Wrapped Cauchy Distributed Angular Softmax for Long-Tailed Visual Recognition

Boran Han [1]

Abstract

Addressing imbalanced or long-tailed data is a major challenge in visual recognition tasks due to disparities between training and testing distributions and issues with data noise. We propose the Wrapped Cauchy Distributed Angular Softmax (WCDAS), a novel softmax function that incorporates data-wise Gaussian-based kernels into the angular correlation between feature representations and classifier weights, effectively mitigating noise and sparse sampling concerns. The class-wise distribution of angular representation becomes a sum of these kernels. Our theoretical analysis reveals that the wrapped Cauchy distribution excels the Gaussian distribution in approximating mixed distributions. Additionally, WCDAS uses trainable concentration parameters to dynamically adjust the compactness and margin of each class. Empirical results confirm label-aware behavior in these parameters and demonstrate WCDAS's superiority over other state-of-the-art softmax-based methods in handling long-tailed visual recognition across multiple benchmark datasets. The code is public available.

1. Introduction

Deep convolutional neural networks are the leading methods for computer vision tasks, including visual recognition. This strength is largely due to their robust representation learning, a technique that simplifies target images into a vector space with fewer dimensions. This crucial step is facilitated by the penultimate layer and subsequently fed into the final classifier, followed by a softmax function, which calculates the probability of an input image being in the j-th class: $P(y = j \mid \mathbf{x})$ (Bridle, 1989; Goodfellow

et al., 2016). However, most image recognition tasks have been demonstrated on well-balanced datasets. In contrast, most real-world data comes with an imbalanced distribution: a few high-frequency classes contain many training examples, while many low-frequency classes have insufficient training examples. This scenario is referred to as long-tailed recognition (Liu et al., 2019), and standard methods trained with such datasets tend not to yield the same performance as balanced ones (Liu et al., 2019; Lin et al., 2017; Cui et al., 2021).

Numerous studies have focused on long-tailed recognition by attempting to re-balance the data distribution through class-balanced sampling or class re-weighting (Han et al., 2005; Kang et al., 2020; Kubát & Matwin, 1997; Huang et al., 2016; 2020; Hong et al., 2023). However, they may under-represent the majority class (Han et al., 2005; Kang et al., 2020; Kubát & Matwin, 1997) or destabilized the network during optimization (Huang et al., 2016; 2020). In addition to direct sampling, focal loss (Lin et al., 2017) adopts loss function emphasizing samples with larger loss value. However, it inevitably involves hyperparameters tuning by cross-validation. An alternative method is to adopt a label-aware correction via introducing a class-wise generalization error bound, such as Label-Distribution-Aware Margin Loss (LDAM) (Cao et al., 2019) and Balanced Meta-Softmax (BALMS) (Ren et al., 2020). Cao, *et. al.* have proved that to improve the accuracy in recognizing long-tailed distributed data, classes with fewer training examples should have a higher generalization error bound (Cao et al., 2019). However, both LDAM and BALMS can be vulnerable when the number of examples per class is unknown and constantly changing. Therefore, further corrections are required for continuous training. Meta-Weight-Net (Shu et al., 2019) and Equilibrium loss (Feng et al., 2021) are developed for class re-weighting and inter-class margin correction, which require no visibility to the underlying data distribution. However, those methods can either be subject to lengthy training time due to the nature of meta-learning (Shu et al., 2019) or high space complexity because of the memory module (Feng et al., 2021). Lastly, using angular softmax, Kobayashi has proposed applying von Mises-Fisher distribution for compact feature space via a user-defined concentration

[1]Amazon Web Services, AI. Work done while at Shell.. Correspondence to: Boran Han <boranhan@amazon.com>.

parameter (κ) (Kobayashi, 2021). However, such a method leads to lengthy hyper-parameter tuning with isotopic κ for all classes. Meanwhile, their trainable class-wise κ approach shows inferior performance compared with the user-defined counterpart for an optimal performance (Kobayashi, 2021). In addition, data noise also exists in long-tail problem (Tong Wu & Lin, 2021; Cao et al., 2021; Zhang et al., 2023).

In light of these challenges, we propose the Wrapped Cauchy Distributed Angular Softmax (WCDAS) for long-tailed visual recognition based on (Kobayashi, 2021). We presume that the data-wise probability distribution follows the wrapped Normal distribution and deduce that WCDAS can be a better fit for mixed distributions comprised of individual distributions. We also demonstrate that WCDAS has several desirable features, such as adaptive regulation of the margins between classes via a concentration parameter, exhibiting label-aware behavior. Upon evaluation on several benchmark long-tailed image classification datasets, WCDAS outperforms state-of-the-art softmax-based methods.

