Prototype-Anchored Learning for Learning with Imperfect Annotations

Xiong Zhou [1] **Xianming Liu** [1 2] **Deming Zhai** [1] **Junjun Jiang** [1 2] **Xin Gao** [3 2] **Xiangyang Ji** [4]

Abstract

The success of deep neural networks greatly relies on the availability of large amounts of high-quality annotated data, which however are difficult or expensive to obtain. The resulting labels may be class imbalanced, noisy or human biased. It is challenging to learn unbiased classification models from imperfectly annotated datasets, on which we usually suffer from overfitting or underfitting. In this work, we thoroughly investigate the popular softmax loss and margin-based loss, and offer a feasible approach to tighten the generalization error bound by maximizing the minimal sample margin. We further derive the optimality condition for this purpose, which indicates how the class prototypes should be anchored. Motivated by theoretical analysis, we propose a simple yet effective method, namely prototype-anchored learning (PAL), which can be easily incorporated into various learning-based classification schemes to handle imperfect annotation. We verify the effectiveness of PAL on class-imbalanced learning and noise-tolerant learning by extensive experiments on synthetic and real-world datasets.

1. Introduction

Over the past few years, deep neural networks have achieved impressive performance in various tasks of computer vision (Goodfellow et al., 2016), such as image classification, segmentation, object detection, etc. One critical factor that attributes to the success of deep learning is the availability of large amounts of annotated training data. However, in many real-world scenarios, it would be difficult to attain perfect supervision for fully supervised learning due to a variety of limitations, such as inaccurate human annotation (Han et al., 2020), difficulty in collecting enough samples for all classes (Zhang et al., 2021c), etc. In this work, we specially consider the task of image classification under two imperfect annotation scenarios that are commonly encountered in practical applications:

Imbalanced Annotation. Real-world data usually exhibits highly-skewed class distribution, due to the nature that some classes are easy to collect sufficient examples while many classes are associated with only a few ones (Zhang et al., 2021c). For instance, in plant disease classification, it is comparatively easier to obtain images for common diseases than rare diseases during data collection. This results in that a few majority classes contain most of the data while massive minority classes only contain a scarce amount of instances. This issue would make naive learning biased towards the majority classes while with poor accuracy on the minority ones, posing a challenge for generalization.

Inaccurate Annotation. Data annotation tasks that require a high level of competency make the acquisition of hand-labeled data time-consuming and expensive. One common and economical way to collect large amounts of training data is through online queries (Liu et al., 2011) or crowdsourcing (Arpit et al., 2017), which however inevitably introduce noisy labels due to that no domain expertise is involved and a variety of human biases exist. This issue would lead deep neural networks to overfit mislabeled data, which seriously hampers the generalization ability of neural networks.

Learning unbiased classification models from imperfectly annotated datasets is a challenging problem. Specifically, for learning with imbalanced annotation, the main concern is how to prevent overfitting to the majority classes and achieve a better trade-off between the majority and minority classes. The difficulty lies in that the spanned feature space of the majority classes is spread while that of the minority classes is concentrated, which hampers the generalization of the learned classifier boundaries due to the lack of intra-class compactness and inter-class separation, as shown in Fig. 1(b). For learning with noisy labels, the key is to prevent overfitting to mislabeled data, as well as achieve a good trade-off between robustness and the fitting ability. The softmax loss enjoys the advantage of fitting ability due to that the optimizer puts more emphasis on ambiguous samples, which however suffers from the overfitting effect

[1]Harbin Institute of Technology, Harbin, China [2]Peng Cheng Laboratory, Shenzhen, China [3]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia [4]Tsinghua University, Beijing, China. Correspondence to: Xianming Liu <csxm@hit.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

on noisy labels. The symmetric losses, such as MAE (Ghosh et al., 2017) and normalized losses (Ma et al., 2020), have theoretical guarantees to be noise-tolerant, which however usually suffer from underfitting on complicated datasets.

In this work, we thoroughly investigate the softmax loss and its variant—margin-based losses, and claim that to tighten the generalization error bound one can maximize the minimal sample margin $\gamma_{\min}$ in Sec. 3.1.2. We further derive the optimality condition of prototypes to obtain the largest $\gamma_{\min}$, which indicates that the class prototypes $\boldsymbol{w}_1, ..., \boldsymbol{w}_k$ should satisfy $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}, \forall i \neq j$, where k is the class number. Motivated by the theoretical analysis, we propose a simple yet effective method, namely *prototype-anchored learning* (PAL). Specifically, we derive a classifier composed of anchored prototypes, which are predefined by a simple implementation that only requires knowing the number of classes and the feature dimension. Due to its simplicity, PAL can be easily incorporated into various learning-based classification schemes. For class-imbalanced learning, PAL can be used in the feature representation learning of decoupling methods (Kang et al., 2020; Zhang et al., 2021a), as well as in tandem with other performance-boosting approaches and modules. For noise-tolerant learning, the application of PAL is not such straightforward as class-imbalanced learning. Instead, we suggest the *feature normalized and prototype-anchored learning* (FNPAL) strategy, which can be combined with traditional losses to boost their ability in handling noisy labels. Extensive experiments are provided to demonstrate the superiority of PAL in handling imperfect annotations.

Our main contributions are highlighted as follows:

- We propose a theoretically sound, simple yet effective scheme—prototype-anchored learning, which can be easily embedded into various learning based classification under imperfect annotations.
- For class-imbalanced learning, we theoretically prove that PAL can implicitly guarantee balanced feature representations, and empirically verify that PAL can combine with existing methods to boost their performance significantly.
- For noise-tolerant learning, we extend the classical symmetric condition to a more general theorem and reveals that PAL can lead to a tighter bound than that without PAL. We provide extensive experimental results to demonstrate its superiority.

2. Related Works

2.1. Class-imbalanced Learning

A mainstream paradigm in class-imbalanced learning is class re-balancing, which seeks to balance the training sample numbers of different classes, such as re-sampling (Han et al., 2005; Liu et al., 2008), re-weighting (Huang et al., 2016; Ren et al., 2018; Cui et al., 2019; Byrd & Lipton, 2019; Shu et al., 2019), and logit adjustment (Ren et al., 2020; Hong et al., 2021; Menon et al., 2021). However, the main drawback of class re-balancing methods is that most of them usually improve tail-class performance while impairing head-class performance (Zhang et al., 2021d). Recently, two-stage approaches (Cao et al., 2019; Kang et al., 2020; Zhou et al., 2020a; Zhong et al., 2021) composed of representation learning and classifier learning have achieved significant improvement compared to one-stage methods. For the two-stage learning framework, Kang et al. (2020) and Zhou et al. (2020a) found that instance-balanced sampling gives the best and most generalizable representation while explicit class-balanced strategies may introduce adverse effects. However, balanced representations and decision boundaries cannot be guaranteed by instance-balanced sampling, while they being more balanced would improve the performance. For instance, Zhong et al. (2021) found that mixup can benefit the learning of minority classes and remedy over-confidence by balancing classifier weight norms. In this work, we theoretically propose an implicit re-balancing method to achieve balanced representations.

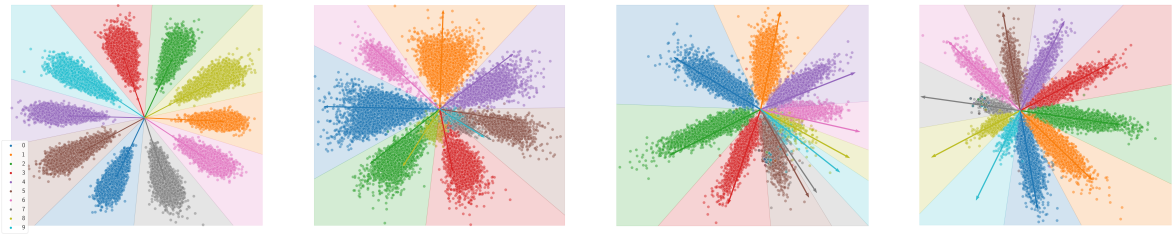
2.2. Noise-tolerant Learning

Noise-tolerant learning, or called learning with noisy labels, aims to train a robust model in the presence of noisy labels. To alleviate the impact of label noise, one popular research line is to design robust loss functions. This approach has been pursued in a large body of work (Long & Servedio, 2008; Wang et al., 2019a; Liu & Guo, 2020; Lyu & Tsang, 2020; Menon et al., 2020; Feng et al., 2020) that embraces new loss functions, especially symmetric losses and their variants (van Rooyen et al., 2015; Ghosh et al., 2017; Zhang & Sabuncu, 2018; Wang et al., 2019b; Ma et al., 2020; Zhou et al., 2021a). However, the fitting ability of the existing symmetric loss functions is restricted by the symmetric condition (Zhang & Sabuncu, 2018) such that other strategies to enhance the fitting ability are necessary (Charoenphakdee et al., 2019; Ma et al., 2020; Zhou et al., 2021b). In a sense, the existing methods of learning with noisy labels mainly seeks a trade-off between fitting ability and robustness.

3. Preliminaries and Motivation

3.1. Preliminaries

For multi-class classification problem, let $\mathcal{X} \subseteq \mathbb{R}^m$ denote the feature space and $\mathcal{Y} = \{1, \cdots, k\}$ denote the label space, we are usually given a labeled dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ to train models, where $(\boldsymbol{x}_i, y_i)$ are drawn from the joint distribution $\mathcal{D}$.



(a) (b) (c) (d)

Figure 1. Visualization on MNIST (a) and long-tailed MNIST (b-d) with imbalance ratio 0.1. The learned representations, class prototypes and decision regions are represented by points, vectors and areas in different colors, respectively. The corresponding color for each class is provided in (a), which is also used in (b-d). Specifically, (a) denotes the class-balanced case by CE, where features and prototypes are optimized to be perfectly balanced. (b) denotes the class-imbalanced case by CE, where the majority classes ("0-3") occupy most of the feature space, the representations of minority classes ("7-9") are narrow, and the majority classes have larger norms and angular distance from other prototypes, and the reverse on the minority classes. (c) denotes the class-imbalanced case by CE with feature and prototype normalization, where minority classes are squeezed into a small space, while the majority classes take up a lot, but the learned representations and decision regions are more separated clearly than (b). (d) denotes the class-imbalanced case by CE with proposed feature normalized and prototype-anchored learning, the decision regions and feature distribution are more balanced.

Softmax Loss. The most popular loss used for classification is the softmax loss, for $(\boldsymbol{x}_i, y_i)$ which is formulated as:

$$L_i = -\log \frac{\exp(\boldsymbol{w}_{y_i}^\top \boldsymbol{z}_i)}{\sum_{j=1}^{k} \exp(\boldsymbol{w}_j^\top \boldsymbol{z}_i)}, \tag{1}$$

where $\boldsymbol{z}_i = \boldsymbol{\phi}_\Theta(\boldsymbol{x}_i) \in \mathbb{R}^d$ is the learned feature representation, $\boldsymbol{\phi}_\Theta$ denotes the feature extraction network with parameters Θ, and $W = (\boldsymbol{w}_1, ..., \boldsymbol{w}_k) \in \mathbb{R}^{d\times k}$ represents the classifier implemented with a linear layer. For brevity, we use "prototypes" to denote the weight vectors of classifier.

Intuitively, the softmax loss promotes the learned feature representation $\boldsymbol{z}_i$ to be close to the corresponding prototype $\boldsymbol{w}_{y_i}$ and apart away from other prototypes. However, it may lead to feature representations without enough discriminativeness by overly enlarging the norm $\|\boldsymbol{z}_i\|_2$ or $\|\boldsymbol{w}_j\|_2$, especially when imperfect annotation exists. For instance, when $\mathbb{P}(y)$ is highly skewed, *i.e.*, imbalanced distribution, the majority classes occupy most of the feature space while the minority classes are narrow, as illustrated in Fig. 1(b). An approach to remedy this effect is performing normalization on both feature vectors and prototypes (Wang et al., 2017), *i.e.*, restricting them on the unit sphere $\mathbb{S}^{d-1}$, which can improve discriminativeness to some extent (see Fig. 1(c)).

3.1.1. Margin-based Loss

Based on the spherical constraint, a popular research line to encourage more discriminative features is to introduce an additional margin into softmax loss to draw a more strict decision boundary (Wang et al., 2018; Deng et al., 2019; Cao et al., 2019). The margin-based loss is formulated as

$$L_{\boldsymbol{\alpha}} = -\log \frac{\exp(s\boldsymbol{w}_y^{\mathrm{T}} \boldsymbol{z} + \alpha_y)}{\exp(s\boldsymbol{w}_y^{\mathrm{T}} \boldsymbol{z} + \alpha_y) + \sum_{j\neq y} \exp(s\boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{z})}, \tag{2}$$

where both $\boldsymbol{w}_j$ and $\boldsymbol{z} \in \mathbb{S}^{d-1}$, s is the inversion of the temperature parameter, and α_y is the introduced per-class margin that depends on the distribution of y. For instance, in one well-known variant of margin-based loss—the label-distribution-aware margin (LDAM) loss (Cao et al., 2019), the per-class margin is defined as $\alpha_y \propto \mathbb{P}(y)^{-1/4}$.

3.1.2. Sample Margin

According to the definition in (Koltchinskii et al., 2002; Cao et al., 2019), for the network $f(\boldsymbol{x}; \Theta, W) = W^\top \boldsymbol{\phi}_\Theta(\boldsymbol{x}) : \mathbb{R}^m \to \mathbb{R}^k$ that outputs k logits, the margin of a sample $(\boldsymbol{x}, y)$ is defined as

$$\gamma(\boldsymbol{x}, y) = f(\boldsymbol{x})_y - \max_{j\neq y} f(\boldsymbol{x})_j = \boldsymbol{w}_y^\top \boldsymbol{z} - \max_{j\neq y} \boldsymbol{w}_j^\top \boldsymbol{z}, \tag{3}$$

Let $S_j = \{i : y_i = j\}$ denote the sample indices corresponding to class j. We further define the sample margin for class j as $\gamma_j = \min_{i\in S_j} \gamma(\boldsymbol{x}_i, y_i)$, and the minimal sample margin over the entire dataset is $\gamma_{\min} = \min\{\gamma_1, ..., \gamma_k\}$.

Let $L_{\gamma,j}[f] = \mathbb{P}_{\boldsymbol{x}\sim\mathcal{P}_j}[\max_{j'\neq j} f(\boldsymbol{x})_{j'} > f(\boldsymbol{x})_j - \gamma]$ denote the hard margin loss on samples from class j, and let $\hat{L}_{\gamma,j}$ denote its empirical variant. When the training dataset is separable (which means that there exists f such that $\gamma_{\min} > 0$ for all training samples), Cao et al. (2019) provided a class-balanced generalization error bound: for $\gamma_j > 0$ and all $f \in \mathcal{F}$, with a high probability, one have

$$\begin{aligned} &\mathbb{P}_{(\boldsymbol{x},y)}[f(\boldsymbol{x})_y < \max_{l\neq y} f(\boldsymbol{x})_l] \\ \leq &\frac{1}{k}\sum_{j=1}^{k}\left(\hat{L}_{\gamma_j,j}[f] + \frac{4}{\gamma_j}\hat{\mathfrak{R}}_j(\mathcal{F}) + \varepsilon_j(\gamma_j)\right). \end{aligned} \tag{4}$$

In the right-hand side, $\frac{4}{\gamma_j}\hat{\mathfrak{R}}_j(\mathcal{F})$ that denotes the empirical Rademacher complexity has a big impact on the value of the

generalization bound. Cao et al. (2019) suggest that large margins $\{\gamma_j\}_{j=1}^k$ for all classes should be encouraged in order to tighten the generalization bound.

