CIFAR-10 dataset, the distributions of model's loss value on: noisy data and clean data from the majority class (head class), and clean data from the minority class (tail class). The model has undergone a short training using simple empirical risk minimization. (b) On the class-imbalanced CIFAR with real-world noise (Wei et al. 2022), there is a mismatch of label frequencies between the observed label and the ground-truth label.

complish this, our proposed method initially performs label refurbishment (Reed et al. 2014) based on the noisy data and the model's perception. Specifically, it obtains corrected labels by leveraging two sources of supervision: the given noisy labels and model-generated pseudo-labels. In a convex form, the clean label probabilities associated with individual samples are dynamically estimated to determine the optimal weight of the two supervisory signals. Importantly, the corrected labels serve not only as clean supervision to combat label noise, but also play a crucial role in estimating the label frequencies. We utilize the corrected labels to online-update the label frequency statistics. These statistics serve as the label-dependent offsets to the logit adjustment technique, which encourages a large relative margin between a pair of minority and majority labels (Menon et al. 2021). By adopting this approach, we enable the model to receive robust and class-balanced supervision even in the absence of ground-truth labels. The contributions and novelties of this work are summarized as follows:

- We analyze traditional methodologies that address noisy labels or class imbalance in their realistic yet underexplored conjunction - learning from long-tailed noisy data - and identify the causes of suboptimal performance.
- We propose a novel method called *Robust Logit Adjustment*. It focuses on the core connection of two problems – the absence of ground-truth labels, for which we compensate by performing label refurbishment. In the next, robust and class-balanced supervision signals can be obtained directly based on the corrected labels and online-updated logit adjustment.
- To validate the effectiveness of our proposed method, we conducted extensive experiments with various noise and imbalance rates. Our method demonstrated significant improvements, achieving up to a 13% accuracy improvement on the noisy long-tailed CIFAR dataset and up to a 1.6% accuracy improvement on real-world noisy datasets with class imbalance, namely long-tailed Animal-10N and Food-101N. We also conduct systematic ablation analysis that leads to an improved understanding of our approach.

Problem Setup and Related Work

Notations and Preliminaries

In learning with noisy labels, we are given N training samples $(\boldsymbol{x}, \tilde{\boldsymbol{y}})$ i.i.d. sampled from a noisy distribution, where each training sample is a pair consists of the input feature $\boldsymbol{x} \in \mathcal{X}$ and the noisy labels $\tilde{\boldsymbol{y}} \in \Delta^{C-1}$, where Δ^{C-1} is the probability simplex (For convenience, we also use $\tilde{y} \in [C]$ represents the corresponding categorical label). The goal of learning with noisy labels is to robustly train a parametric model $f(\boldsymbol{x}; \boldsymbol{\theta}) : \mathcal{X} \mapsto \Delta^{C-1}$, while mitigating the impact of label noise in the training samples. The performance is evaluated on the test samples $(\boldsymbol{x}, \boldsymbol{y})$ sampled from the training-time-inaccessible clean distribution. Learning with long-tailed noisy data (LNL-LT) further extends this setting by considering the existence of class imbalance, where the number of samples for tail classes is significantly smaller than that of head classes.

Learning with Noisy Labels

Label noise is a common occurrence in real-world datasets, and previous research has focused on addressing different types of noise. These noise models are based on three main assumptions: uniform label noise, class-dependent noise, and instance-dependent noise. Uniform noise assumes that the labels are uniformly corrupted and assigned to other classes (Manwani and Sastry 2013). Theoretical studies have shown that certain loss functions can achieve optimal solutions on uniformly noisy datasets as on clean datasets (Manwani and Sastry 2013). Additionally, commonly used regularization techniques such as Dropout (Srivastava et al.

2014), weight decay (Krogh and Hertz 1991), and the intrinsic robustness of deep networks (Zhang et al. 2016) can help mitigate the impact of this type of noise. The class-dependent noise assumption states that the noisy label is conditionally independent of data features when the true labels is given, i.e., $\mathrm{p}(\tilde{y}|y, \boldsymbol{x}) = \mathrm{p}(\tilde{y}|y)$. Several methods have been developed to adapt to this noise corruption. For example, in Chen and Gupta (2015), a linear adaptation layer was inserted between the base network and the cross-entropy loss calculation. This layer modifies the classification model using a parameterized label transition matrix, aligning the predicted results with the label distribution of the noisy data. Instance-dependent label noise, i.e., the corruption probability is assumed to be dependent on both the data features and class labels, is more challenging to directly model. Therefore, researchers have leveraged the inherent noise tolerance of deep neural networks (Arpit et al. 2017). Recent LNL methods, such as DivideMix (Li, Socher, and Hoi 2020), are typically based on the memorization effect. This assumption suggests that the model first learns the clean labels with a common mapping relationship and then memorizes the incorrect labels. For instance, the Co-teaching algorithm (Han et al. 2018) maintains two models during training, with one model selecting samples with smaller losses for parameter updates. However, long-tailed class distributions would render most of these LNL algorithms unable to discover or correct noise from the behavior of the class-imbalanced model, making them vulnerable to long-tailed noisy data.