In summary, our contributions include: 1) proposing a model that considers noise-induced uncertainty in the form of data-wise wrapped Normal distributed kernels; 2) proving that WCDAS can more effectively fit the mixed distribution of such kernels; 3) showing that under a specific condition, our method also significantly enhances inter-class margins, resulting in compact clustering; and 4) demonstrating that the concentration parameter can be adaptive, with classes with fewer training samples having a higher concentration parameter and a larger margin.

2. Related works

Angular-based Softmax. Angular softmax (Liu et al., 2016) and its mutant approaches (Deng et al., 2019; Liu et al., 2017) have recently been proposed to improve the softmax loss in face verification tasks. Unlike conventional softmax, these methods allow neural networks to learn features in an angular manner by focusing on the cosine similarity between classifier weights and features. Among these, Large-margin softmax (Liu et al., 2016) directly enforces inter-class separability on the dot-product similarity, while SphereFace (Liu et al., 2017) and ArcFace (Deng et al., 2019) enforce multiplicative and additive angular margins on the hypersphere manifold, respectively. These margins are controlled by a hyperparameter, m: the larger the value of m, the larger the margin. Consequently, larger margins between classes can lead to compact clusters, resulting in enhanced performance over conventional softmax. (Liu et al., 2017; Deng et al., 2019; Liu et al., 2016).

Long-tailed recognition. Datasets with long-tailed distribution (Liu et al., 2019) not only have an imbalanced class with respect to the number of examples per class but also have a long tail of classes with only a few examples (<10), i. e., tail class. Two predominant approaches for such a problem are (1) loss function improvement and (2) data re-balancing. The former approach exploits aggressive learning in the tail classes (Lin et al., 2017; Jingru Tan, 2020; Cui et al., 2019) or forcing large margin between classes, especially tail classes(Cao et al., 2019; Ren et al., 2020; Ye et al., 2020). In particular, Cao *et. al.* (Cao et al., 2019) theoretically prove that the generalization error bound could be minimized by increasing the margins of tail classes. In addition to margin correction, Feng *et. al.* also balances the classification via a Feature Memory Module (Feng et al., 2021). At the same time, a handful of studies focus on data re-balancing during training, the second approach for imbalance training. Data rebalancing can be achieved by data re-sampling (Han et al., 2005; Kang et al., 2020; Kubát & Matwin, 1997) or class re-weighting (Huang et al., 2016; 2020). However, data re-balancing-based strategies can lead to overfitting the tail classes and less efficient learning of the over-representative ones. The sampling strategies include fixed samplers (Kang et al., 2020) and meta-based samplers (Ren et al., 2020; Shu et al., 2019). Decoupled training (Kang et al., 2020) is a simple yet effective solution that could significantly improve the generalization issue on long-tailed datasets. During this two-stage training, the representation learning is trained by instance-balanced sampler (Kang et al., 2020) while the classifier is further fine-tuned by class-balanced sampler (Kang et al., 2020) and meta sampler (Ren et al., 2020).

Parametric modeling of feature distribution. Despite the emergence of deep learning being attributed to non-parametric non-linearity modeling, effectively training a network can prove challenging when dealing with certain real-world datasets that present issues such as class imbalance and insufficient examples. Parametric modeling, based on certain assumptions, can greatly assist learning in these adverse situations (Yang et al., 2021; Hayat et al., 2019). One such approach involves approximating the Gaussian distribution of feature representation in few-shot learning to enhance generalizability (Yang et al., 2021). In the context of imbalanced classes, studies have shown that Gaussian distribution (Hayat et al., 2019) and von Mises-Fisher distribution (Kobayashi, 2021) modeling of feature representation, or angles between weights and features, can significantly improve performance. Parametric modeling of the feature space can also better handle uncertainty caused by noise in the data. Popular methods of utilizing parametric models to account for uncertainty include Variational Auto-encoder (Kingma & Welling, 2013), Bayesian-based dropout (Gal & Ghahramani, 2016), and DUL (Chang et al., 2020), among others.

Inspired by these three distinct approaches, we propose a method that parametrically models the feature representation. This method uses data-wise Gaussian kernels as basis and it includes class-wise parameters that are trainable, providing an adaptable framework for various types of data."