3.2. Theoretical Motivation

Inspired by Zhou et al. (2022), we turn to *tighten the generalization bound by maximizing the minimal sample margin* $\gamma_{\min}$. In this way, we can achieve the larger margins $\{\gamma_j\}$ for all classes. To this end, we have the following theorems:

Lemma 3.1 (**The optimality condition of prototypes to maximize** $\gamma_{\min}$**).** *If* $\boldsymbol{w}_1, ..., \boldsymbol{w}_k, \boldsymbol{z}_1, ..., \boldsymbol{z}_N \in \mathbb{S}^{d-1}$ *(*$2 \leq k \leq d+1$*), then the maximum of the minimal sample margin* $\gamma_{\min}$ *is* $\frac{k}{k-1}$*, which is uniquely obtained if* $\boldsymbol{z}_i = \boldsymbol{w}_{y_i}, \forall i$*, and* $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}, \forall i \neq j$.

This lemma indicates that, once the maximum of $\gamma_{\min}$ is obtained, the class prototypes $\boldsymbol{w}_1, ..., \boldsymbol{w}_k$ will satisfy $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}, \forall i \neq j$. In other words, the angular between any two class prototypes will be $\arccos \frac{-1}{k-1}$, which is actually also the solution of the *best-packing problem* on the sphere (Borodachov et al., 2019). This conclusion can also be derived for the loss in Eq. (2) with the same per-label margins when learning from balanced datasets.

Theorem 3.2. *For balanced datasets (i.e., each class has the same number of samples), if* $\boldsymbol{w}_1, ..., \boldsymbol{w}_k, \boldsymbol{z}_1, ..., \boldsymbol{z}_N \in \mathbb{S}^{d-1}$*,* $(2 \leq k \leq d+1)$*, then learning with* $L_{\boldsymbol{\alpha}}$ *that has the same per-class margins (i.e.,* $\alpha_j = \alpha, \forall j \in [k]$*) will deduce* $\boldsymbol{z}_i = \boldsymbol{w}_{y_i}, \forall i$*, and* $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}, \forall i \neq j$.

This theorem shows that, on balanced datasets, by minimizing the margin-based loss with the same per-class margins, we can achieve the maximum of $\gamma_{\min}$, and thus make the distribution of prototypes the most discriminative, and also make each feature concentrate on the corresponding prototype. However, on imbalanced datasets, this wonderful effect cannot be enjoyed anymore. As shown in Fig. 1(c), the learned representation of minority classes are squeezed into a narrow space, while that of majority classes occupy a large part of the feature space.

In the class-imbalanced setting, LDAM (Cao et al., 2019) intuitively assigns different margins to each class to enforce a large margin between the minority and majority classes. However, there is no theoretical guarantee that we will obtain class-balanced features and prototypes such as those used in learning with balanced datasets. Moreover, different per-class margins in LDAM cannot guarantee Fisher consistency (Lin, 2004) (or called classification calibration (Bartlett et al., 2006)), since it allows shifting the decision boundary away from the minority classes. We can derive the following theorem:

Theorem 3.3. *Under class-imbalanced data distribution, LDAM is not classification-calibrated.*

As a corroboration, Menon et al. (2021) empirically claims that the existence of different per-class margins would sacrifice the consistency with the Bayes-optimal solution and thus result in sub-optimal solutions even in simple settings.

3.3. Prototype-anchored Learning

As analyzed above, when learning with imperfect annotations, the current softmax loss and margin-based losses suffer from the lack of enough discrimination or Fisher consistency. Lemma 3.1 provides the optimality condition of prototypes to maximize the minimal sample margin. Inspired by the derived condition $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}, \forall i \neq j$, we propose a simple yet effective method, namely *prototype-anchored learning* (PAL). Specifically, we derive a classifier composed of predefined prototypes satisfying the optimality condition in Lemma 3.1, and the classifier is anchored during training, *i.e.*, performs no gradient updates. In practice, to generate these prototypes we randomly initialize $\{\boldsymbol{z}_i\}_{i=1}^N$ and $\{\boldsymbol{w}_i\}_{i=1}^k$ in $L_{\boldsymbol{\alpha}}$ ($\boldsymbol{\alpha} = 0$) with a balanced setting (*i.e.*, $N = k$ and $y_i = i$), and then directly minimize $L_{\boldsymbol{\alpha}}$ to obtain the optimal prototypes $\{\boldsymbol{w}_i\}_{i=1}^k$ according to Theorem 3.2. The pseudo code can be seen in Tab. 7.

Prototype-anchored learning can be regarded as a parameterized loss with respect to the feature representations. As shown in Eq. (1), the softmax loss parameterized by prototypes $(\boldsymbol{w}_1, ..., \boldsymbol{w}_k)$ enlarges the inner product $\boldsymbol{w}_{y_i}^\top \boldsymbol{z}_i$ and shrinks other inner products $\boldsymbol{w}_j^\top \boldsymbol{z}_i, \forall j \neq y_i$. When these prototypes are anchored, the classification problem is transformed to a feature alignment problem, which is intuitively more stable than jointly learning feature extractor and classifier prototypes, especially for learning with imperfect data.

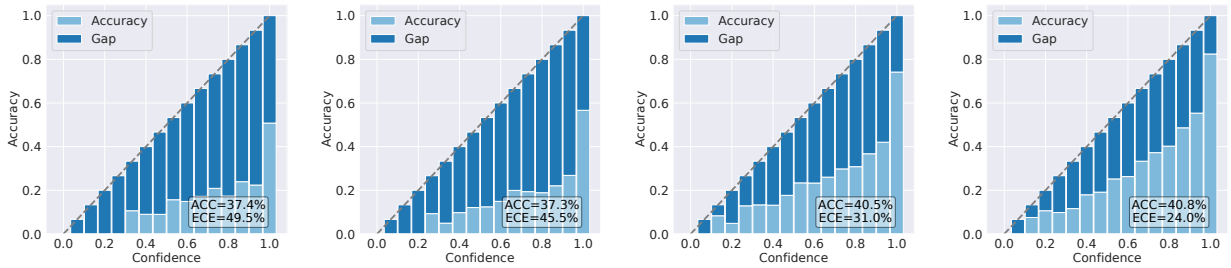
4. Applications of PAL

Due to its simplicity, PAL can be easily incorporated into various learning-based classification schemes. In this section, we present how to apply PAL in handling imperfect annotations. We specially consider two scenarios—class-imbalanced learning and learning with noisy labels.

4.1. Class-imbalanced Learning

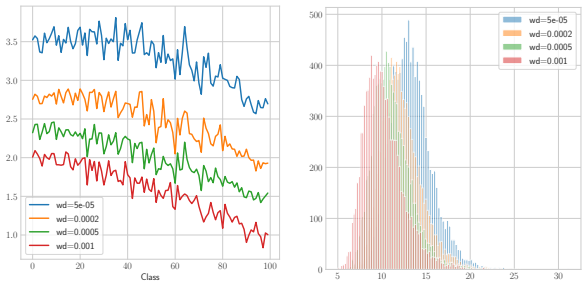
Theorem 3.2 proves that learning with L_α will lead to prototypes that satisfy $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}, \forall i \neq j$, which however works for balanced datasets. For imbalanced datasets, a phenomenon termed *Minority Collapse* would occur (Fang et al., 2021). In the following, we demonstrate in theory that the anchored prototypes can alleviate minority collapse in imbalanced training and acquire the maximum of $\gamma_{\min}$:

Theorem 4.1. *For imbalanced or balanced datasets, if* $\boldsymbol{z}_1, ..., \boldsymbol{z}_N, \boldsymbol{w}_1, ..., \boldsymbol{w}_k \in \mathbb{S}^{d-1}$ $(2 \leq k \leq d+1)$*, where* $\boldsymbol{w}_1, ..., \boldsymbol{w}_k$ *are anchored and satisfy that* $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}$*,* $\forall i \neq j$*, then learning with* $L_{\boldsymbol{\alpha}}$ *will deduce* $\boldsymbol{z}_i = \boldsymbol{w}_{y_i}, \forall i$*,*



(a) CE with wd=0.0 (b) CE with wd=5e-5 (c) CE with wd=5e-4 (d) CE with wd=1e-3

Figure 2. Reliability diagrams of ResNet-32 trained by CE on CIFAR-100-LT with imbalance ratio 100 under different weight decays (wds). As can be seen, an appropriate larger weight decay can improve both accuracy and confidence.



(a) prototype norms (b) histogram of feature norms

Figure 3. Illustration of prototypes norms and feature norms by CE trained on CIFAR-100-LT with imbalance ratio 100 under different weight decays. As can be seen, the larger weight decay usually leads to smaller prototype norms and feature norms.

and obtain the maximum of the minimal sample margin $\gamma_{\min} = \frac{k}{k-1}$.

Theorem 4.1 offers a solid theoretical guarantee to obtain the maximal $\gamma_{\min}$ on imbalanced datasets by PAL.

In class-imbalanced learning, a fruitful avenue of exploration is the strategy of decoupling the learning procedure into representation learning and classification training (Kang et al., 2020), which works in a two-stage manner: firstly learn feature representation from the original imbalanced data, and then retrain the classifier using class-balanced sampling with the first-stage representation extractor frozen. The state-of-the-art decoupling methods usually utilize the softmax loss for the representation learning (Kang et al., 2020; Zhong et al., 2021), which does not explicitly perform normalization on features. This seems to invalidate the normalization condition of Theorem 3.1. Fortunately, in deep neural networks, there are some implicit biases that actually conduct restriction on features and prototypes, such as weight decay (Fang et al., 2021). Figure 3 shows that weight decay can facilitate not only explicitly learning small weights of networks but also implicitly producing small feature representations. Moreover, we empirically find that *an appropriately large weight decay can mitigate class imbalance* by preventing the excessively large norms that cause over-confidence (see more analysis in Appendix B.1) in Fig. 2. This makes our PAL still work well for the softmax loss in imbalanced setting with limited feature norms. Formally, we have the following theorem:

Theorem 4.2. *For imbalanced or balanced datasets, if* $\|\boldsymbol{z}_i\| \leq B, \forall i \in [1, n]$, *and the class prototypes* $\boldsymbol{w}_1, ..., \boldsymbol{w}_k \in \mathbb{S}^{d-1}$ $(2 \leq k \leq d+1)$ *are anchored to satisfy* $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}, \forall i \neq j$, *then learning with the softmax loss will deduce* $\frac{\boldsymbol{z}_i^\top \boldsymbol{w}_{y_i}}{\|\boldsymbol{z}_i\|_2} = 1, \forall i$, *and obtain the maximum of* $\gamma_{\min}$.

According to this theoretical guarantee, PAL can be used in the feature representation learning of the first stage, as well as in tandem with other performance-boosting approaches and modules. Through such a simple operation, as demonstrated by extensive experiments in Section 5.1, PAL can significantly boost the performance of the existing long-tailed classification methods.

4.2. Learning with Noisy Labels

In noise-tolerant learning, the most popular family of loss functions is the symmetric loss (Manwani & Sastry, 2013; van Rooyen et al., 2015; Ghosh et al., 2017), which satisfies

$$\sum_{i=1}^{k} L(f(\boldsymbol{x}), i) = C, \forall \boldsymbol{x} \in \mathcal{X}, \forall f. \tag{5}$$

where C is a constant. This symmetric condition theoretically guarantees the noise tolerance by risk minimization on a symmetric loss function (Ghosh et al., 2017) under symmetric label noise, *i.e.*, the global minimizer of the noisy L-risk $R_L^\eta(f) = \mathbb{E}_{\mathcal{D}}[(1 - \eta_{\boldsymbol{x}})L(f(\boldsymbol{x}), y) + \sum_{i \neq y} \eta_{\boldsymbol{x},i} L(f(\boldsymbol{x}), i)]$ also minimizes the L-risk $R_L(f) = \mathbb{E}_{\mathcal{D}} L(f(\boldsymbol{x}), y)$, where $\eta_{\boldsymbol{x},\boldsymbol{i}}$ denotes the probability (or called noise rate) of flipping label y into label i.

Negative-signed Sample Logit Loss. In this work, we propose a novel loss for noise-tolerant learning based on sample logit $f(\boldsymbol{x}) = W^\top \phi_\Theta(\boldsymbol{x})$, called *negative-signed*

sample logit loss (NSL), which is defined as:

$$L_{\mathrm{NSL}}(f(\boldsymbol{x}), i) = -f(\boldsymbol{x})_i = -\boldsymbol{w}_i^\top \boldsymbol{\phi}_{\boldsymbol{\Theta}}(\boldsymbol{x}). \quad (6)$$

Importantly, when $\sum_{i=1}^k \boldsymbol{w}_i = 0$, we can simply derive that $\sum_{i=1}^k L_{\mathrm{NSL}}(f(\boldsymbol{x}), i) = 0$. Thus, NSL also serves as a symmetric loss. However, NSL only contains the fitting term that encourages each feature to be close to the corresponding prototype, which would lead to a trivial solution when jointly training Θ and W. Fortunately, when prototypes are anchored, this issue can be easily handled as the solution of the best packing problem (which implicitly satisfies $\sum_{i=1}^k \boldsymbol{w}_i = 0$). Specifically, we have:

Proposition 4.3. *If the prototypes $\boldsymbol{w}_1, ..., \boldsymbol{w}_k \in \mathbb{S}^{d-1}$ ($2 \leq k \leq d+1$) are anchored to satisfy $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}$, $\forall i \neq j$, $L_{\mathrm{NSL}}(f(\boldsymbol{x}), i)$ is symmetric. More specifically, we have $\sum_{i=1}^k L_{\mathrm{NSL}}(f(\boldsymbol{x}), i) = 0$, and learning with L_{NSL} will lead to the maximum of $\gamma_{\min}$ under symmetric label noise.*

More generally, given the anchored prototypes $\boldsymbol{w}_1, ..., \boldsymbol{w}_k$, we can provide a risk bound for the loss functions satisfying that $L_W(\boldsymbol{z}) = \sum_{i=1}^k L(W^\top \boldsymbol{z}, i)$ is λ-Lipschitz[1]:

Theorem 4.4. *In a multi-class classification problem, given $\boldsymbol{w}_1, ..., \boldsymbol{w}_k$, if $\boldsymbol{z} = \boldsymbol{\phi}_\Theta(x)$ is norm-bounded by B, i.e., $\|\boldsymbol{z}\|_2 \leq B$, then for any loss $L(\boldsymbol{z}, i)$ satisfying $L_W(\boldsymbol{z}) = \sum_{i=1}^k L(W^\top \boldsymbol{z}, i)$ is λ-Lipschitz, we have the following risk bound under symmetric label noise with $\eta < \frac{k-1}{k}$:*

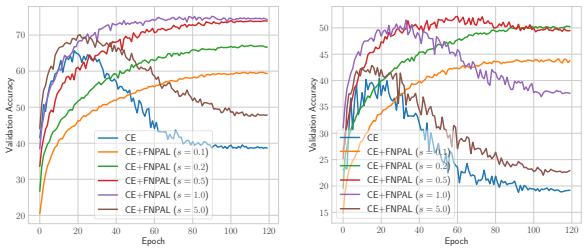
$$R_L(\hat{f}) - R_L(f^*) \leq \frac{2\eta\lambda B}{(1-\eta)k - 1}, \quad (7)$$

where $\hat{f}$ and f^* denote a global minimize of $R_L^\eta(f)$ and $R_L(f)$, respectively.

Theorem 4.4 naturally encompasses the symmetric condition in (Ghosh et al., 2017, Theorem 1). We know that if the loss L is symmetric, then $\lambda = 0$, and thus we have $R_L(\hat{f}) = R_L(f^*)$.

For general losses, when B is fixed, the risk bound will depend on the Lipschitz constant of $L_W(\boldsymbol{z})$ and noise rate η. As can be seen, the Lipschitz constant depends on the choice of W. Actually, our PAL strategy in which the prototypes are anchored can provide a smaller λ than learning with unanchored prototypes, especially for some losses like CE that are usually not Lipschitz continuous. This will be described later. We also note that the feature norm B can be naturally adjusted by tuning the scaled parameter s under the spherical constraint on features and prototypes (*e.g.*, $L_{\boldsymbol{\alpha}}$ in Eq. (2)), which can be served as a parameter to trade off fitting ability and robustness.