Long-Tailed Learning

Real-world data usually have long-tailed class imbalance. Class re-balancing is a mainstream method used to address this issue. For instance, re-sampling methods (Liu, Wu, and Zhou 2009) adjust the number of samples per class used in each training iteration to ensure class balance. In our work, we also employ logit adjustment to achieve class rebalancing from the perspective of loss calculation. This approach involves adjusting the model's logits based on the label frequencies:

$$\mathcal{L}_{\mathrm{LA}}(\boldsymbol{x}, y=j) = -\log \frac{\exp(f_j(\boldsymbol{x}) + \log \mathrm{p}(y=j))}{\sum_i \exp(f_i(\boldsymbol{x}) + \log \mathrm{p}(y=i))}. \tag{1}$$

The effectiveness of logit adjustment has been theoretically demonstrated to minimize the average per-class error in a Fisher-consistent manner (Menon et al. 2021). Additionally, data augmentation serves as a powerful technique for long-tailed learning. One approach is SMOTE (Liu, Wu, and Zhou 2009), which generates more samples for tail classes by mixing intra-class neighbors. Remix (Chou et al. 2020), building on the Mixup (Zhang et al. 2017) approach, assigns the label in favor of the minority class after mixing the features of two randomly selected samples. In recent years, robust long-tailed learning has emerged as an important topic (Zhang et al. 2022a; Wei et al. 2021). For example, open-set long-tailed learning (Liu et al. 2019) aims to optimize performance across a dataset comprising head, tail, and open classes, with the goal of reducing confusion between tail and open classes. Moreover, Zhang et al. (2022b) relax the original assumption by targeting at unknown class distributions in the test set. However, the problem of long-tailed data under label noise is underexplored, even though it easily appears in the real world.

Learning with Long-Tailed Noisy Data

Recently, a few works delve into the problem of learning with long-tailed noisy data (Yi et al. 2022; Zhang et al. 2022a; Wei et al. 2021), which aims to address the label noise while maintaining a balanced model across classes. This research direction places particular emphasis on the tail class, which is often more demanding to handle due to the limited availability of samples. For instance, Yi et al. (2022) initially train a noise identifier that remains unaffected by the long-tailed distribution, followed by the optimization of the classification model. (Wei et al. 2021) propose a prototype-based metric that achieves better noise detection for tail classes. They then employ semi-supervised learning to incorporate the noisy data for training. Another approach, SSBL (Zhang et al. 2022a), utilizes class-aware sample selection to distinguish clean samples from noisy ones, subsequently employing a balanced loss for robust training. Compared to previous LNL-LT methods that solve the issues of label noise and long-tailed class distribution separately, our work explores a unified solution by correcting the noisy labels with label refurbishment. This naturally enables robust training against label noise and the use of logit adjustment for class rebalancing. Besides, the proposed method is simple and effective, requiring only a single training target and minimal design or hyperparameter tuning. This allows for further improvement or customization using additional LNL-LT techniques to enhance performance.

Method

Training Framework

Our algorithm follows an iterative process 1) correcting noisy labels and label frequency statistics based on the model's perception, and 2) training the model using more robust and class-balanced supervision. The algorithm begins with a warm-up period, during which the model is trained briefly on the given noisy dataset following a standard supervised training schema. In each iteration of robust training, we correct the original noisy and imbalanced supervisions from two perspectives. First, we employ label refurbishment to handle label noise. This involves estimating the clean probabilities of the given noisy labels and performing a convex refurbishment using pseudo-labels to generate clean supervision. Second, the label frequency statistics are continuously updated using the training-time corrected target labels. The logit adjustment technique based on the updated label frequencies is then used to ensure a balance between head and tail classes.

Warm-up Training

As shown in the empirical study in Zhang et al. (2016); Arpit et al. (2017) and many LNL literatures (Arazo et al. 2019; Li, Socher, and Hoi 2020), the model would learn the commonly shared correct patterns within the dataset while hasn't

had time to overfit unique noise pattern. Therefore, we first warm up the model on the given noisy dataset following the mini-batch stochastic gradient descent supervised training schema. Specifically, mini-batches of data are randomly drawn from the noisy dataset for loss calculation, the loss of individual sample is:

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}, \tilde{y}=j) = -\log \frac{\exp(f_j(\boldsymbol{x}))}{\sum_i \exp(f_i(\boldsymbol{x}))}, \tag{2}$$

where $\tilde{y}$ represents the noisy label. After warm-up, the model acquires a certain level of generalization ability, and its perception, i.e., predicted classes and confidence on samples, become useful in the subsequent robust training stage.