We further validate the new trade-off by adjusting the inversion of temperature parameter. As illustrated in Fig. 4, an appropriate small feature norm can achieve good robustness and accuracy. However, it can be observed that the same parameter s at different noise rates does not perform exactly the same. For example, when $s = 0.2$, it obtains the best accuracy at $\eta = 0.8$ but encounters underfitting at $\eta = 0.6$. This is mainly due to that smaller noise rates require more fitting power. We propose to set $s = O(\frac{1}{\eta+\epsilon})$, which makes the risk bound in Eq. (7) be noise-independent.



(a) $\eta = 0.6$ (b) $\eta = 0.8$

Figure 4. Validation accuracies of CE and CE+FNPAL (with different scaled parameter s) on CIFAR-10 with symmetric label noise. As can be seen, CE exhibits significant overfitting after epoch 20 while CE+FNPAL shows more robustness and achieve better performance. To conclude, as s decreases, the robustness increases. This empirically demonstrates that small feature norms can significantly mitigate label noise when prototypes are anchored. On the other hand, the fitting ability decreases as s increases. That is, s introduces a trade-off between fitting ability and robustness. Thus, an appropriate small s is preferred for better performance.

Based on the above discussion, finally we suggest the *feature normalized and prototype-anchored learning* (FNPAL) strategy that performs ℓ_2 normalization on features and anchored prototypes, which can be combined with traditional losses to boost their ability on noise-tolerant learning. Specifically, for cross-entropy loss, we have the following proposition:

Proposition 4.5. *In a multi-class classification problem, let $\boldsymbol{w}_1, ..., \boldsymbol{w}_k \in \mathbb{S}^{d-1}$ ($2 \leq k \leq d+1$) satisfy $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}$, $\forall i \neq j$, if $\boldsymbol{z} = \boldsymbol{\phi}_\Theta(x)$ is norm-bounded by B, i.e., $\|\boldsymbol{z}\|_2 \leq B$, then we have the following risk bound for the CE loss defined in Eq. (1) under symmetric label noise with $\eta < \frac{k-1}{k}$:*

$$R_L(\hat{f}) - R_L(f^*) \leq \frac{2c\eta k(1-t)B}{k - 1 + t(k-1)^2}, \quad (8)$$

where $c = \frac{k-1}{(1-\eta)k-1}$, $t = \exp(-\frac{kB}{k-1})$, $\hat{f}$ and f^* denote the global minimum of $R_L^\eta(f)$ and $R_L(f)$, respectively.

Remark. Proposition 4.5 provides a FNPAL-based risk bound that relies on the Lipschitz constant $\lambda_{\mathrm{PAL}} = \frac{k(1-\exp(-\frac{kB}{k-1}))}{1+(k-1)\exp(-\frac{kB}{k-1})}$. Actually, FNPAL indicates a tighter risk bound by a smaller Lipschitz constant than unnormalized or unanchored cases: (1) when $\boldsymbol{w}$ is not normalized, we can easily know that $L_W(\boldsymbol{z})$ is not Lips-

[1] that is, $\|L_W(\boldsymbol{z}_1) - L_W(\boldsymbol{z}_2)\|_2 \leq \lambda \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2$

chitz continuous as $\boldsymbol{w}_i = t\boldsymbol{z}$ and $t \to \infty$; (2) when $\boldsymbol{w}_i$ is normalized and unanchored (*w.l.o.g*, $\|\boldsymbol{w}_i\|_2 = 1$) and $\boldsymbol{z}$ is unnormalized, we have the Lipschitz constant $\sup_{W,\boldsymbol{z}} \|\frac{\partial L_W(\boldsymbol{z})}{\partial \boldsymbol{z}}\|_2 \geq k > \lambda_{\text{PAL}}$; (3) when $\boldsymbol{w}_i$ and $\boldsymbol{z}$ are normalized and unanchored ($\|\boldsymbol{w}_i\|_2 = 1$, $\|\boldsymbol{z}\|_2 = B$), we have $\sup_{W,\boldsymbol{z}} \|\frac{\partial L_W(\boldsymbol{z})}{\partial \boldsymbol{z}}\|_2 \geq \frac{2(\exp(2B)-1)}{\frac{\exp(2B)}{k-1}+1} > \lambda_{\text{PAL}}$. More details are provided in Appendix A.7.2.

Similar risk bounds can also be derived for GCE (Zhang & Sabuncu, 2018) and Focal loss (Lin et al., 2017). Please see Appendix A.7 for more details.

Table 1. Validation accuracy (%) on ImageNet-LT. The results with positive gains are **boldfaced** and the best one is underlined.

Method	Many	Medium	Few	All
CE	66.8	36.9	7.1	43.6
FL	64.3	37.1	8.2	43.7
OLTR	51.0	40.8	20.8	41.9
Causal Norm	65.2	47.7	29.8	52.0
Balanced Softmax	63.6	48.4	32.9	52.1
LADE	65.1	48.9	33.4	53.0
cRT+mixup	63.9	49.1	30.2	51.7
LWS+mixup	62.9	49.8	31.6	52.0
MiSLAS	61.7	51.3	35.8	52.7
CE+PAL	**69.0**	**42.5**	**11.0**	**47.6**
MiSLAS+PAL	**64.0**	**51.6**	32.4	**53.3**

5. Experiments

In this section, we empirically investigate the effectiveness of prototype-anchored learning on two tasks: long-tailed classification and learning with noisy labels. The benchmark datasets include benchmark datasets: MNIST (Lecun et al., 1998), CIFAR-10/-100 (Krizhevsky & Hinton, 2009) and ImageNet (Deng et al., 2009) as well as a real-world long-tailed and noisy dataset WebVision (Li et al., 2017).

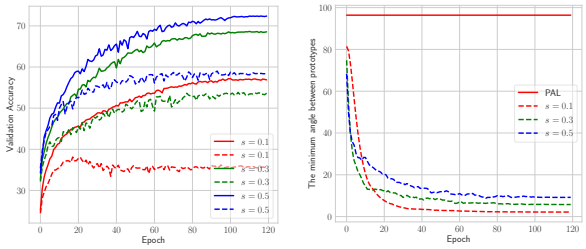
5.1. Evaluation on Long-tailed Classification

We evaluate our approach on benchmarks CIFAR-100-LT and ImageNet-LT with artificially created long-tailed settings. We follow the controllable data imbalance strategy in (Cao et al., 2019; Zhong et al., 2021) by reducing the number of training examples per class and keeping the validation set unchanged. The imbalance ratio $\rho = \max_i n_i / \min_i n_i$ denotes the ratio between sample sizes of the most frequent and the least frequent classes.

Baselines. We compare the proposed method against previous one-stage methods of Focal loss (FL) (Lin et al., 2017), LDAM (Cao et al., 2019), OLTR (Liu et al., 2019), and LADE (Hong et al., 2021), against two-stage methods including LDAM+DRW (Cao et al., 2019), cRT (Kang et al., 2020), BNN (Zhou et al., 2020b), and MiSLAS (Zhong et al., 2021), and an ensemble method TADE (Zhang et al., 2021b). We simply embed PAL into one-stage methods CE and LDAM, and a two-stage method MiSLAS. For MiSLAS, we only use PAL in the representation learning stage.

Table 2. Validation accuracy (%) on CIFAR-100-LT. The results with positive gains are **boldfaced** and the best one is underlined.

Method	CIFAR-100-LT		
	Imbalance Ratio		
	100	50	10
CE	38.4	43.9	55.8
FL	38.4	44.3	55.8
CE+mixup	39.6	45.0	58.2
LDAM	41.8	45.6	59.0
BBN	42.6	47.1	59.2
cRT+mixup	45.1	50.9	62.1
LWS+mixup	44.2	50.7	62.3
PCSoftmax	45.3	49.5	61.2
LADE	45.4	49.5	59.0
LDAM+DRW	45.2	50.0	60.8
MiSLAS	47.0	52.3	63.2
TADE	49.8	53.9	63.6
CE+PAL	**39.2**	**44.4**	**57.3**
CE+mixup+PAL	**41.9**	**47.2**	**60.2**
LDAM+PAL	**42.7**	**46.9**	**59.4**
LDAM+DRW+PAL	**45.9**	**50.1**	**61.4**
MiSLAS+PAL	**50.6**	**55.3**	**64.2**



(a) Validation accuracy (b) The minimum angle

Figure 5. Validation of FNPAL in mitigating label noise. Solid and dashed lines denote the curve of CE+FNPAL and CE+FNPN, respectively. (a) and (b) denote the curves of validation accuracy and the minimum angle between prototypes, respectively.

Results. The experiments follow the settings in (Zhong et al., 2021). Tables 1 and 2 report the validation accuracy results of each method on CIFAR-100-LT and ImageNet-LT, respectively. As can be seen, our proposed PAL obviously improve the performance on both one-stage and two-stage methods. Particularly, the PAL-based MiSLAS outperform the ensemble method TADE and achieve the best results on CIFAR-100-LT, which demonstrate that PAL can significantly improve representation learning.

5.2. Evaluation on Learning with Noisy Labels

We evaluate the proposed method on MNIST, CIFAR-10/-100 with synthesis noisy labels. We follow the noise generation, networks and training settings in (Ma et al., 2020).

Table 3. Validation accuracies (%) of different methods on benchmark datasets with clean or symmetric label noise ($\eta \in [0.2, 0.4, 0.6, 0.8]$). The results (mean $\pm$ std) are reported over 3 random runs. The results with positive gains are **boldfaced** and the best one is underlined.

Dataset	Method	Clean ($\eta = 0.0$)	Symmetric Noise Rate (η) 0.2	0.4	0.6	0.8
MNIST	CE	99.17 ± 0.04	91.40 ± 0.11	74.36 ± 0.29	49.32 ± 0.70	22.32 ± 0.15
	FL	99.16 ± 0.02	91.49 ± 0.20	75.28 ± 0.10	50.25 ± 0.70	22.68 ± 0.14
	GCE	99.15 ± 0.02	98.90 ± 0.03	96.81 ± 0.23	81.39 ± 0.64	33.07 ± 0.31
	SCE	99.28 ± 0.07	98.91 ± 0.12	97.60 ± 0.22	88.00 ± 0.50	47.32 ± 0.99
	NCE+MAE	99.42 ± 0.02	99.18 ± 0.08	98.47 ± 0.21	95.52 ± 0.04	73.05 ± 0.59
	NCE+RCE	99.40 ± 0.04	99.24 ± 0.01	98.44 ± 0.11	95.77 ± 0.09	74.80 ± 0.28
	NFL+RCE	99.37 ± 0.01	99.16 ± 0.03	98.55 ± 0.05	95.62 ± 0.24	74.67 ± 0.97
	NSL	99.24 ± 0.03	98.99 ± 0.03	98.58 ± 0.11	95.99 ± 0.24	59.77 ± 1.98
	CE+FNPAL	99.24 ± 0.05	**99.05 ± 0.04**	**98.66 ± 0.04**	**97.62 ± 0.15**	**79.23 ± 0.87**
	SCE+FNPAL	99.27 ± 0.04	**99.06 ± 0.05**	**98.76 ± 0.09**	**97.94 ± 0.07**	**88.56 ± 1.07**
	NCE+RCE+FNPAL	99.29 ± 0.04	99.04 ± 0.07	98.11 ± 0.09	94.84 ± 0.08	**79.70 ± 1.06**
	NFL+RCE+FNPAL	99.29 ± 0.06	99.02 ± 0.05	98.32 ± 0.14	95.38 ± 0.11	**76.06 ± 0.58**
CIFAR10	CE	90.36 ± 0.25	74.78 ± 0.68	57.95 ± 0.12	38.21 ± 0.12	18.89 ± 0.43
	FL	89.69 ± 0.25	74.19 ± 0.23	57.35 ± 0.27	38.11 ± 0.76	19.39 ± 0.44
	GCE	89.37 ± 0.29	87.05 ± 0.21	82.43 ± 0.10	68.05 ± 0.07	25.21 ± 0.28
	SCE	91.24 ± 0.19	87.34 ± 0.01	79.84 ± 0.43	61.09 ± 0.19	27.19 ± 0.34
	NCE+MAE	89.02 ± 0.09	87.06 ± 0.17	83.92 ± 0.16	76.47 ± 0.25	45.01 ± 0.31
	NCE+RCE	91.12 ± 0.14	89.21 ± 0.00	86.03 ± 0.14	80.04 ± 0.26	51.67 ± 1.38
	NFL+RCE	91.03 ± 0.15	89.10 ± 0.16	86.20 ± 0.19	79.58 ± 0.08	50.03 ± 2.78
	NSL	88.07 ± 0.12	86.46 ± 0.02	83.27 ± 0.13	76.17 ± 0.40	46.74 ± 0.72
	CE+FNPAL	90.69 ± 0.11	**86.34 ± 0.37**	**81.30 ± 0.29**	**72.77 ± 0.41**	**51.46 ± 1.10**
	SCE+FNPAL	91.11 ± 0.13	87.30 ± 0.06	**82.68 ± 0.22**	**73.49 ± 0.42**	**51.99 ± 1.10**
	NCE+RCE+FNPAL	90.88 ± 0.10	**89.34 ± 0.15**	**86.65 ± 0.21**	**80.28 ± 0.07**	**57.21 ± 0.22**
	NFL+RCE+FNPAL	91.16 ± 0.25	**89.49 ± 0.32**	**86.66 ± 0.08**	**80.33 ± 0.15**	**56.23 ± 0.15**
CIFAR100	CE	70.41 ± 1.17	55.64 ± 0.17	40.39 ± 0.46	22.00 ± 1.23	7.37 ± 0.16
	FL	70.56 ± 0.59	56.02 ± 0.80	40.41 ± 0.39	22.11 ± 0.30	7.70 ± 0.20
	GCE	63.06 ± 1.00	62.15 ± 0.66	57.11 ± 1.43	45.99 ± 1.00	18.32 ± 0.36
	SCE	70.41 ± 0.63	55.05 ± 0.68	39.60 ± 0.14	21.53 ± 0.72	7.82 ± 0.30
	NCE+MAE	67.16 ± 0.13	52.34 ± 0.12	35.81 ± 0.42	19.29 ± 0.29	7.31 ± 0.23
	NCE+RCE	68.09 ± 0.26	64.32 ± 0.40	58.11 ± 0.63	45.94 ± 1.31	25.22 ± 0.08
	NFL+RCE	67.58 ± 0.39	64.48 ± 0.50	57.86 ± 0.12	46.74 ± 0.59	24.55 ± 0.47
	NSL	70.08 ± 0.19	65.30 ± 0.36	56.77 ± 0.52	41.21 ± 1.01	12.16 ± 0.96
	CE+FNPAL	71.69 ± 0.27	**65.38 ± 0.17**	**57.24 ± 0.36**	**41.35 ± 0.19**	**12.12 ± 0.88**
	SCE+FNPAL	70.87 ± 0.45	**65.30 ± 0.15**	**55.10 ± 0.45**	**39.73 ± 0.04**	**11.70 ± 0.53**
	NCE+RCE+FNPAL	69.29 ± 0.32	**65.53 ± 0.30**	**60.53 ± 0.27**	**49.73 ± 0.64**	24.54 ± 0.28
	NFL+RCE+FNPAL	69.53 ± 0.05	**65.94 ± 0.32**	**60.89 ± 0.60**	**50.10 ± 0.40**	24.15 ± 1.06

Baselines. We consider several state-of-the-art methods: Generalized Cross Entropy (GCE) (Zhang & Sabuncu, 2018), Symmetric Cross Entropy (SCE) (Wang et al., 2019b), and the Active Passive Loss (APL) (Ma et al., 2020) including NCE+MAE, NCE+RCE, and NFL+RCE. We also compare the commonly used losses CE and Focal Loss (FL). The best parameter settings for all baselines are verified by (Ma et al., 2020), thus we follow them. For our feature-normalized PAL (FNPAL) with CE and SCE, we utilize the noise-aware way, *i.e.*, $s = \frac{0.25}{0.05+\eta}$, and $s = 10$ for APLs.