Label Refurbishment

To correct the noisy labels, we employ the label refurbishment technique. It utilizes convex form to leverage adaptively estimated per-sample clean probabilities, effectively balancing the noisy supervision signals with the self-supervision signals to obtain the refurbished label $\boldsymbol{y}^*$:

$$\boldsymbol{y}^* = w\tilde{\boldsymbol{y}} + (1-w)\hat{\boldsymbol{y}}\,, \quad \text{where} \begin{array}{l} w = \text{clean probability;} \\ \tilde{\boldsymbol{y}} = \text{noisy label;} \\ \hat{\boldsymbol{y}} = \text{pseudo-label.} \end{array} \tag{3}$$

Eq. 3 represents a balance between the given noisy supervision and self-supervision. When the given label is likely to be correct, the clean probability w approaches 1, resulting in a target supervision that aligns with the given label. Conversely, when the given supervision is likely to be incorrect, the clean probability w approaches 0, causing the target supervision to lean towards the self-supervision. Note that all labels are probability vectors, i.e., $\boldsymbol{y}^*, \tilde{\boldsymbol{y}}, \hat{\boldsymbol{y}} \in \Delta^{C-1}$. The fine-grained targets help prevent potential overconfidence and improve model calibration. The generation of pseudo-labels $\hat{\boldsymbol{y}}$ and the estimation of clean probabilities w is described in the next section.

Pseudo-Labeling In scenarios where the given labels are noisy, incorporating self-training signals can be beneficial. Pseudo-labeling is a widely employed technique in weakly supervised learning domains, including learning with noisy labels (Huang, Zhang, and Zhang 2020) and semi-supervised learning (Sohn et al. 2020). In this study, we utilize the class distribution predictions obtained from the training-time model as pseudo-labels, denoted as:

$$\hat{\boldsymbol{y}} = f(\boldsymbol{x}; \boldsymbol{\theta}). \tag{4}$$

However, relying solely on pseudo-labels for training disregards the external supervision provided by the given labels. Therefore, it is beneficial to combine the pseudo-labels with the given labels to create the final refurbished labels.

Clean Probability Estimation The primary objective of learning with noisy data is to determine the correctness of given labels. In this study, we estimate the clean probability $\mathrm{p}(y = \tilde{y}|\boldsymbol{x})$ for each label, where y represents the inaccessible ground-truth label. To accomplish this, we resort to the "small-loss assumption", which directly utilizes the memorization effect to identify noise: Deep models tend to exhibit

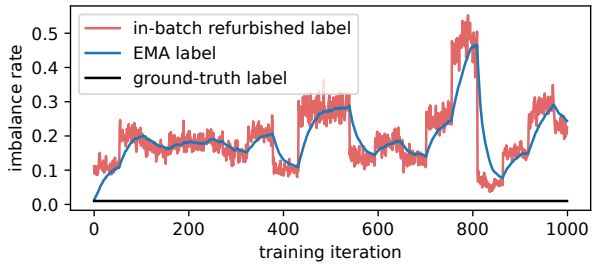

Figure 2: The change of the imbalance ratio of refurbished labels and in-batch target labels along training.

smaller losses on clean samples and larger losses on noisy samples (Arpit et al. 2017; Yao et al. 2020). This is because deep models initially learn simple correct patterns shared by the majority of samples. Guided by this assumption, we compute the per-sample loss values and cluster them into two groups. The cluster with the smaller mean loss value corresponds to the clean data and the cluster with the bigger mean loss value corresponds to the mislabeled data. Fig. 1 (a) illustrates the loss distribution of clean and noisy data, which can be approximated by two Gaussian distributions, respectively. To model the mixture distribution, we employ the probabilistic clustering method GMM (Gaussian Mixture Model). The optimization of GMM follows the standard expectation-maximization algorithm. We refer to our source code for detailed implementation. After fitting the GMM, it provides the clean probability by identifying the probabilities of samples belonging to the Gaussian distribution with the smaller mean loss value.

The clean probability estimation approach discussed above is effective for long-tailed class distributions only when the model can provide balanced predictions, as discussed in the Introduction Section. We address this issue by adopting an online-updated logit adjustment schema during training. Furthermore, we discovered that reweighting the loss values with respect to the inverse class frequency before modeling the distribution of loss using GMM leads to improved class-balanced results. We visualize the distribution of loss values and demonstrate the superiority of reweighted loss compared to the original loss in the supplementary material.