Feature-normalized PAL is better than one without PAL. As aforementioned, feature normalization is important to mitigate label noise but is not sufficient for fitting ability. Proposition 4.5 shows that anchored prototypes will provide a tighter risk bound. To verify this, we run experiments on CIFAR-10 with 0.6 symmetric label noise by CE with feature normalization and prototype normalization (CE+FNPN) and CE with feature-normalized PAL (CE+FNPAL). Under the same s, CE+FNPAL always shows better performance than CE+FNPN as shown in Fig. 5(a). The main reason is that the prototypes of CE+FNPN are defective during training. In Fig. 5(b), the minimum angle between prototypes of CE+FNPN is very small during training, *i.e.*, there are at least two classes that are hard to classify.

Results. Tables 3 and 4 report the validation accuracy results of different methods on the benchmark datasets with symmetric label noise and asymmetric label noise, respectively. As can be seen, our proposed FNPAL can significantly improve the baselines in most label noise settings on MNIST, CIFAR10/-100. For example, CE+FNPAL per-

Table 4. Validation accuracies (%) of different methods on benchmark datasets with asymmetric label noise ($\eta \in [0.2, 0.3, 0.4]$). The results (mean $\pm$ std) are reported over 3 random runs. C10 and C100 denote CIFAR-10 and CIFAR-100, respectively. The results with positive gains are **boldfaced** and the best one is underlined. * denotes the method training with FNPAL.

Dataset	Method	Asymmetric Noise Rate (η) 0.2	0.3	0.4
MNIST	CE	94.42 ± 0.28	88.74 ± 0.07	81.45 ± 0.09
	FL	94.18 ± 0.05	88.87 ± 0.30	81.79 ± 0.36
	GCE	96.57 ± 0.15	89.20 ± 0.11	81.92 ± 0.45
	SCE	98.13 ± 0.17	93.59 ± 0.24	84.42 ± 0.57
	NCE+MAE	98.87 ± 0.10	97.00 ± 0.15	91.33 ± 0.65
	NCE+RCE	98.91 ± 0.08	97.37 ± 0.20	86.95 ± 4.27
	NFL+RCE	98.87 ± 0.05	97.41 ± 0.12	92.35 ± 0.26
	NSL	98.79 ± 0.05	95.85 ± 0.12	87.41 ± 0.62
	CE*	$\mathbf{98.58 \pm 0.05}$	$\mathbf{96.94 \pm 0.27}$	$\mathbf{91.77 \pm 0.94}$
	SCE*	$\mathbf{99.08 \pm 0.02}$	$\mathbf{98.43 \pm 0.13}$	$\mathbf{92.11 \pm 0.65}$
	NCE+RCE*	$\mathbf{98.93 \pm 0.04}$	$\mathbf{97.91 \pm 0.29}$	$\mathbf{94.37 \pm 0.54}$
	NFL+RCE*	$\mathbf{99.02 \pm 0.10}$	$\mathbf{98.18 \pm 0.19}$	$\mathbf{94.86 \pm 0.18}$
C10	CE	83.34 ± 0.32	79.58 ± 0.22	74.61 ± 0.42
	FL	83.31 ± 0.23	79.67 ± 0.08	74.19 ± 0.40
	GCE	85.91 ± 0.02	80.56 ± 0.30	74.49 ± 0.34
	SCE	86.47 ± 0.14	81.61 ± 0.17	75.52 ± 0.28
	NCE+MAE	86.52 ± 0.28	83.74 ± 0.19	76.75 ± 0.33
	NCE+RCE	88.61 ± 0.27	85.55 ± 0.01	79.25 ± 0.33
	NFL+RCE	88.58 ± 0.15	85.53 ± 0.26	79.61 ± 0.07
	NSL	86.23 ± 0.23	83.56 ± 0.15	78.63 ± 0.28
	CE*	$\mathbf{86.11 \pm 0.19}$	$\mathbf{83.06 \pm 0.06}$	$\mathbf{77.41 \pm 0.27}$
	SCE*	$\mathbf{87.32 \pm 0.21}$	$\mathbf{84.18 \pm 0.10}$	$\mathbf{78.15 \pm 0.34}$
	NCE+RCE*	$\mathbf{89.01 \pm 0.18}$	$\mathbf{85.97 \pm 0.19}$	$\mathbf{79.51 \pm 0.19}$
	NFL+RCE*	$\mathbf{88.64 \pm 0.17}$	$\mathbf{86.20 \pm 0.09}$	$\mathbf{79.66 \pm 0.60}$
C100	CE	57.67 ± 0.89	50.63 ± 0.57	41.95 ± 0.12
	FL	58.37 ± 0.56	51.44 ± 0.79	42.16 ± 0.53
	GCE	59.56 ± 0.64	54.22 ± 1.18	42.18 ± 0.77
	SCE	58.15 ± 0.54	50.58 ± 0.27	41.57 ± 0.19
	NCE+MAE	52.03 ± 0.15	44.45 ± 0.22	36.86 ± 0.46
	NCE+RCE	62.93 ± 0.42	55.70 ± 0.14	42.73 ± 0.58
	NFL+RCE	62.94 ± 0.39	55.83 ± 0.46	42.63 ± 0.13
	NSL	58.16 ± 0.46	50.35 ± 0.41	40.38 ± 0.60
	CE*	$\mathbf{57.93 \pm 0.18}$	$\mathbf{51.45 \pm 0.92}$	$\mathbf{47.53 \pm 0.94}$
	SCE*	$\mathbf{57.61 \pm 0.22}$	$\mathbf{50.74 \pm 0.51}$	$\mathbf{44.31 \pm 0.22}$
	NCE+RCE*	$\mathbf{64.63 \pm 0.32}$	$\mathbf{57.43 \pm 0.53}$	$\mathbf{49.77 \pm 1.11}$
	NFL+RCE*	$\mathbf{64.32 \pm 0.59}$	$\mathbf{57.92 \pm 0.65}$	$\mathbf{50.19 \pm 1.60}$

forms much better than the original CE in all noisy cases. Moreover, we also add FNPAL to APLs (NCE+RCE and NFL+RCE), which surprisingly brings obvious positive gains and achieves the state-of-the-art methods in most cases. The results demonstrate that FNPAL can be robust enough to embed into the existing methods and get the outstanding performance for symmetric and asymmetric label noise.

5.3. Evaluation on a Real-world Dataset

We also evaluate our proposed prototype-anchored learning on a real-world imbalanced and noisy dataset WebVision 1.0 (Li et al., 2017). We follow the "mini" setting in (Jiang et al., 2018; Ma et al., 2020) that takes the first 50 concepts of the Google resized image subset as the training dataset and further validates the trained ResNet-50 (He et al., 2016) on the same 50 concepts in validation set.

Table 5. Top-1 validation accuracies (%) on mini-WebVision.

Method	CE	FL	NCE+RCE	**NSL**	**CE+PAL**	**CE+FNPAL**
Acc	62.60	63.80	66.32	**69.56**	**68.92**	**69.69**

Results. The top-1 validation accuracies of different methods are reported in Table 5. Based on PAL, our proposed NSL shows very good performance, while CE+PAL significantly improve CE and outperform NCE+RCE a lot. Moreover, CE with feature-normalized PAL (CE+FNPAL) also brings 0.77 gains to CE+PAL. The results demonstrate that PAL can effectively help the trained model against real-world class imbalance and label noise.

6. Conclusion

In this paper, we presented a simple yet effective method for learning with imperfect annotations, which is also theoretically sound. We formulate the goal of tightening the margin-based generalization bound as maximizing the minimal sample margin $\gamma_{\min}$, and provide the optimality condition of prototypes to maximize $\gamma_{\min}$, from which the proposed prototype-anchored learning (PAL) is derived. Our PAL strategy can be easily embedded into various learning based classification under imperfect annotations. We specially consider two scenarios–class-imbalanced learning and learning with noisy labels. Extensive experiments are provided to demonstrate the advantage of PAL. In future work, we will further explore the application of PAL on other classification problems under imperfect annotation.

Acknowledgments

This work was supported by National Key Research and Development Project under Grant 2019YFE0109600, National Natural Science Foundation of China under Grants 61922027 and 6207115, and Beijing Municipal Science and Technology Commission Grant Z201100005820005.

References

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pp. 233–242. PMLR, 2017.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Borodachov, S. V., Hardin, D. P., and Saff, E. B. *Discrete energy on rectifiable sets*. Springer, 2019.

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Byrd, J. and Lipton, Z. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pp. 872–881. PMLR, 2019.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Charoenphakdee, N., Lee, J., and Sugiyama, M. On symmetric losses for learning from corrupted labels. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 961–970, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Fang, C., He, H., Long, Q., and Su, W. J. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

Feng, L., Shu, S., Lin, Z., Lv, F., Li, L., and An, B. Can cross entropy loss be robust to label noise? In *IJCAI*, 2020.

Ghosh, A., Kumar, H., and Sastry, P. S. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Han, B., Yao, Q., Liu, T., Niu, G., Tsang, I., Kwok, J. T., and Sugiyama, M. A survey of label-noise representation learning: Past, present and future. *ArXiv*, abs/2011.04406, 2020.

Han, H., Wang, W.-Y., and Mao, B.-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pp. 878–887. Springer, 2005.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., and Chang, B. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6626–6636, 2021.

Huang, C., Li, Y., Loy, C. C., and Tang, X. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5375–5384, 2016.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2304–2313. PMLR, 2018.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2020.

Koltchinskii, V., Panchenko, D., et al. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of statistics*, 30(1):1–50, 2002.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Computer Science Department, University of Toronto, Tech. Rep*, 1, 01 2009.

Krogh, A. and Hertz, J. A. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pp. 950–957, 1992.

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

Li, W., Wang, L., Li, W., Agustsson, E., and Van Gool, L. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

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

Lin, Y. A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1):73–82, 2004.

Liu, W., Jiang, Y.-G., Luo, J., and Chang, S.-F. Noise resistant graph ranking for improved web image search. In *CVPR 2011*, pp. 849–856. IEEE, 2011.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

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

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2537–2546, 2019.

Long, P. M. and Servedio, R. A. Random classification noise defeats all convex potential boosters. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 608–615, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054.

Lyu, Y. and Tsang, I. Curriculum loss: Robust learning and generalization against label corruption. *ICLR*, 2020.

Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., and Bailey, J. Normalized loss functions for deep learning with noisy labels. In *ICML*, 2020.

Manwani, N. and Sastry, P. S. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics*, 43(3): 1146–1151, 2013. doi: 10.1109/TSMCB.2012.2223460.

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

Menon, K. A., Rawat, S. A., Reddi, J. S., and Kumar, S. Can gradient clipping mitigate label noise. *ICLR*, 2020.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2233–2241, 2017. doi: 10.1109/CVPR.2017.240.

Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., and Li, H. Balanced meta-softmax for long-tailed visual recognition. In *Proceedings of Neural Information Processing Systems(NeurIPS)*, Dec 2020.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343. PMLR, 10–15 Jul 2018.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. How does batch normalization help optimization? In *Proceedings of the 32nd international conference on neural information processing systems*, pp. 2488–2498, 2018.

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., and Meng, D. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

van Rooyen, B., Menon, A., and Williamson, R. C. Learning with symmetric label noise: The importance of being unhinged. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28, pp. 10–18. Curran Associates, Inc., 2015.

Wang, F., Xiang, X., Cheng, J., and Yuille, A. L. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1041–1049, 2017.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018.

Wang, X., Kodirov, E., Hua, Y., and Robertson, N. Improving mae against cce under label noise. *ArXiv*, abs/1903.12141, 2019a.

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In *IEEE International Conference on Computer Vision*, 2019b.

Zhang, S., Li, Z., Yan, S., He, X., and Sun, J. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021a.

Zhang, Y., Hooi, B., Hong, L., and Feng, J. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv preprint arXiv:2107.09249*, 2021b.

Zhang, Y., Kang, B., Hooi, B., Yan, S., and Feng, J. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021c.

Zhang, Y., Kang, B., Hooi, B., Yan, S., and Feng, J. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021d.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 8778–8788. Curran Associates, Inc., 2018.

Zhong, Z., Cui, J., Liu, S., and Jia, J. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16489–16498, 2021.

Zhou, B., Cui, Q., Wei, X.-S., and Chen, Z.-M. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9716–9725, 2020a. doi: 10.1109/CVPR42600.2020.00974.

Zhou, B., Cui, Q., Wei, X.-S., and Chen, Z.-M. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9719–9728, 2020b.

Zhou, X., Liu, X., Jiang, J., Gao, X., and Ji, X. Asymmetric loss functions for learning with noisy labels. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12846–12856. PMLR, 18–24 Jul 2021a.

Zhou, X., Liu, X., Wang, C., Zhai, D., Jiang, J., and Ji, X. Learning with noisy labels via sparse regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 72–81, 2021b.

Zhou, X., Liu, X., Zhai, D., Jiang, J., Gao, X., and Ji, X. Learning towards the largest margins. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hqkhcFHOeKD.

Appendix for "Prototype-anchored Learning for Learning with Imperfect Data"

A. Proof for Theorems and Propositions

A.1. Proof for Lemma 3.1

Proof. According to the definition of $\gamma_{\min}$, we have

$$\begin{aligned}\arg\max_{\boldsymbol{w}}\max_{\boldsymbol{z}}\gamma_{\min} &= \arg\max_{\boldsymbol{w}}\max_{\boldsymbol{z}}\min_{i}\boldsymbol{w}_{y_i}^\top\boldsymbol{z}_i-\max_{j\neq y_i}\boldsymbol{w}_j^\top\boldsymbol{z}_i\\ &= \arg\max_{\boldsymbol{w}}\min_{i}\max_{\boldsymbol{z}_i}\boldsymbol{w}_{y_i}^\top\boldsymbol{z}_i-\max_{j\neq y_i}\boldsymbol{w}_j^\top\boldsymbol{z}_i\\ &= \arg\max_{\boldsymbol{w}}\min_{i}\max_{\boldsymbol{z}_i}\boldsymbol{w}_{y_i}^\top\boldsymbol{z}_i-\boldsymbol{w}_k^\top\boldsymbol{z}_i\\ &= \arg\max_{\boldsymbol{w}}\min_{i}\|\boldsymbol{w}_{y_i}-\boldsymbol{w}_k\|_2\end{aligned},$$

where $k\in\arg\max_{j\neq y_i}\boldsymbol{w}_j^\top\boldsymbol{z}_i$, and $\boldsymbol{z}_i=\frac{\boldsymbol{w}_{y_i}-\boldsymbol{w}_k}{\|\boldsymbol{w}_{y_i}-\boldsymbol{w}_k\|_2}$. Notice that $\boldsymbol{w}_k^\top\boldsymbol{z}_i=-\frac{1}{2}\|\boldsymbol{w}_{y_i}-\boldsymbol{w}_k\|_2$, then $k=\arg\min_{j\neq y_i}\|\boldsymbol{w}_{y_i}^\top-\boldsymbol{w}_j\|_2$. Therefore, we have

$$\arg\max_{\boldsymbol{w}}\max_{\boldsymbol{z}}\gamma_{\min}=\max_{\boldsymbol{w}}\min_{i}\min_{k\neq y_i}\|\boldsymbol{w}_{y_i}-\boldsymbol{w}_k\|_2=\arg\max_{\boldsymbol{w}}\min_{i\neq j}\|\boldsymbol{w}_i-\boldsymbol{w}_j\|_2,$$

i.e., maximizing $\gamma_{\min}$ will provide the solution of the Tammes Problem. According to the proof of Borodachov et al. (2019, Theorem 3.3.1), we have the above solution satisfies that $\boldsymbol{w}_i^\top\boldsymbol{w}_j=\frac{-1}{k-1},\forall i\neq j$. Then we have $\boldsymbol{z}_i=\arg\max_{\boldsymbol{z}\in\mathbb{S}^{d-1}}\boldsymbol{w}_{y_i}^\top\boldsymbol{z}-\max_{j\neq y}\boldsymbol{w}_j^\top\boldsymbol{z}=\boldsymbol{w}_{y_i}$, and $\gamma_{\min}=-\frac{k}{k-1}$.

□

A.2. Proof for Theorem 3.2

Proof. Since the function $\exp$ is strictly convex, using the Jensen's inequality, we have

$$\begin{aligned}L&=\frac{1}{N}\sum_{i=1}^{N}-\log\frac{\exp(s\boldsymbol{w}_{y_i}^\top\boldsymbol{z}_i+\alpha)}{\exp(s\boldsymbol{w}_{y_i}^\top\boldsymbol{z}_i+\alpha)+\sum_{j\neq i}\exp(s\boldsymbol{w}_j^\top\boldsymbol{z}_i)}\\ &\geq\frac{1}{N}\sum_{i=1}^{N}-\log\frac{\exp(s\boldsymbol{w}_{y_i}^\top\boldsymbol{z}_i+\alpha)}{\exp(s\boldsymbol{w}_{y_i}^\top\boldsymbol{z}_i+\alpha)+(k-1)\exp(\frac{s}{k-1}\sum_{j\neq i}\boldsymbol{w}_j^\top\boldsymbol{z}_i)}\end{aligned}$$

Let $\overline{\boldsymbol{w}}=\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{w}_i$, and $\sigma=\frac{k}{k-1}$, then we have

$$\begin{aligned}L&\geq\frac{1}{N}\sum_{i=1}^{N}\log\left[1+(k-1)\exp(s\sigma(\overline{\boldsymbol{w}}-\boldsymbol{w}_{y_i})^\top\boldsymbol{z}_i-\alpha\right]\\ &\geq\frac{1}{N}\sum_{i=1}^{N}\log[1+(k-1)\exp(-s\sigma\|\overline{\boldsymbol{w}}-\boldsymbol{w}_{y_i}\|_2-\alpha)]\,,\\ &=\frac{1}{k}\sum_{i=1}^{k}\log[1+(k-1)\exp(-s\sigma\|\overline{\boldsymbol{w}}-\boldsymbol{w}_i\|_2-\alpha)]\end{aligned}$$

where we use the facts that $(\overline{\boldsymbol{w}} - \boldsymbol{w}_{y_i})^\top \boldsymbol{z}_i \geq -\|\overline{\boldsymbol{w}} - \boldsymbol{w}_i\|_2$ when $\boldsymbol{z}_i \in \mathbb{S}^{d-1}$. Due the convexity of the function $\log[1+\exp(ax+b)]$ $(a>0)$, we use the Jensen's inequality and obtain that

$$\begin{aligned}
L &\geq \log\left[1+(k-1)\exp\left(-\frac{s\sigma}{k}\sum_{i=1}^{k}\|\overline{\boldsymbol{w}}-\boldsymbol{w}_i\|_2-\alpha\right)\right] \\
&\geq \log\left[1+(k-1)\exp\left(-\frac{s}{k-1}\sqrt{k\sum_{i=1}^{k}\|\overline{\boldsymbol{w}}-\boldsymbol{w}_i\|_2^2}-\alpha\right)\right], \\
&= \log\left[1+(k-1)\exp\left(-\frac{s}{k-1}\sqrt{k(k-k\|\overline{\boldsymbol{w}}\|_2^2\|_2^2)}-\alpha\right)\right] \\
&\geq \log[1+(k-1)\exp(-s\sigma-\alpha)]
\end{aligned}$$

where in the second inequality we used the Cauchy–Schwarz inequality, and the third inequality is based on that $k(k-k\|\overline{\boldsymbol{w}}\|_2^2\|_2^2 \leq k^2$. According to the above derivation, the equality holds if and only if $\forall i$, $\boldsymbol{w}_1^\top \boldsymbol{z}_i = ... = \boldsymbol{w}_{y_i-1}^\top \boldsymbol{z}_i = \boldsymbol{w}_{y_i+1}^\top \boldsymbol{z}_i = ... = \boldsymbol{w}_k^\top \boldsymbol{z}_i$, $\boldsymbol{w}_{y_i}^\top \boldsymbol{z}_i = 1$, $\boldsymbol{z}_i = -\frac{\overline{\boldsymbol{w}}-\boldsymbol{w}_{y_i}}{\|\overline{\boldsymbol{w}}-\boldsymbol{w}_{y_i}\|_2}$, $\|\overline{\boldsymbol{w}}-\boldsymbol{w}_1\|_2 = ... = \|\overline{\boldsymbol{w}}-\boldsymbol{w}_k\|_2$, and $\overline{\boldsymbol{w}}=0$. The condition can be simplified as $\forall i \neq j$, $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}$, and $\boldsymbol{z}_i = \boldsymbol{w}_{y_i}$ when $2 \leq d$ and $2 \leq k \leq d+1$. □

A.3. Proof for Theorem 3.3

Proof. Considering the binary classification problem, the conditional risk with respect to the margin-based loss in Eq. 2 is

$$\begin{aligned}
C(\eta_{\boldsymbol{x}}; \phi_\Theta(\boldsymbol{x}), \boldsymbol{w}_+, \boldsymbol{w}_-) = &-\eta_{\boldsymbol{x}} \log \frac{\exp(s\boldsymbol{w}_+^{\mathrm{T}}\phi_\Theta(\boldsymbol{x})+\alpha_+)}{\exp(s\boldsymbol{w}_+^{\mathrm{T}}\phi_\Theta(\boldsymbol{x})+\alpha_+)+\exp(s\boldsymbol{w}_-^{\mathrm{T}}\phi_\Theta(\boldsymbol{x}))} \\
&-(1-\eta_{\boldsymbol{x}}) \log \frac{\exp(s\boldsymbol{w}_-^{\mathrm{T}}\phi_\Theta(\boldsymbol{x})+\alpha_-)}{\exp(s\boldsymbol{w}_+^{\mathrm{T}}\phi_\Theta(\boldsymbol{x}))+\exp(s\boldsymbol{w}_-^{\mathrm{T}}\phi_\Theta(\boldsymbol{x})+\alpha_-)}
\end{aligned} \tag{9}$$

where $\boldsymbol{z} = \phi_\Theta(\boldsymbol{x}), \boldsymbol{w}_+, \boldsymbol{w}_- \in \mathbb{S}^{d-1}$, $\alpha_+ = \frac{C}{\pi_+^{1/4}}$, and $\alpha_- = \frac{C}{\pi_-^{1/4}}$. The Lagrangian function is $\mathcal{L}(\eta_{\boldsymbol{x}}; \boldsymbol{z}, \boldsymbol{w}_+, \boldsymbol{w}_-) = C(\eta_{\boldsymbol{x}}; \phi_\Theta(\boldsymbol{x}), \boldsymbol{w}_+, \boldsymbol{w}_-) + \lambda(\|\boldsymbol{z}\|_2^2 - 1)$, and the gradient of $\mathcal{L}$ with respect to $\boldsymbol{z}$ can be derived as follows:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\eta_{\boldsymbol{x}}; \boldsymbol{z}, \boldsymbol{w}_+, \boldsymbol{w}_-)}{\partial \boldsymbol{z}} =& 2\lambda\boldsymbol{z} + \frac{s\eta_{\boldsymbol{x}}\exp(s(\boldsymbol{w}_- - \boldsymbol{w}_+)^{\mathrm{T}}\boldsymbol{z}-\alpha_+)}{1+\exp(s(\boldsymbol{w}_- - \boldsymbol{w}_+)^{\mathrm{T}}\boldsymbol{z}-\alpha_+)}(\boldsymbol{w}_- - \boldsymbol{w}_+) \\
&+ \frac{s(1-\eta_{\boldsymbol{x}})\exp(s(\boldsymbol{w}_+ - \boldsymbol{w}_-)^{\mathrm{T}}\boldsymbol{z}-\alpha_-)}{1+\exp(s(\boldsymbol{w}_+ - \boldsymbol{w}_-)^{\mathrm{T}}\boldsymbol{z}-\alpha_-)}(\boldsymbol{w}_+ - \boldsymbol{w}_-)
\end{aligned}, \tag{10}$$

therefore, the minimum of $C(\eta_{\boldsymbol{x}}; \phi_\Theta, \boldsymbol{w}_+, \boldsymbol{w}_-)$ will be obtained at $\frac{\boldsymbol{w}_+ - \boldsymbol{w}_-}{\|\boldsymbol{w}_+ - \boldsymbol{w}_-\|_2}$ or $-\frac{\boldsymbol{w}_+ - \boldsymbol{w}_-}{\|\boldsymbol{w}_+ - \boldsymbol{w}_-\|_2}$.

As can be noticed, the prediction followed LDAM is +1 if $\frac{\exp(s\boldsymbol{w}_+^\top \boldsymbol{z})}{\exp(s\boldsymbol{w}_+^\top \boldsymbol{z})+\exp(s\boldsymbol{w}_-^\top \boldsymbol{z})} > \frac{\exp(s\boldsymbol{w}_-^\top \boldsymbol{z})}{\exp(s\boldsymbol{w}_+^\top \boldsymbol{z})+\exp(s\boldsymbol{w}_-^\top \boldsymbol{z})}$, and vice versa. This indicates that the Bayes-optimal prediction of LDAM is dependent on the sign of $(\boldsymbol{w}_+ - \boldsymbol{w}_-)^{\mathrm{T}}\boldsymbol{z}$ with $\boldsymbol{z} = \arg\min_{\boldsymbol{z}\in\mathbb{S}^{d-1}} C(\eta_{\boldsymbol{x}}; \boldsymbol{z}, \boldsymbol{w}_+, \boldsymbol{w}_-)$. In the following, we are going to confirm the sign.

Assume that the minimizer of $C(\eta_{\boldsymbol{x}}; \phi_\Theta, \boldsymbol{w}_+, \boldsymbol{w}_-)$ is $t \cdot \frac{\boldsymbol{w}_+ - \boldsymbol{w}_-}{\|\boldsymbol{w}_+ - \boldsymbol{w}_-\|_2}$, where $t \in \{-1, +1\}$. Then the Bayes-optimal prediction of LDAM depends on the value of t. Actually, the value of t relies not only on $\eta_{\boldsymbol{x}}$, but also on α_+ and α_-.

Let $r = s\|\boldsymbol{w}_+ - \boldsymbol{w}_-\|_2 \in [0, 2s]$, and $g(t; \eta_{\boldsymbol{x}}, r, \alpha_+, \alpha_-) = \eta_{\boldsymbol{x}} \log[1+\exp(-tr-\alpha_+)] + (1-\eta_{\boldsymbol{x}})\log[1+\exp(tr-\alpha_-)]$, then we have

$$\begin{aligned}
t &= \mathrm{sign}(g(-1; \eta_{\boldsymbol{x}}, r, \alpha_+, \alpha_-) - g(1; \eta_{\boldsymbol{x}}, r, \alpha_+, \alpha_-)) \\
&= \mathrm{sign}\left(\eta_{\boldsymbol{x}} \log\frac{1+\exp(r-\alpha_+)}{1+\exp(-r-\alpha_+)} + (1-\eta_{\boldsymbol{x}})\log\frac{1+\exp(-r-\alpha_-)}{1+\exp(r-\alpha_-)}\right) \\
&= \mathrm{sign}\left(\eta_{\boldsymbol{x}} - \frac{\log\frac{1+\exp(r-\alpha_-)}{1+\exp(-r-\alpha_-)}}{\log\frac{1+\exp(r-\alpha_+)}{1+\exp(-r-\alpha_+)} + \log\frac{1+\exp(r-\alpha_-)}{1+\exp(-r-\alpha_-)}}\right)
\end{aligned} \tag{11}$$

where $\mathrm{sign}(\cdot)$ denotes the signum function.

As can be seen, if $a_+ = a_-$, then the Bayes-optimal prediction of LDAM is $t = \text{sign}(\eta_{\boldsymbol{x}} - \frac{1}{2})$, which is the Bayes-optimal classifier, then LDAM is calibrated. However, under imbalanced data distribution, the definition is $\alpha_+ \neq \alpha_-$, we know that $\frac{\log \frac{1+\exp(r-\alpha_-)}{1+\exp(-r-\alpha_-)}}{\log \frac{1+\exp(r-\alpha_+)}{1+\exp(-r-\alpha_+)} + \log \frac{1+\exp(r-\alpha_-)}{1+\exp(-r-\alpha_-)}}$ will not be equal to $\frac{1}{2}$, so LDAM is not calibrated. □

A.4. Proof for Theorem 4.1

Proof. For the margin-based loss $L_{\boldsymbol{\alpha}} = -\log \frac{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z}+\alpha_y)}{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z}+\alpha_y)+\sum_{j\neq y}\exp(s\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{z})}$, we have

$$
\begin{aligned}
L_{\boldsymbol{\alpha}}(\boldsymbol{z}, y) &= -\log \frac{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} + \alpha_y)}{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} + \alpha_y) + \sum_{j\neq y}\exp(s\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{z})} \\
&\geq -\log \frac{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} + \alpha_y)}{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} + \alpha_y) + (k-1)\exp[\frac{s}{k-1}(\sum_{j\neq y}\boldsymbol{w}_j)^{\mathrm{T}}\boldsymbol{z}]} \\
&= -\log \frac{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} + \alpha_y)}{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} + \alpha_y) + (k-1)\exp(-\frac{s}{k-1}\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z})} \\
&= \log\left[1 + (k-1)\exp\left(-\frac{sk}{k-1}\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} - \alpha_y\right)\right] \\
&\geq \log\left[1 + (k-1)\exp\left(-\frac{sk}{k-1} - \alpha_y\right)\right]
\end{aligned}
$$

where in the first inequality we used the Jensen's inequality when $\exp(\cdot)$ is convex, the second equality is according to the condition that $\sum_{j=1}^{k}\boldsymbol{w}_j = 0$ since $\boldsymbol{w}_i^{\top}\boldsymbol{w}_j = \frac{-1}{k-1}, \forall i \neq j$, and the last equality comes from the fact that $\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} \leq 1$.

Therefore, we have the lower bound of the risk $\frac{1}{N}\sum_{i=1}^{N} L_{\boldsymbol{\alpha}}(\boldsymbol{z}_i, y_i; W) \geq \frac{1}{N}\sum_{i=1}^{N}\left[1 + (k-1)\exp\left(-\frac{sk}{k-1} - \alpha_y\right)\right]$, where the equality holds if and only if $\forall i$, $\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{z}_i = \cdots = \boldsymbol{w}_{y_i-1}^{\mathrm{T}}\boldsymbol{z}_i = \boldsymbol{w}_{y_i+1}^{\mathrm{T}}\boldsymbol{z}_i = \cdots = \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{z}_i$, and $\boldsymbol{w}_{y_i}^{\mathrm{T}}\boldsymbol{z}_i = 1$. Then we have $\gamma_{\min} = \frac{k}{k-1}$. □

A.5. Proof for Theorem Theorem 4.2

Proof. Similar to the proof for Theorem 3.1, for the margin-based loss $L_{\boldsymbol{\alpha}} = -\log \frac{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z}+\alpha_y)}{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z}+\alpha_y)+\sum_{j\neq y}\exp(s\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{z})}$, we have

$$
\begin{aligned}
L_{\boldsymbol{\alpha}}(\boldsymbol{z}, y) &\geq \log\left[1 + (k-1)\exp\left(-\frac{sk}{k-1}\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} - \alpha_y\right)\right] \\
&\geq \log\left[1 + (k-1)\exp\left(-\frac{sk\|\boldsymbol{z}\|_2}{k-1} - \alpha_y\right)\right] \\
&\geq \log\left[1 + (k-1)\exp\left(-\frac{skB}{k-1} - \alpha_y\right)\right]
\end{aligned}
$$

where the derivation above comes from the fact that $\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} \leq \|\boldsymbol{z}\|_2 \leq B$.