Robust Logit Adjustment

The original form of logit adjustment, as shown in Eq. 1, addresses the class imbalance problem under two assumptions. First, it assumes that all data are correctly labeled, and the label frequency is known. Second, it assumes that the training targets remain constant throughout the training process. However, in the case of LNL-LT, there is no access to the frequency of the ground-truth labels and the given label frequency is biased, as illustrated in Fig. 1(b). Furthermore, as depicted in Fig. 2, when a robust algorithm attempts to correct the training targets during training, the distribution of refurbished labels, along with the imbalance ratio, changes. Therefore, leveraging the distribution of the

current training target (i.e., the refurbished labels) to construct the logit adjustment can provide more class-balanced supervision. Additionally, considering that statistical measures obtained within a batch may fluctuate significantly due to the influence of sub-sampling in mini-batch gradient descent, incorporating previous statistics for smoothing can yield more stable results. Based on the above discussion, we first initialize the class distribution, denoted as $\boldsymbol{p}^*$, using the given noisy label frequency at the beginning of training:

$$\boldsymbol{p}^* = \frac{1}{|\tilde{\mathcal{Y}}|} \sum_{\tilde{\boldsymbol{y}} \in \tilde{\mathcal{Y}}} \tilde{\boldsymbol{y}}, \tag{5}$$

where we use $\tilde{\mathcal{Y}}$ to denote all noisy training labels for convenience. During training, we use the refurbished labels $\boldsymbol{y}^*$ to update $\boldsymbol{p}^*$ in every batch in an Exponential Moving Average (EMA) fashion, i.e.,

$$\boldsymbol{p}^* = \alpha \boldsymbol{p}^* + (1-\alpha) \frac{1}{|\mathcal{B}^*|} \sum_{\boldsymbol{y}^* \in \mathcal{B}^*} \boldsymbol{y}^*, \tag{6}$$

where α is the smoothing coefficient hyper-parameter, $\mathcal{B}^*$ consists of all the refurbished labels for a batch of samples. As shown in Fig. 2, the change in the imbalance ratio of our estimated label frequencies is smoother, and therefore provides a smoother supervision. It is worth noting that the estimated imbalance ratio is higher than the true imbalance ratio due to the used class rebalance technique, which encourages the model to make high-confidence predictions for the tail classes.

Final Training Target

After obtaining the refurbished label $\boldsymbol{y}^*$ and the robust label frequency $\boldsymbol{p}^*$, the final training target of our method includes a simple cross-entropy term with the refurbished label after logit adjustment:

$$\mathcal{L}_{\mathrm{RLA}} = -\frac{1}{C} \sum_j \boldsymbol{y}_j^* \log \frac{\exp(f_j(\mathcal{A}(\boldsymbol{x}))) + \log \boldsymbol{p}_j^*)}{\sum_i \exp(f_i(\mathcal{A}(\boldsymbol{x})) + \log \boldsymbol{p}_i^*)}, \tag{7}$$

where $\mathcal{A}$ denotes the augmentation function RandAugment (Cubuk et al. 2020; Zhang et al. 2022a), which is used in successful semi-supervised learning methods FixMatch (Sohn et al. 2020) to enhance generalization. For convenience, we use $\boldsymbol{y}_i^*$ or $\boldsymbol{p}_i^*$ to refer to the i-th element in the vectors $\boldsymbol{y}^*$ or $\boldsymbol{p}^*$. Our simple training target is effective without bells and whistles, which makes it ready to be combined with other techniques, e.g., the class-invariant noise identifier (Yi et al. 2022) or other self-supervised techniques (Karthik, Revaud, and Chidlovskii 2021), to further boost performance and facilitate customization.

Experiments

Experimental Details

Dataset Construction Our experiments are conducted on two types of datasets, one is obtained by corrupting the labels and create class imbalance on correctly labeled dataset CIFAR (Krizhevsky, Hinton et al. 2009). Another is obtained by creating class imbalance on the real-world noisy datasets, namely Animal-10N (Song, Kim, and Lee 2019) and Food-101N (Lee et al. 2018).

To construct the imbalanced noisy dataset, we follow (Yi et al. 2022; Karthik, Revaud, and Chidlovskii 2021; Cao et al. 2021) and decide the number of samples per class according to the exponential function:

$$N_c = N_{max} \frac{1}{\eta}^{\frac{c-1}{C-1}}, \quad c \in \{1, \ldots, C\}, \tag{8}$$

where η is the imbalance ratio, N_{max} is the number of the samples from the majority class. For synthetic label noise, we corrupt the data according using the noise transition matrix T:

$$T_{i,j} = \mathrm{p}(\tilde{y} = j | y = i) = \begin{cases} 1 - r & i = j \\ r \frac{N_j}{\sum_{k \neq i} N_k} & i \neq j \end{cases}, \tag{9}$$

where r is the noise rate. $T_{i,j}$ represents the probability of a sample in class i being corrupted into class j. In this way, the imbalance ratio is kept after the corruption and all classes have the same noise rate.