Therefore, we have the lower bound of the risk $\frac{1}{N}\sum_{i=1}^{N} L_{\boldsymbol{\alpha}}(\boldsymbol{z}_i, y_i; W) \geq \frac{1}{N}\sum_{i=1}^{N}\left[1 + (k-1)\exp\left(-\frac{skB}{k-1} - \alpha_y\right)\right]$, where the equality holds if and only if $\forall i$, $\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{z}_i = \cdots = \boldsymbol{w}_{y_i-1}^{\mathrm{T}}\boldsymbol{z}_i = \boldsymbol{w}_{y_i+1}^{\mathrm{T}}\boldsymbol{z}_i = \cdots = \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{z}_i$, $\frac{\boldsymbol{w}_{y_i}^{\mathrm{T}}\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|_2} = 1$, and $\|\boldsymbol{z}_i\|_2 = B$ Then we have $\gamma_{\min} = \frac{kB}{k-1}$. □

A.6. Proof for Theorem 4.4

Proof. Recall that for any f, we have $L(f(\boldsymbol{x}), i) = L(W^\top \boldsymbol{\phi}_\Theta(\boldsymbol{x}), i)$, $R_L(f) = \mathbb{E}_{\boldsymbol{x},y} L(f(\boldsymbol{x}), y)$, and $L_W(\boldsymbol{\phi}_\Theta(\boldsymbol{x})) = \sum_{i=1}^k L(W^\top \boldsymbol{\phi}_\Theta(\boldsymbol{x}), i)$. Then for symmetric or uniform label noise, the noisy risk $R_L^\eta(f)$ satisfies that

$$\begin{aligned} R_L^\eta(f) &= \mathbb{E}_{\boldsymbol{x},\tilde{y}} L(f(\boldsymbol{x}), \tilde{y}) \\ &= \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{y|\boldsymbol{x}} \mathbb{E}_{\tilde{y}|\boldsymbol{x},y} L(f(\boldsymbol{x}), \tilde{y}) \\ &= \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{y|\boldsymbol{x}} \left[(1-\eta) L(f(\boldsymbol{x}, y) + \frac{\eta}{k-1} \sum_{i \neq y} L(f(\boldsymbol{x}), y) \right], \\ &= (1-\eta) R_L(f) + \frac{\eta}{k-1} \left[\mathbb{E}_{\boldsymbol{x}} L_W(\boldsymbol{\phi}_\Theta(\boldsymbol{x})) - R_L(f) \right] \end{aligned}$$

that is,

$$(1 - \frac{\eta k}{k-1}) R_L(f) = R_L^\eta(f) - \frac{\eta}{k-1} \mathbb{E}_{\boldsymbol{x}} L_W(\boldsymbol{\phi}_\Theta(\boldsymbol{x})).$$

For $f^* \in \arg\min_{f \in \mathcal{H}} R_L(f)$ and $\hat{f} \in \arg\min_{f \in gH} R_L^\eta(f)$, we then obtain

$$\begin{aligned} (1 - \frac{\eta k}{k-1})[R_L(\hat{f}) - R_L(f^*)] &= R_L^\eta(\hat{f}) - R_L^\eta(f^*) - \frac{\eta}{k-1} [\mathbb{E}_{\boldsymbol{x}} L_W(\boldsymbol{\phi}_{\hat{\Theta}}(\boldsymbol{x})) - \mathbb{E}_{\boldsymbol{x}} L_W(\boldsymbol{\phi}_{\Theta^*}(\boldsymbol{x}))] \\ &\leq \frac{\eta}{k-1} \mathbb{E}_{\boldsymbol{x}} |L_W(\boldsymbol{\phi}_{\hat{\Theta}}(\boldsymbol{x})) - L_W(\boldsymbol{\phi}_{\Theta^*}(\boldsymbol{x}))| \\ &\leq \frac{\eta\lambda}{k-1} \mathbb{E}_{\boldsymbol{x}} \|\boldsymbol{\phi}_{\hat{\Theta}}(\boldsymbol{x}) - \boldsymbol{\phi}_{\Theta^*}(\boldsymbol{x})\|_2 \leq \frac{2\eta\lambda B}{k-1} \end{aligned},$$

where we use the facts that $R_L^\eta(\hat{f}) - R_L^\eta(f^*) \leq 0$, $L_W(\boldsymbol{z})$ is λ-Lipschitz, and $\|\boldsymbol{\phi}_{\hat{\Theta}}(\boldsymbol{x}) - \boldsymbol{\phi}_{\Theta^*}(\boldsymbol{x})\|_2 \leq 2B$ when $\|\boldsymbol{\phi}_\Theta(\boldsymbol{x})\|_2 = B$. Therefore, rearranging the above equation will obtain the risk bound in Eq. (7). □

A.7. Proof for Proposition 4.5

Proof. For the softmax loss in Eq. (1), we have

$$L_W(\boldsymbol{z}) = -\sum_{i=1}^k \log \frac{\exp(\boldsymbol{w}_i^\top \boldsymbol{z})}{\sum_{j=1}^k \exp(\boldsymbol{w}_j^\top \boldsymbol{z})}, \tag{12}$$

then the derivative of $L_W(\boldsymbol{z})$ with respect to $\boldsymbol{z}$ is

$$\frac{\partial L_W(\boldsymbol{z})}{\partial \boldsymbol{z}} = \sum_{i=1}^k \left[\boldsymbol{w}_i - \sum_{j=1}^k \frac{\exp(\boldsymbol{w}_j^\top \boldsymbol{z})}{\sum_{t=1}^k \exp(\boldsymbol{w}_t^\top \boldsymbol{z})} \boldsymbol{w}_j \right],$$

since $\boldsymbol{w}_1, ..., \boldsymbol{w}_k \in \mathbb{S}^{d-1}$ satisfy $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}$, then $\sum_{i=1}^k \boldsymbol{w}_i = \boldsymbol{0}$ and $\frac{\partial L_W(\boldsymbol{z})}{\partial \boldsymbol{z}} = k \sum_{j=1}^k \frac{\exp(\boldsymbol{w}_i^\top \boldsymbol{z})}{\sum_{i=1}^k \exp(\boldsymbol{w}_j^\top \boldsymbol{z})} \boldsymbol{w}_i$. The Lipschitz constant of $L_W(\boldsymbol{z})$ will depend on the upper bound of $\|\frac{\partial L_W(\boldsymbol{z})}{\partial \boldsymbol{z}}\|_2$, and this bound exists when $\boldsymbol{w}_i \in \mathbb{S}^{d-1}$ and $\|\boldsymbol{z}\|_2 \leq B$. Actually, let $\alpha_i = \frac{\exp(\boldsymbol{w}_i^\top \boldsymbol{z})}{\sum_{j=1}^k \exp(\boldsymbol{w}_j^\top \boldsymbol{z})}$, then $\alpha_i \in [\frac{1}{1+(k-1)\exp(\frac{kB}{k-1})}, \frac{1}{1+(k-1)\exp(-\frac{kB}{k-1})}]$, $\sum_{i=1}^k \alpha_i = 1$, and we have

$$\begin{aligned} \left\| \frac{\partial L_W(\boldsymbol{z})}{\partial \boldsymbol{z}} \right\|_2^2 &= k^2 \| \sum_{i=1}^k \alpha_i \boldsymbol{w}_i \|_2^2 = k^2 \sum_{i=1}^k \sum_{j=1}^k \alpha_i \alpha_j \boldsymbol{w}_i^\top \boldsymbol{w}_j \\ &= k^2 \left[\sum_{i=1}^k \alpha_i^2 - \frac{1}{k-1} \sum_{i \neq j} \alpha_i \alpha_j \right] = k^2 \left[\frac{k}{k-1} \sum_{i=1}^k \alpha_i^2 - \frac{1}{k-1} \right] \\ &\leq k^2 \left[\frac{k}{k-1} \left(\frac{1}{[1+(k-1)\exp(-\frac{kB}{k-1})]^2} + \frac{k-1}{[k-1+\exp(\frac{kB}{k-1})]^2} \right) - \frac{1}{k-1} \right], \\ &= k^2 \left(\frac{1 - \exp(-\frac{kB}{k-1})}{1 + (k-1)\exp(-\frac{kB}{k-1})} \right)^2 \end{aligned}$$

thus, the Lipschitz constant of $L_W(\boldsymbol{z})$ is $\lambda_{\text{PAL}} = \frac{k(1-\exp(-\frac{kB}{k-1}))}{1+(k-1)\exp(-\frac{kB}{k-1})}$, then we have the following risk according to Theorem 4.4:

$$R_L(\hat{f}) - R_L(f^*) \leq \frac{2c\eta k(1-t)B}{k-1+t(k-1)^2}, \tag{13}$$

where $c = \frac{k-1}{(1-\eta)k-1}$, and $t = \exp(-\frac{kB}{k-1})$. □

A.7.1. More Analysis about Other Losses

For a general loss, we usually have $L_i = \ell(\alpha_i)$, where $\alpha_i = \frac{\exp(\boldsymbol{w}_i^\top \boldsymbol{z})}{\sum_{j=1}^k \exp(\boldsymbol{w}_j^\top \boldsymbol{z})} \in [0,1]$ denotes the predicted probability of class i. Then the derivative of $L = \sum_{i=1}^k L_i$ with respect to $\boldsymbol{z}$ can be derived as

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{z}} &= \sum_{i=1}^k \frac{\partial L_i}{\partial \alpha_i}\frac{\partial \alpha_i}{\partial \boldsymbol{z}} = \sum_{i=1}^k \frac{\partial \ell(\alpha_i)}{\partial \alpha_i}\cdot \alpha_i \left(\boldsymbol{w}_i - \sum_{j=1}^k \alpha_j \boldsymbol{w}_j\right) \\ &= \left(\sum_{i=1}^k \alpha_i \cdot \frac{\partial \ell(\alpha_i)}{\partial \alpha_i}\right)\left(\sum_{i=1}^k \frac{\alpha_i \cdot \frac{\partial \ell(\alpha_i)}{\partial \alpha_i}}{\sum_{j=1}^k \alpha_i \cdot \frac{\partial \ell(\alpha_j)}{\partial \alpha_j}}\boldsymbol{w}_i - \sum_{i=1}^k \alpha_i \boldsymbol{w}_i\right).\end{aligned} \tag{14}$$

As can be seen, to make L be Lipschitz continuous, the key point is to guarantee that $\sum_{i=1}^k \alpha_i \cdot \frac{\partial \ell(\alpha_i)}{\partial \alpha_i}$ and $\frac{\alpha_i \cdot \frac{\partial \ell(\alpha_i)}{\partial \alpha_i}}{\sum_{j=1}^k \alpha_i \cdot \frac{\partial \ell(\alpha_j)}{\partial \alpha_j}}$ are also bounded. In the following, we will analyze GCE (Zhang & Sabuncu, 2018) and Focal loss (Lin et al., 2017).

GCE. GCE is formulated as $\ell_i(\alpha_i) = \frac{1-\alpha_i^q}{q}$ $(0 < q \leq 1)$. The derivative of ℓ_i w.r.t α_i is $\frac{\partial \ell_i}{\partial \alpha_i} = -\alpha_i^{q-1}$, then $\frac{\alpha_i \cdot \frac{\partial \ell(\alpha_i)}{\partial \alpha_i}}{\sum_{j=1}^k \alpha_i \cdot \frac{\partial \ell(\alpha_j)}{\partial \alpha_j}} = \frac{\alpha_i^q}{\sum_{j=1}^k \alpha_j^q} \in [0,1]$ is bounded, and we also have $\sum_{i=1}^k \alpha_i \cdot \frac{\partial \ell(\alpha_i)}{\partial \alpha_i} = -\sum_{i=1}^k \alpha_i^q \in [-1,0]$. Thus, GCE loss is Lipschitz continuous.

Focal Loss. Focal loss is formulated as $\ell_i = -(1-\alpha_i)^\gamma \log \alpha_i$ $(\gamma > 0)$. We have $\alpha_i \frac{\partial \ell_i}{\partial \alpha_i} = -(1-\alpha_i)^{\gamma-1}(1-\alpha_i - \gamma \alpha_i \log \alpha_i) < 0$. As can be seen, when $\alpha_i \to 0$ $(\gamma > 0)$ or $\alpha_i \to 1$ $(\gamma < 1)$, $\alpha_i \frac{\partial \ell_i}{\partial \alpha_i}$ will be unbounded. This issue can be alleviated when $\boldsymbol{w}_1, ..., \boldsymbol{w}_k \in \mathbb{S}^{d-1}$ satisfy $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \frac{-1}{k-1}$, since we have $\alpha_i \in [\frac{1}{1+(k-1)\exp(\frac{kB}{k-1})}, \frac{1}{1+(k-1)\exp(-\frac{kB}{k-1})}]$, where $\|\boldsymbol{z}\|_2 \leq B$.

A.7.2. About the Tighter Bound Indicated by PAL

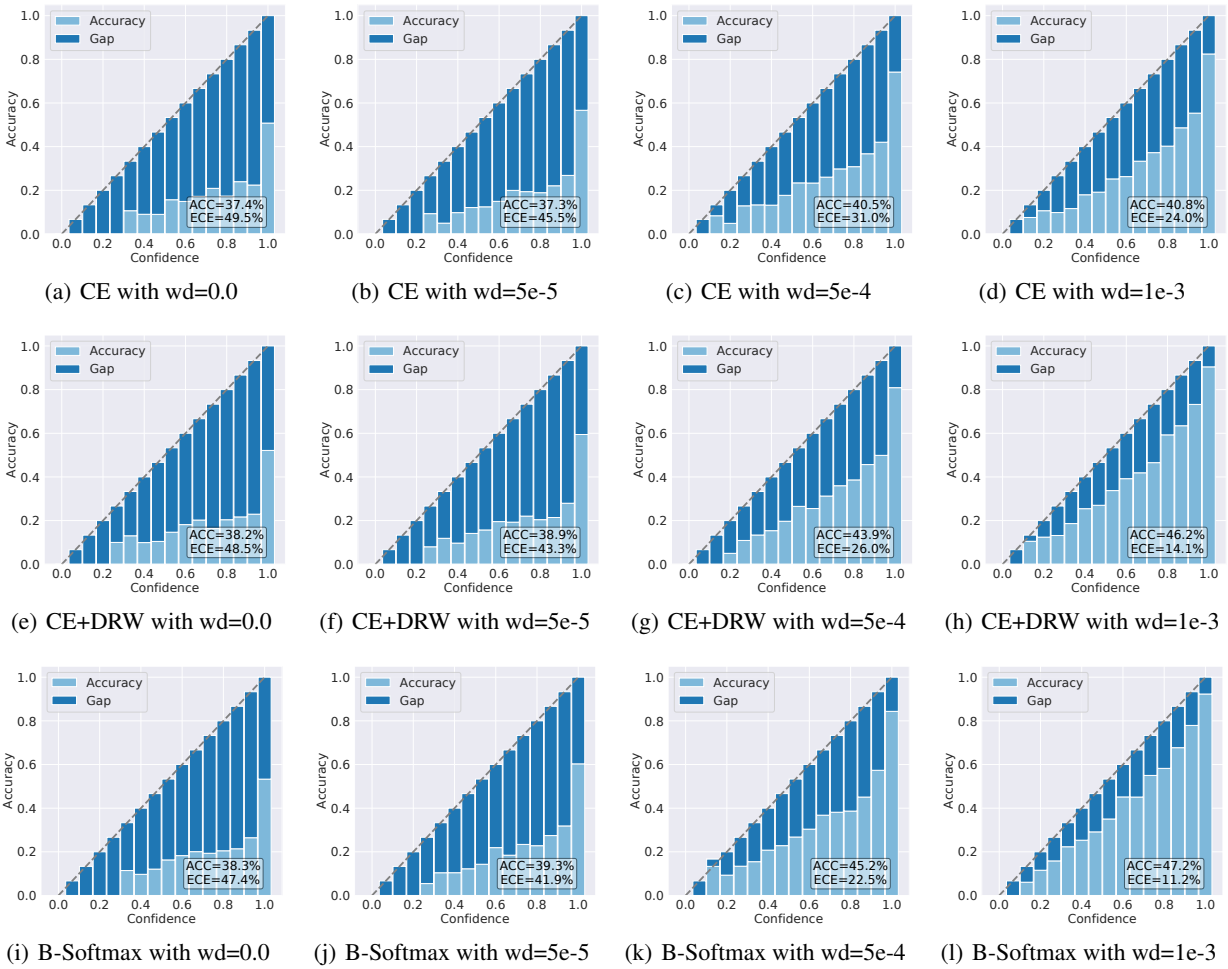
In this section, we will prove that PAL will lead to a smaller Lipschitz constant of L_W in Eq. (12) than unnormalized or unacnhored cases. Note that $\lambda_{\text{PAL}} = \frac{k(1-\exp(-\frac{kB}{k-1}))}{1+(k-1)\exp(-\frac{kB}{k-1})}$, we have:

1. **When $\boldsymbol{w}$ is unnormalized**. The Lipschitz constant of L_W is $\sup_{W,\boldsymbol{z}} \left\| \sum_{i=1}^k \boldsymbol{w}_i - k\sum_{j=1}^k \frac{\exp(\boldsymbol{w}_j^\top \boldsymbol{z})}{\sum_{t=1}^k \exp(\boldsymbol{w}_t^\top \boldsymbol{z})}\boldsymbol{w}_j \right\|_2$, which can be infinitely large as $\boldsymbol{w}_i = t\boldsymbol{z}$ $(t \to \infty)$ and $\boldsymbol{w}_j$ $(j \neq i)$ is fixed.

2. **When $\boldsymbol{w}_i$ is normalized but unanchored (*w.l.o.g*, $\|\boldsymbol{w}_i\|_2 = 1$ and $\boldsymbol{z}$ is unnormalized).** We have the Lipschitz constant $\sup_{W,\boldsymbol{z}} \|\frac{\partial L_W(\boldsymbol{z})}{\partial \boldsymbol{z}}\|_2 \geq k$ (which can be obtained at $\sum_{i=1}^k \boldsymbol{w}_i = 0$ and $\max_i \frac{\exp(\boldsymbol{w}_i^T \boldsymbol{z})}{\sum_{j=1}^k \exp(\boldsymbol{w}_j^T \boldsymbol{z})} = 1$) that is larger than λ_{PAL};

3. **When both $\boldsymbol{w}_i$ and $\boldsymbol{z}$ are normalized and unanchored ($\|\boldsymbol{w}_i\|_2 = 1$, $\|\boldsymbol{z}\|_2 = B$).** Considering that $\boldsymbol{w}_1 = \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}$, and $\boldsymbol{w}_2 = ...\boldsymbol{w}_k = -\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}$, we have

$$\left\|\frac{\partial L_W(\boldsymbol{z})}{\partial \boldsymbol{z}}\right\|_2 = \frac{2(\exp(2B)-1)}{\frac{\exp(2B)}{k-1}+1} > \lambda_{\text{PAL}},$$



(a) CE with wd=0.0 (b) CE with wd=5e-5 (c) CE with wd=5e-4 (d) CE with wd=1e-3

(e) CE+DRW with wd=0.0 (f) CE+DRW with wd=5e-5 (g) CE+DRW with wd=5e-4 (h) CE+DRW with wd=1e-3

(i) B-Softmax with wd=0.0 (j) B-Softmax with wd=5e-5 (k) B-Softmax with wd=5e-4 (l) B-Softmax with wd=1e-3

Figure 6. Reliability diagrams of ResNet-32 trained on CIFAR-100-LT with imbalance ratio 100 under different weight decays (wds). (a-d), (e-h) and (i-l) denotes the results by CE, CE+DRW and balanced Softmax (Ren et al., 2020), respectively.

where $\frac{2(\exp(2B)-1)}{\frac{\exp(2B)}{k-1}+1} > \frac{k(1-\exp(-\frac{kB}{k-1}))}{1+(k-1)\exp(-\frac{kB}{k-1})}$ is equivalent to

$$\begin{aligned}
&(\exp(2B)-1)\left[1+(k-1)\exp\left(\frac{-k}{k-1}B\right)\right] > \frac{k}{2}\left[1-\exp\left(\frac{-k}{k-1}B\right)\right]\left[\frac{\exp(2B)}{k-1}+1\right] \\
\Leftrightarrow &\left(1-\frac{k}{2k-2}\right)\exp(2B)+\left(k-1+\frac{k}{2k-2}\right)\exp\left(\frac{k-2}{k-1}B\right) > \frac{k}{2}+1+\left(k-1-\frac{k}{2}\right)\exp\left(-\frac{k}{k-1}B\right),
\end{aligned} \tag{15}$$

since $k \geq 2$ and $B > 0$, we have LHS $> (1-\frac{k}{2k-2})+(k-1+\frac{k}{2k-2}) = k \geq \frac{k}{2}+1+\left(k-1-\frac{k}{2}\right)\exp\left(-\frac{k}{k-1}B\right) =$ RHS.

Therefore, PAL will lead to a smaller Lipschitz constant, and then a tighter risk bound.

B. More Experimental Details, Analysis and Results

B.1. Rethinking the Role of Weight Decay in Mitigating Class Imbalance

Based on the restriction of the norms of features and classifier weights, we rethink the role of weight decay that is often overlooked in mitigating class imbalance. As is well-known, a simple weight decay can improve generalization,

Table 6. Top-1 validation accuracy on CIFAR-100-LT with imbalance ratio 0.01 under different values of weight decay in the SGD optimizer. We keep everything else the same except the weight decay.

Method	weight decay									
	0	5e-5	1e-4	2e-4	3e-4	5e-4	7e-4	1e-3	5e-3	1e-2
CE	37.36	37.31	37.76	38.69	39.65	40.49	40.61	40.78	24.10	13.19
CE+DRW	38.16	38.88	39.41	40.59	42.20	43.19	44.68	46.18	31.49	18.10
Balanced Softmax	38.27	39.31	41.16	42.68	43.89	45.19	46.71	47.21	30.02	17.5

since it suppresses any irrelevant components of model weights by choosing weights with the smallest norms when solving the learning problem (Krogh & Hertz, 1992). Actually, weight decay can facilitate not only explicitly learning smaller weights of networks, but also implicitly resulting in smaller feature representations. For a network $f(x) = L_n(L_{n-1}(\cdots L_2(L_1(x); W_1); W_2) \cdots ; W_n)$ with bounded input x and weights $W_1, ..., W_n$, its output $f(x)$ is usually bounded by $W_1, ..., W_n$. Back to long-tailed recognition, **an appropriately large weight decay can mitigate class imbalance** by preventing the excessively large norms of prototypes and features that cause over confidence, since networks trained on imbalanced datasets are usually miscalibrated and over-confident due to the imbalanced composition ratio of each class (Zhong et al., 2021).

As shown in Fig. 3(a), the norms of both prototypes obviously decrease as the weight decay increases, and the distribution of feature norms gradually moves to the left in Fig. 3(b). This is beneficial to alleviate the over-confident results and improve the performance of classification. We conduct an experiment on CIFAR-100-LT with imbalance ratio 100 to validate that an appropriately large weight decay can obviously improve confidence and accuracy, where Fig. 6 shows that an appropriate large weight decay can not only largely enhance the network calibration but also greatly improve the performance for long-tailed recognition. For example, CE+DRW with weight decay $1e-3$ has obviously larger accuracy and smaller expected calibration error (ECE) than CE+DRW with weight decay $5e-5$. However, the fitting ability decreases as weight decay increases. as shown in Tab. 6.

Moreover, another easily overlooked to restrict weights and features is batch normalization (BatchNorm) (Ioffe & Szegedy, 2015), which intuitively limits the distribution of features. Santurkar et al. (2018) also prove that BatchNorm has minimax bound on weight-space Lipschitzness. Recall to the softmax loss in Eq. (1), $\boldsymbol{z}_i$ and $\boldsymbol{w}_j$ implicitly have a small norm constraint with the role of both batch normalization and weight decay, *i.e.*, $\exists r \in \mathbb{R}$, *s.t.*, $\|\boldsymbol{z}_i\|, \|\boldsymbol{w}_j\| \leq r$, which shows that the domain of $\boldsymbol{z}_i$ and $\boldsymbol{w}_j$ belongs to a 2-norm ball.

B.2. Prototypes Generation

To generate these prototypes, we randomly initialize $\{\boldsymbol{z}_i\}_{i=1}^N$ and $\{\boldsymbol{w}_i\}_{i=1}^k$ in $L_{\boldsymbol{\alpha}}$ ($\boldsymbol{\alpha} = 0$) with a balanced setting (*i.e.*, $N = k$ and $y_i = i$), and then directly minimize $L_{\boldsymbol{\alpha}}$ to obtain the optimal prototypes $\{\boldsymbol{w}_i\}_{i=1}^k$ according to Theorem 3.2. We train $\{\boldsymbol{z}_i\}_{i=1}^N$ and $\{\boldsymbol{w}_i\}_{i=1}^k$ by SGD optimizer with momentum 0.9 and initial learning rate 0.1 for 100,000 epochs. We also use cosine learning rate annealing with $T_{\max} = 20000$. The pseudo-code can be seen in Tab. 7

B.3. Details of Imbalanced Learning

Imbalanced CIFAR-10 and CIFAR-100. The original version of CIFAR-10 and CIFAR-100 contains 50,000 training images and 10,000 test images of size 32×32 with 10 and 100 classes, respectively. To create their imbalanced version, we follow the setting in (Buda et al., 2018; Cui et al., 2019; Cao et al., 2019), where we reduce the number of training examples per class, and keep the test set unchanged. To ensure that our methods apply to a variety of settings, we consider two types of imbalance: long-tailed imbalance (Cui et al., 2019) and step imbalance (Buda et al., 2018). We use the imbalance ratio ρ to denote the ratio between sample sizes of the most frequent and least frequent class, *i.e.*, $\rho = \max_i\{n_i\}/\min_i\{n_i\}$. Long-tailed imbalance utilizes an exponential decay in sample sizes across different classes.

Networks and training. We follow the settings in (Zhong et al., 2021): ResNet-32 for CIFAR-100-LT and ResNet-50 for ImageNet-LT, but we train MiSLAS on CIFAR-100 with initial learning rate 0.15 and weight decay $5e-4$ for stage-1, and initial learning rate 0.01 for stage-2. The experiments with the same settings in (Cao et al., 2019) are reported in Tab. 8.

Table 7. Pseudocode of prototypes generation in a PyTorch-like style

```
def generate_weight(n_classes, n_hiddens, use_relu=False):
    n_samples = n_classes
    scale = 5
    Z = torch.randn(n_samples, n_hiddens).cuda()
    Z.requires_grad = True
    W = torch.randn(n_classes, n_hiddens).cuda()
    W.requires_grad = True
    nn.init.kaiming_normal_(W)

    optimizer = SGD([Z, W], lr=0.1, momentum=0.9, weight_decay=1e-4)
    scheduler = CosineAnnealingLR(optimizer, T_max=20000, eta_min=0)

    criterion = nn.CrossEntropyLoss()
    for i in range(epochs):
        if use_relu:
            z = F.relu(Z)
        else:
            z = Z
        w = W
        L2_z = F.normalize(z, dim=1)
        L2_w = F.normalize(w, dim=1)
        out = F.linear(L2_z, L2_w)
        loss = criterion(out * scale, labels)
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()
        scheduler.step()
    return W
```

Table 8. Top-1 validation accuracy of ResNet-32 on imbalanced CIFAR-10 and CIFAR-100.

Dataset	Imbalanced CIFAR-10				Imbalanced CIFAR-100			
Imbalance Type	long-tailed		step		long-tailed		step	
Imbalance Ratio	100	10	100	10	100	10	100	10
CE	70.68	86.63	64.54	84.61	38.59	56.38	38.90	54.71
CE-PAL	70.60	86.39	62.81	84.58	**39.13**	**57.89**	**39.15**	**55.47**
CE-DRW	74.85	87.69	69.14	86.89	39.38	58.01	42.64	58.45
CE-DRW-PAL	74.14	**87.72**	**70.18**	**87.18**	**42.48**	**58.66**	**44.57**	**60.14**
Norm	70.65	86.15	62.39	84.41	35.68	54.79	39.72	53.76
Norm-PAL	**71.27**	**86.83**	**62.81**	**85.05**	**38.14**	**56.09**	39.28	**54.10**
Norm-DRW	73.38	87.38	68.52	86.22	38.04	56.89	42.23	56.90
Norm-DRW-PAL	**74.27**	**87.67**	**69.28**	**87.00**	**39.58**	**56.78**	**43.08**	**58.25**
LDAM	73.53	86.14	64.87	84.84	37.23	57.36	39.82	54.01
LDAM-PAL	**73.66**	**86.79**	**65.37**	**85.57**	**39.70**	**57.76**	39.09	53.30
LDAM-DRW	76.00	87.78	72.22	87.69	41.28	58.98	45.43	59.11
LDAM-DRW-PAL	**77.79**	**87.82**	**75.54**	**88.30**	**42.98**	**59.26**	**46.17**	58.69

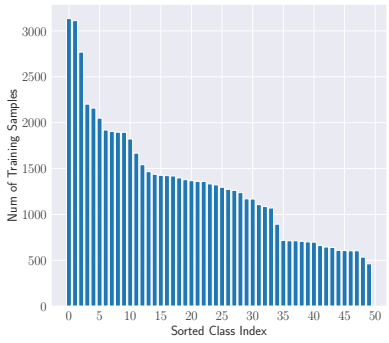


Figure 7. The label distribution on mini-WebVision.

B.4. Details of Learning with Noisy Labels

Noise generation. The noisy labels are generated following standard approaches in previous works (Ma et al., 2020; Patrini et al., 2017). For symmetric noise, we corrupt the training labels by flipping labels in each class randomly to incorrect labels to other classes with flip probability $\eta \in \{0.2, 0.3, 0.6, 0.8\}$. For asymmetric noise, we flip the labels within a specific set of classes. For MNIST, flipping $7 \rightarrow 1$, $2 \rightarrow 7$, $5 \leftrightarrow 6$, $3 \rightarrow 8$. For CIFAR-10, flipping TRUCK $\rightarrow$ AUTOMOBILE, BIRD $\rightarrow$ AIRPLANE, DEER $\rightarrow$ HORSE, CAR $\leftrightarrow$ DOG. For CIFAR-100, the 100 classes are grouped into 20 super-classes with each having 5 sub-classes, and each class are flipped within the same super-class into the next in a circular fashion.

Networks and training. We follow the settings in (Ma et al., 2020): 4-layer CNN for MNIST, an 8-layer CNN for CIFAR-10 and a ResNet-34 (He et al., 2016) for CIFAR-100. The networks are trained for 50, 120, 200 epochs for MNIST, CIFAR-10, CIFAR-100, respectively. For all the training, we use SGD optimizer with momentum 0.9 and cosine learning rate annealing. The initial learning rate is set to 0.01, 0.01 and 0.1 for MNIST, CIFAR-10 and CIFAR-100, respectively. Weight decay is set to 1×10^{-3}, 1×10^{-4} and 1×10^{-5} for MNIST, CIFAR-10 and CIFAR-100, respectively. The initial learning rate is set to 0.01 for MNIST/CIFAR-10 and 0.1 for CIFAR-100. Batch size is set to 128. Typical data augmentations including random width/height shift and horizontal flip are applied.