To create class imbalance on real-world noise dataset Animal-10N and Food101N, we randomly choose samples from each class following the imbalance from Eq. 8. For comparison, results of DivideMix-LA and DivideMix-DRW (Li, Socher, and Hoi 2020), Logit Adjustment (Menon et al. 2021), and SSBL_2 (Zhang et al. 2022a) are reported.

Compared Methods For comparison on the long-tailed noisy datasets, we report the results of:

- ERM (Empirical Risk Minimization) trained model.
- LNL methods: Co-teaching (Han et al. 2018), ELR+ (Liu et al. 2020), and DivideMix (Li, Socher, and Hoi 2020).
- LT method: Logit Adjustment (Menon et al. 2021).
- LNL-LT method: SSBL_2 (Zhang et al. 2022a), DivideMix-LA, and DivideMix-DRW. The latter two are the combinations of DivideMix with long-tailed learning Methods Logit Adjustment (Menon et al. 2021) and Deferred Re-Weighting (Cao et al. 2019), respectively.

The compared methods all use the same training protocol and model architecture. The experiments are conducted on long-tailed noisy datasets that are constructed using consistent methods, as in previous work (Menon et al. 2021; Cao et al. 2019). Both the best test accuracy across all epochs and the averaged test accuracy over the last 10 epochs are reported on CIFAR. Additional training details are provided in the supplementary material. It is important to note that the test datasets are class-balanced (For example, there are 10,000 test images evenly distributed across the 10 classes on CIFAR-10). Given the class-balanced nature of the test dataset, metrics such as accuracy are sufficient to evaluate whether the prediction results are biased towards the head classes. There is no need to rely on metrics like F1-score or AUC. It is also worth noting that the problem studied in this paper is different from class-dependent label noise (Chen and Gupta 2015). The latter focuses on a noise model where each sample of a class is corrupted with a certain probability to another class, without considering class imbalance.

Dataset		CIFAR-10						CIFAR-100					
Noise Rate		0.2			0.5			0.2			0.5		
Method/Imbalance Rate		0.1	0.02	0.01	0.1	0.02	0.01	0.1	0.02	0.01	0.1	0.02	0.01
ERM	Best	76.90	65.35	60.82	65.75	48.76	39.70	45.83	35.05	29.96	28.96	19.88	16.80
	Last	73.02	61.35	54.48	45.85	33.05	28.79	45.64	34.93	29.88	24.33	17.77	14.47
Co-teaching	Best	78.50	45.86	39.59	34.60	23.58	17.45	43.81	30.58	28.08	14.58	11.62	9.69
	Last	77.51	44.91	38.07	31.71	22.57	14.88	43.69	30.22	28.08	14.49	10.54	9.35
DivideMix	Best	91.03	83.07	70.08	85.97	69.73	52.06	63.08	49.76	43.71	55.98	41.79	35.03
	Last	90.71	82.16	69.82	85.79	68.19	51.72	62.54	49.45	42.84	55.56	41.25	34.51
DivideMix-DRW	Best	90.44	84.64	80.50	85.40	73.90	47.31	63.65	50.28	45.23	56.77	42.24	36.17
	Last	89.33	82.99	80.40	84.88	72.92	45.90	63.24	50.04	44.95	56.58	41.61	35.74
DivideMix-LA	Best	91.68	85.16	80.06	85.84	56.28	49.48	64.76	53.31	47.52	57.11	40.72	35.18
	Last	91.65	84.15	78.40	85.47	54.67	46.66	64.33	52.83	46.21	56.57	40.21	34.73
Logit Adjustment	Best	90.02	84.70	81.92	87.22	81.72	78.74	63.12	52.93	48.58	61.35	51.46	47.16
	Last	90.02	84.65	80.57	87.22	81.72	78.74	62.01	52.69	47.94	59.02	50.40	46.06
$SSBL_2$	Best	92.47	87.14	81.98	89.41	78.65	72.96	65.09	53.52	47.87	57.95	44.37	39.49
	Last	92.25	86.87	81.29	89.09	76.44	69.11	64.60	53.35	47.64	57.80	43.47	38.64
Robust Logit Adjustment	Best	**93.56**	**89.64**	**86.39**	**93.70**	**89.82**	**86.71**	**69.36**	**59.21**	**54.61**	**70.93**	**57.85**	**50.84**
	Last	**93.24**	**89.39**	**84.71**	**93.36**	**89.82**	**86.49**	**69.11**	**58.72**	**54.10**	**70.32**	**57.34**	**50.35**

Table 1: Comparison with state-of-the-art methods on long-tailed CIFAR with synthetic noise. The results of other methods are from Zhang et al. (2022a).