B.5. Details of experiments on WebVision

WebVision 1.0 (Li et al., 2017) contains 2.4 million images of real-world noisy labels, crawled from the web using 1,000 concepts in ImageNet ILSVRC12. Since the dataset is very big, for quick experiments, we follow the training setting in (Jiang et al., 2018; Ma et al., 2020) that only takes the first 50 classes of the Google resized image subset. We evaluate the trained networks on the same 50 classes of WebVision 1.0 validation set, where each concept in validation set contains 50 instances and the labels can be considered as a clean validation. ResNet-50 (He et al., 2016) is the model to be learned. The distribution of labels in mini-WebVision is shown in Fig. 7, where the imbalance ratio is $3134/462$.

We add our proposed PAL and FNPAL in some commonly-used but not sufficiently robust loss functions, such as CE, Focal loss (FL), SCE and compare them with GCE, SCE and NCE+RCE. The training details follow (Ma et al., 2020), where for each method, we train a ResNet-50 (He et al., 2016) using SGD for 250 epochs with initial learning rate 0.4, nesterov momentum 0.9 and weight decay 3×10^{-5} and batch size 256. The learning rate is multiplied by 0.97 after every epoch of training. All the images are resized to 224×224. Typical data augmentations including random width/height shift, color jittering, and random horizontal flip are applied. More detailed results than Table 5 are reported in Table 9.

Table 9. Top-1 validation accuracies (%) on WebVision validation set using different loss functions.

Method	CE	FL	GCE	SCE	NCE+RCE	**NSL**	**CE+PAL**	**CE+FNPAL**	**FL+FNPAL**	**SCE+FNPAL**
Accuracy	62.60	63.80	61.76	66.92	66.32	**69.56**	**68.92**	**69.69**	**69.64**	**69.92**

Table 10. Validation accuracies (%) of different methods on benchmark datasets with clean or symmetric label noise ($\eta \in [0.2, 0.4, 0.6, 0.8]$). The results (mean±std) are reported over 3 random runs. The results with positive gains are **boldfaced** and the best one is underlined.

Dataset	Method	Clean ($\eta = 0.0$)	Symmetric Noise Rate (η) 0.2	0.4	0.6	0.8
MNIST	CE	99.17 ± 0.04	91.40 ± 0.11	74.36 ± 0.29	49.32 ± 0.70	22.32 ± 0.15
	FL	99.16 ± 0.02	91.49 ± 0.20	75.28 ± 0.10	50.25 ± 0.70	22.68 ± 0.14
	GCE	99.15 ± 0.02	98.90 ± 0.03	96.81 ± 0.23	81.39 ± 0.64	33.07 ± 0.31
	SCE	99.28 ± 0.07	98.91 ± 0.12	97.60 ± 0.22	88.00 ± 0.50	47.32 ± 0.99
	NCE+MAE	99.42 ± 0.02	99.18 ± 0.08	98.47 ± 0.21	95.52 ± 0.04	73.05 ± 0.59
	NCE+RCE	99.40 ± 0.04	99.24 ± 0.01	98.44 ± 0.11	95.77 ± 0.09	74.80 ± 0.28
	NFL+RCE	99.37 ± 0.01	99.16 ± 0.03	98.55 ± 0.05	95.62 ± 0.24	74.67 ± 0.97
	NSL	99.24 ± 0.03	98.99 ± 0.03	98.58 ± 0.11	95.99 ± 0.24	59.77 ± 1.98
	CE+FNPAL	99.24 ± 0.05	**99.05 ± 0.04**	**98.66 ± 0.04**	**97.62 ± 0.15**	**79.23 ± 0.87**
	FL+FNPAL	99.16 ± 0.03	**98.97 ± 0.09**	**98.61 ± 0.07**	**97.69 ± 0.03**	**80.63 ± 0.28**
	GCE+FNPAL	99.07 ± 0.06	**99.04 ± 0.01**	**98.62 ± 0.06**	**97.89 ± 0.13**	**94.65 ± 0.52**
	SCE+FNPAL	99.27 ± 0.04	**99.06 ± 0.05**	**98.76 ± 0.09**	**97.94 ± 0.07**	**88.56 ± 1.07**
	NCE+MAE+FNPAL	99.34 ± 0.06	99.10 ± 0.09	98.40 ± 0.09	**95.70 ± 0.05**	**80.21 ± 0.43**
	NCE+RCE+FNPAL	99.29 ± 0.04	99.04 ± 0.07	98.11 ± 0.09	94.84 ± 0.08	**79.70 ± 1.06**
	NFL+RCE+FNPAL	99.29 ± 0.06	99.02 ± 0.05	98.32 ± 0.14	95.38 ± 0.11	**76.06 ± 0.58**
CIFAR10	CE	90.36 ± 0.25	74.78 ± 0.68	57.95 ± 0.12	38.21 ± 0.12	18.89 ± 0.43
	FL	89.69 ± 0.25	74.19 ± 0.23	57.35 ± 0.27	38.11 ± 0.76	19.39 ± 0.44
	GCE	89.37 ± 0.29	87.05 ± 0.21	82.43 ± 0.10	68.05 ± 0.07	25.21 ± 0.28
	SCE	91.24 ± 0.19	87.34 ± 0.01	79.84 ± 0.43	61.09 ± 0.19	27.19 ± 0.34
	NCE+MAE	89.02 ± 0.09	87.06 ± 0.17	83.92 ± 0.16	76.47 ± 0.25	45.01 ± 0.31
	NCE+RCE	91.12 ± 0.14	89.21 ± 0.00	86.03 ± 0.14	80.04 ± 0.26	51.67 ± 1.38
	NFL+RCE	91.03 ± 0.15	89.10 ± 0.16	86.20 ± 0.19	79.58 ± 0.08	50.03 ± 2.78
	NSL	88.07 ± 0.12	86.46 ± 0.02	83.27 ± 0.13	76.17 ± 0.40	46.74 ± 0.72
	CE+FNPAL	90.69 ± 0.11	**86.34 ± 0.37**	**81.30 ± 0.29**	**72.77 ± 0.41**	**51.46 ± 1.10**
	FL+FNPAL	90.02 ± 0.26	**86.17 ± 0.49**	**80.96 ± 0.15**	**72.36 ± 0.34**	**51.54 ± 0.13**
	SCE+FNPAL	91.11 ± 0.13	87.30 ± 0.06	**82.68 ± 0.22**	**73.49 ± 0.42**	**51.99 ± 1.10**
	NCE+MAE+FNPAL	89.18 ± 0.05	**87.26 ± 0.07**	**84.55 ± 0.02**	**78.39 ± 0.49**	**52.15 ± 0.65**
	NCE+RCE+FNPAL	90.88 ± 0.10	**89.34 ± 0.15**	**86.65 ± 0.21**	**80.28 ± 0.07**	**57.21 ± 0.22**
	NFL+RCE+FNPAL	91.16 ± 0.25	**89.49 ± 0.32**	**86.66 ± 0.08**	**80.33 ± 0.15**	**56.23 ± 0.15**
CIFAR100	CE	70.41 ± 1.17	55.64 ± 0.17	40.39 ± 0.46	22.00 ± 1.23	7.37 ± 0.16
	FL	70.56 ± 0.59	56.02 ± 0.80	40.41 ± 0.39	22.11 ± 0.30	7.70 ± 0.20
	GCE	63.06 ± 1.00	62.15 ± 0.66	57.11 ± 1.43	45.99 ± 1.00	18.32 ± 0.36
	SCE	70.41 ± 0.63	55.05 ± 0.68	39.60 ± 0.14	21.53 ± 0.72	7.82 ± 0.30
	NCE+MAE	67.16 ± 0.13	52.34 ± 0.12	35.81 ± 0.42	19.29 ± 0.29	7.31 ± 0.23
	NCE+RCE	68.09 ± 0.26	64.32 ± 0.40	58.11 ± 0.63	45.94 ± 1.31	25.22 ± 0.08
	NFL+RCE	67.58 ± 0.39	64.48 ± 0.50	57.86 ± 0.12	46.74 ± 0.59	24.55 ± 0.47
	NSL	70.08 ± 0.19	65.30 ± 0.36	56.77 ± 0.52	41.21 ± 1.01	12.16 ± 0.96
	CE+FNPAL	71.69 ± 0.27	**65.38 ± 0.17**	**57.24 ± 0.36**	**41.35 ± 0.19**	**12.12 ± 0.88**
	FL+FNPAL	71.79 ± 0.26	**65.24 ± 0.05**	**56.05 ± 0.15**	**40.41 ± 0.47**	**13.16 ± 0.25**
	GCE+FNPAL	70.22 ± 0.05	**63.40 ± 0.12**	50.86 ± 0.38	34.01 ± 0.44	15.08 ± 0.47
	SCE+FNPAL	70.87 ± 0.45	**65.30 ± 0.15**	**55.10 ± 0.45**	**39.73 ± 0.04**	**11.70 ± 0.53**
	NCE+MAE+FNPAL	67.30 ± 0.07	**62.00 ± 0.20**	**52.03 ± 0.18**	**33.78 ± 0.12**	**10.92 ± 0.77**
	NCE+RCE+FNPAL	69.29 ± 0.32	**65.53 ± 0.30**	**60.53 ± 0.27**	**49.73 ± 0.64**	24.54 ± 0.28
	NFL+MAE+FNPAL	69.53 ± 0.05	**65.94 ± 0.32**	**60.89 ± 0.60**	**50.10 ± 0.40**	24.15 ± 1.06

Table 11. Validation accuracies (%) of different methods on benchmark datasets with asymmetric label noise ($\eta \in [0.1, 0.2, 0.3, 0.4]$). The results (mean±std) are reported over 3 random runs. The results with positive gains are **boldfaced** and the best one is underlined.

Dataset	Method	Asymmetric Noise Rate (η) 0.1	0.2	0.3	0.4
MNIST	CE	97.66 ± 0.05	94.42 ± 0.28	88.74 ± 0.07	81.45 ± 0.09
	FL	97.68 ± 0.20	94.18 ± 0.05	88.87 ± 0.30	81.79 ± 0.36
	GCE	99.05 ± 0.03	96.57 ± 0.15	89.20 ± 0.11	81.92 ± 0.45
	SCE	99.17 ± 0.02	98.13 ± 0.17	93.59 ± 0.24	84.42 ± 0.57
	NCE+MAE	99.28 ± 0.08	98.87 ± 0.10	97.00 ± 0.15	91.33 ± 0.65
	NCE+RCE	99.26 ± 0.06	98.91 ± 0.08	97.37 ± 0.20	86.95 ± 4.27
	NFL+RCE	99.33 ± 0.02	98.87 ± 0.05	97.41 ± 0.12	92.35 ± 0.26
	NSL	99.15 ± 0.02	98.79 ± 0.05	95.85 ± 0.12	87.41 ± 0.62
	CE+FNPAL	**98.99 ± 0.01**	**98.58 ± 0.05**	**96.94 ± 0.27**	**91.77 ± 0.94**
	FL+FNPAL	**99.02 ± 0.09**	**98.52 ± 0.06**	**97.06 ± 0.05**	**92.17 ± 0.44**
	GCE+FNPAL	99.11 ± 0.07	99.05 ± 0.03	98.82 ± 0.09	97.27 ± 0.23
	SCE+FNPAL	**99.23 ± 0.05**	**99.08 ± 0.02**	**98.43 ± 0.13**	**92.11 ± 0.65**
	NCE+MAE+FNPAL	99.23 ± 0.09	**99.00 ± 0.01**	**97.71 ± 0.20**	**93.95 ± 0.43**
	NCE+RCE+FNPAL	**99.26 ± 0.05**	**98.93 ± 0.04**	**97.91 ± 0.29**	**94.37 ± 0.54**
	NFL+RCE+FNPAL	99.17 ± 0.10	**99.02 ± 0.10**	**98.18 ± 0.19**	**94.86 ± 0.18**
CIFAR-10	CE	86.89 ± 0.04	83.34 ± 0.32	79.58 ± 0.22	74.61 ± 0.42
	FL	86.36 ± 0.30	83.31 ± 0.23	79.67 ± 0.08	74.19 ± 0.40
	GCE	88.25 ± 0.47	85.91 ± 0.02	80.56 ± 0.30	74.49 ± 0.34
	SCE	89.73 ± 0.22	86.47 ± 0.14	81.61 ± 0.17	75.52 ± 0.28
	NCE+MAE	88.41 ± 0.28	86.52 ± 0.28	83.74 ± 0.19	76.75 ± 0.33
	NCE+RCE	90.19 ± 0.20	88.61 ± 0.27	85.55 ± 0.01	79.25 ± 0.33
	NFL+RCE	89.97 ± 0.17	88.58 ± 0.15	85.53 ± 0.26	79.61 ± 0.07
	NSL	87.50 ± 0.04	86.23 ± 0.23	83.56 ± 0.15	78.63 ± 0.28
	CE+FNPAL	**88.55 ± 0.11**	**86.11 ± 0.19**	**83.06 ± 0.06**	**77.41 ± 0.27**
	FL+FNPAL	**88.12 ± 0.18**	**85.56 ± 0.11**	**82.90 ± 0.09**	**76.74 ± 0.52**
	SCE+FNPAL	89.24 ± 0.09	**87.32 ± 0.21**	**84.18 ± 0.10**	**78.15 ± 0.34**
	NCE+MAE+FNPAL	88.21 ± 0.04	**87.13 ± 0.27**	**85.43 ± 0.41**	**78.96 ± 0.34**
	NCE+RCE+FNPAL	89.94 ± 0.15	**89.01 ± 0.18**	**85.97 ± 0.19**	**79.51 ± 0.19**
	NFL+RCE+FNPAL	**90.05 ± 0.06**	**88.64 ± 0.17**	**86.20 ± 0.09**	**79.66 ± 0.60**
CIFAR-100	CE	65.03 ± 0.07	57.67 ± 0.89	50.63 ± 0.57	41.95 ± 0.12
	FL	64.83 ± 1.13	58.37 ± 0.56	51.44 ± 0.79	42.16 ± 0.53
	GCE	62.46 ± 0.50	59.56 ± 0.64	54.22 ± 1.18	42.18 ± 0.77
	SCE	64.55 ± 0.58	58.15 ± 0.54	50.58 ± 0.27	41.57 ± 0.19
	NCE+MAE	60.55 ± 0.42	52.03 ± 0.15	44.45 ± 0.22	36.86 ± 0.46
	NCE+RCE	66.31 ± 0.59	62.93 ± 0.42	55.70 ± 0.14	42.73 ± 0.58
	NFL+RCE	66.41 ± 0.20	62.94 ± 0.39	55.83 ± 0.46	42.63 ± 0.13
	NSL	65.79 ± 0.27	58.16 ± 0.46	50.35 ± 0.41	40.38 ± 0.60
	CE+FNPAL	**65.99 ± 0.06**	**57.93 ± 0.18**	**51.45 ± 0.92**	**47.53 ± 0.94**
	FL+FNPAL	**65.54 ± 0.13**	**57.17 ± 0.43**	**51.74 ± 0.64**	**44.63 ± 0.14**
	SCE+FNPAL	**64.65 ± 0.37**	**57.61 ± 0.22**	**50.74 ± 0.51**	**44.31 ± 0.22**
	NCE+MAE+FNPAL	60.53 ± 0.73	**52.58 ± 0.38**	**44.96 ± 0.29**	**36.88 ± 0.40**
	NCE+RCE+FNPAL	**67.19 ± 0.10**	**64.63 ± 0.32**	**57.43 ± 0.53**	**49.77 ± 1.11**
	NFL+RCE+FNPAL	**67.09 ± 0.30**	**64.32 ± 0.59**	**57.92 ± 0.65**	**50.19 ± 1.60**