Comparison with SOTA Methods on Synthetic Long-Tailed Noisy Datasets

Based on the results presented in Tab. 1 conducted on long-tailed noisy CIFAR, we can draw the following conclusions: Our method achieves the best results across different noise rates and imbalance rates compared to LNL methods, LT methods, and LNL-LT methods. The improvements are significant especially under heavy noise and large imbalance rates. For instance, at a noise rate of 0.5 and an imbalance rate of 0.01, *Robust Logit Adjustment* achieves the best accuracy of 86.71%. This outperforms other methods, such as $SSBL_2$ (72.96%), by a significant margin. We also note that, with a fixed imbalance rate, the increase in noise rate does not necessarily lead to worse performance. For example, at an imbalance rate of 0.01, the accuracy of *Robust Logit Adjustment* remains high even at a noise rate of 0.5 (86.71% vs. 86.39%). We hypothesize that this is because the increase in synthetic noise rate can help mitigate the class imbalance. As a result, the performance remains stable or even improves despite the presence of more noise.

Comparison with SOTA Methods on Long-Tailed Real-World Noisy Datasets

We then conduct experiments on real-world noisy datasets with long-tailed class distributions. As shown in Tab. 2, our method achieves competitive results on the Animal-10N and Food-101N datasets, which have varying levels of class imbalance. For instance, on the Food-101N dataset, our method achieves accuracy values of 75.58%, 68.76%, and 63.88% for imbalance rates of 0.05, 0.02, and 0.01, respectively. These results indicate that our method performs well in handling long-tailed class distributions with real-world noise.

Dataset	Animal-10N			Food-101N		
Method/Imbalance Ratio	0.05	0.02	0.01	0.05	0.02	0.01
DivideMix-DRW	79.32	74.14	66.14	64.21	53.98	53.65
DivideMix-LA	78.62	74.90	66.08	71.48	63.73	59.01
Logit Adjustment	69.38	64.02	54.48	64.21	48.69	46.39
$SSBL_2$	**81.20**	**75.84**	69.36	74.39	**69.06**	63.67
Robust LA	79.14	75.28	**71.00**	**75.58**	68.76	**63.88**

Table 2: Comparison with state-of-the-art methods on long-tailed Animal-10N and Food-101N. The results of other methods are from Zhang et al. (2022a).

While our method may not achieve the best performance compared to other recent methods across all imbalance ratios, it only consists of simple techniques to achieve competitive results. Additionally, it requires minimal hyperparameter tuning. This simplicity brings better versatility and expansibility, making it readily combinable with other noisy label learning and long-tailed learning techniques for potentially improved performance.

Ablation Study

In order to study the effects of removing components of our method, we conduct the ablation study on long-tailed noisy CIFAR-10 and summarize the results in Tab. 3.

To study the effect of logit adjustment for class rebalance, we remove it and only use cross-entropy for training. It significantly impacts the performance, especially under larger class imbalance. The results show that when logit adjustment is omitted, the performance consistently drops across different imbalance rates. For instance, under a noise rate of 0.5 and imbalance rate of 0.01, the accuracy decreases from

Noise Rate		0.2			0.5		
Method/Imbalance Rate		0.1	0.02	0.01	0.1	0.02	0.01
Robust Logit Adjustment	Best	93.56	89.64	86.39	93.70	89.82	86.71
	Last	93.24	89.39	84.71	93.34	89.82	86.49
w/o logit adjustment	Best	93.36	84.66	78.47	92.39	81.85	72.96
	Last	93.36	82.73	76.27	92.39	81.19	71.85
w/o online update	Best	93.13	71.29	64.52	79.10	53.88	44.30
	Last	82.09	55.83	38.82	54.58	30.59	21.60

Table 3: Ablation study on the long-tailed noisy CIFAR-10.

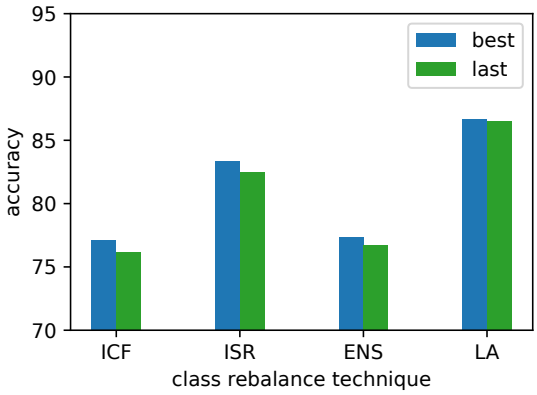

Figure 3: The performance of different class rebalancing techniques.

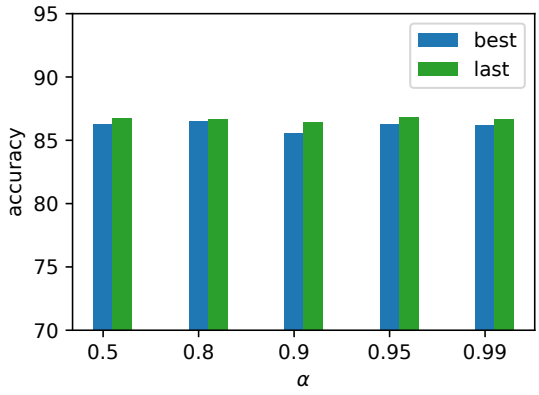

Figure 4: Ablation study of the hyper-parameter α for online update of estimated class distribution.

86.71% to 72.96%. This demonstrates the effectiveness of logit adjustment in handling class imbalance.

To study the effect of updating the estimated class distribution, we remove the update schema and only use the noisy labels' frequency for logit adjustment. Without online update, i.e., not modifying the label-dependent offsets throughout the training process, the performance consistently drops across different noise and imbalance rates. For example, under a noise rate of 0.5 and imbalance rate of 0.01, the accuracy decreases to 44.30%. This suggests that updating the scales of logit adjustment to provide class-balanced supervision within each training iteration is crucial for achieving better performance.

We also conducted an ablation study on the class-rebalance scheme as shown in Fig. 4. The purpose was to compare different methods of addressing class imbalance and evaluate their effectiveness. We explored three other different class-rebalance schemes other than LA (Logit Adjustment):

- ICF: Inverse class frequency assigns higher weights to minority classes and lower weights to majority classes based on their inverse frequencies, i.e., $\text{weight}(c) = \frac{1}{\text{N}_\text{c}}$.
- ISR: Based on the inverse class frequency, inverse class frequency puts a square root on the class frequency, i.e., $\text{weight}(c) = \frac{1}{\sqrt{N_c}}$.
- ENS: Effective number of samples (Cui et al. 2019) reweight the class via $\text{weight}(c) = \frac{(1-\beta^{N_c})}{(1-\beta)}$, where β is a hyper-parameter is set to 0.999 following the originnal paper.

The results of the ablation study indicate that logit adjustment is the most effective LT method. Furthermore, we also examined the sensitivity to the choice of the hyperparameter α in Fig. 3, which is used in the update of label frequency statistics. We varied the value of alpha and evaluated the performance of our method. the results show that our method is not sensitive to the choice of α.

Conclusion

Learning with long-tailed noisy data is a practical yet understudied problem. We propose and report on a simple method based on the understanding of the core connection between two problems: Label refurbishment provides a direct solution to label noise, and the online-updated logit adjustment serves as a natural source for balancing the class. Our method achieves new state-of-the-art robustness in the presence of label noise and class imbalance. Our future work is to dig into the conditional class imbalance distribution in the class-dependent or instance-dependent noise model and hopefully provide more realistic noise-robust algorithms.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (Grant No. 62192783, 62376117), the National Social Science Fund of China (Grant No. 23BJL035), the Science and Technology Major Project of Nanjing (comprehensive category) (Grant No. 202309007), the Collaborative Innovation Center of Novel Software Technology and Industrialization at Nanjing University and Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY224061).

References

Algan, G.; and Ulusoy, I. 2021. Image Classification with Deep Learning in the Presence of Noisy Labels: A Survey. *Knowledge-Based Systems*, 215: 106771.

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N.; and McGuinness, K. 2019. Unsupervised Label Noise Modeling and Loss Correction. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 312–321. PMLR.

Arpit, D.; Jastrzębski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; and Lacoste-Julien, S. 2017. A Closer Look at Memorization in Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 233–242.

Cao, K.; Chen, Y.; Lu, J.; Aréchiga, N.; Gaidon, A.; and Ma, T. 2021. Heteroskedastic and Imbalanced Deep Learning with Adaptive Regularization. In *Proceedings of 9th International Conference on Learning Representations*.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in the 33rd International Conference on Neural Information Processing Systems*.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.*, 16: 321–357.

Chen, X.; and Gupta, A. 2015. Webly Supervised Learning of Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 1431–1439.

Chou, H.-P.; Chang, S.-C.; Pan, J.-Y.; Wei, W.; and Juan, D.-C. 2020. Remix: Rebalanced Mixup. In *European Conference on Computer Vision*, 95–110. Springer.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Advances in Neural Information Processing Systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 702–703.

Cui, Y.; Jia, M.; Lin, T.; Song, Y.; and Belongie, S. J. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 9268–9277. Computer Vision Foundation / IEEE.

Han, B.; Yao, Q.; Liu, T.; Niu, G.; Tsang, I. W.; Kwok, J. T.; and Sugiyama, M. 2020. A Survey of Label-Noise Representation Learning: Past, Present and Future. *arXiv Preprint arXiv:2011.04406*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; and Sugiyama, M. 2018. Co-Teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In *Advances in the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 8536–8546.

Huang, L.; Zhang, C.; and Zhang, H. 2020. Self-Adaptive Training: Beyond Empirical Risk Minimization. In *Advances in Neural Information Processing Systems*, volume 33.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling Representation and Classifier for Long-Tailed Recognition. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Karthik, S.; Revaud, J.; and Chidlovskii, B. 2021. Learning From Long-Tailed Data with Noisy Labels. *arXiv Preprint arXiv:2108.11096*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning Multiple Layers of Features From Tiny Images. Technical report, University of Toronto.

Krogh, A.; and Hertz, J. A. 1991. A Simple Weight Decay Can Improve Generalization. In *Advances in Neural Information Processing Systems*, NIPS'91, 950–957. ISBN 1558602224.

Lee, K.; 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, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022. BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. In *Proceedings of International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900.

Li, J.; Socher, R.; and Hoi, S. C. H. 2020. DivideMix: Learning with Noisy Labels as Semi-Supervised Learning. In *Proceedings of the 8th International Conference on Learning Representations*.

Li, M.; Soltanolkotabi, M.; and Oymak, S. 2020. Gradient Descent with Early Stopping Is Provably Robust to Label Noise for Overparameterized Neural Networks. In *Proceedings of the The 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 4313–4324.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-Learning Regularization Prevents Memorization of Noisy Labels. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA. ISBN 9781713829546.

Liu, X.; Wu, J.; and Zhou, Z. 2009. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Trans. Syst. Man Cybern. Part B*, 39(2): 539–550.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-Scale Long-Tailed Recognition in an Open World. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2537–2546. Computer Vision Foundation / IEEE.

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

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-Tail Learning via Logit Adjustment. In *Proceedings of the 9th International Conference on Learning Representations*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; HAZIZA, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features Without Supervision. *Transactions on Machine Learning Research*.

Park, S.; Lim, J.; Jeon, Y.; and Choi, J. Y. 2021. Influence-Balanced Loss for Imbalanced Visual Classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 715–724. IEEE.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training Deep Neural Networks on Noisy Labels with Bootstrapping. *arXiv Preprint arXiv:1412.6596*.

Roh, Y.; Heo, G.; and Whang, S. E. 2021. A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4): 1328–1347.

Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *Advances in the 34th International Conference on Neural Information Processing Systems*, NIPS'20. ISBN 9781713829546.

Song, H.; Kim, M.; and Lee, J.-G. 2019. Selfie: Refurbishing Unclean Samples for Robust Deep Learning. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 5907–5915.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2020. Learning From Noisy Labels with Deep Neural Networks: A Survey. *arXiv Preprint arXiv:2007.08199*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *The Journal of Machine Learning Research*, 15(1): 1929–1958.

Van Engelen, J. E.; and Hoos, H. H. 2020. A Survey on Semi-Supervised Learning. *Machine Learning*, 109(2): 373–440.

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

Wei, T.; Shi, J.-X.; Tu, W.-W.; and Li, Y.-F. 2021. Robust Long-Tailed Learning Under Label Noise. *arXiv Preprint arXiv:2108.11569*.

Yao, Q.; Yang, H.; Han, B.; Niu, G.; and Kwok, J. T. 2020. Searching to Exploit Memorization Effect in Learning with Noisy Labels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 10789–10798.

Yi, X.; Tang, K.; Hua, X.-S.; Lim, J.-H.; and Zhang, H. 2022. Identifying Hard Noise in Long-Tailed Sample Distribution. In *Proceedings of the European Conference on Computer Vision*, 739–756. Springer.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2016. Understanding Deep Learning Requires Rethinking Generalization. *arXiv Preprint arXiv:1611.03530*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. Mixup: Beyond Empirical Risk Minimization. *arXiv Preprint arXiv:1710.09412*.

Zhang, L.; Tian, Z.-H.; Zhou, W.; and Wang, W. 2022a. Learning From Long-Tailed Noisy Data with Sample Selection and Balanced Loss. *arXiv Preprint arXiv:2211.10906*.

Zhang, Y.; Hooi, B.; Hong, L.; and Feng, J. 2022b. Self-Supervised Aggregation of Diverse Experts for Test-Agnostic Long-Tailed Recognition. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.