3. Wrapped Cauchy Distributed Angular Softmax (WCDAS)

Previous knowledge. For angular softmax, the predicted probability from the linear classifier in CNNs for the j-th class given a sample vector $\mathbf{x}$ and a weighting vector $\mathbf{w}$ is formulated as:

$$P(y = j \mid \theta) = \frac{e^{f(\theta;j)}}{\sum_{c=1}^{C} e^{f(\theta;c)}} = \frac{e^{s\cos\theta_j}}{\sum_{c=1}^{C} e^{s\cos\theta_c}} \quad (1)$$

where,

$$f(\theta; j) = s\cos\theta_j \quad (2)$$

$f(\theta; j)$ calculates the angle between normalized vectors $\mathbf{x}$ and $\mathbf{w}$, $\cos\theta_j = \mathbf{x}^\mathsf{T}\mathbf{w}_j$. For the ease of writing, we refer the angular representation (θ_j) between $\mathbf{x}$ and $\mathbf{w}$ as "angular features". $s \in \mathbb{R}$ is a empirically-defined constant (Deng et al., 2019; Liu et al., 2017) or trainable parameter (Kobayashi, 2021).

Intuition and Overview of WCDAS. The probability function ($f(\theta; j)$) of the angular softmax function (Equation 2) describes the angle between representation features and classifier weights. As such, the classifier weights are optimized to minimize the loss function, given $\cos\theta_j$. However, this approach may potentially lead to overfitting, especially when training with a few examples, as discussed in previous large-margin based cosine softmax studies (Kobayashi, 2021; Liu et al., 2016), or with data containing noise, as reported by other studies (Tong Wu & Lin, 2021; Cao et al., 2021; Zhang et al., 2023). To address these issues, our method seeks an optimal parametric probability density function of θ_j, conditioned on $y = j$, i.e., $P(\theta \mid y = j)$. To achieve this, we initially propose using a data-wise Gaussian-based kernel as a basis. Intuitively, given θ, such a kernel can model the data-wise uncertainty caused by the input noise or sparse sampling, instead of a direct class-wise distribution (Section 3.1). By doing so, we can obtain the class-wise angular feature probability density function by summing the individual basis (Section 3.1). Subsequently, we prove that this class-wise distribution can be more accurately approximated by a Wrapped Cauchy distribution, $f(\boldsymbol{\rho}, \theta; j)$, with a class-wise trainable concentration parameter, $\boldsymbol{\rho} \in \mathbb{R}^\mathbb{C}$ (Section 3.2). We provide insights into why our novel softmax is a better parametric distribution for representation feature modeling (Section 3.2) and how it can create large margins under specific conditions (Section 3.3).
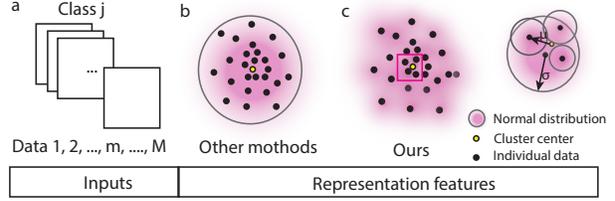

Figure 1: Illustration of our method compared with other methods. Black dot: representation of each data in one class. Yellow dot with a black edge: centroid of the cluster. Gray solid line: Gaussian kernel boundary. (a) Input data 1, 2, ..., M in Class j. (b) parametric modeling of features from *each data* via a wrapped Normal kernel. (Hayat et al., 2019; Kobayashi, 2021) (c) Left panel: parametric modeling of features from *each data* via wrapped Normal distribution. Right panel: zoomed diagram of the magenta box in the left panel.

3.1. Wrapped Normal Basis for Angular Feature Density Estimation.

Assumption. To mitigate overfitting in the representation features, we approximate the uncertainty induced by noise or sparse sampling using a Gaussian distribution. Consequently, the angular feature of each data point follows the probability distribution of a Normal distribution in circular coordinates, i.e., a Wrapped Normal distribution or a von Mises-Fisher distribution. Given that the latter approximates the former distribution, we treat both distributions as equivalent for ease of discussion. This model of noise or sparse sampling-induced uncertainty using a Gaussian distribution has been widely utilized in various studies (Gal & Ghahramani, 2016; Rasmussen & Williams, 2005; Abdar et al., 2021). Following this assumption, the probability distribution function of the angular feature for the m-th data point in the j-th class can be represented in the form of a Symmetric-Wrapped Stable (SWS) distribution (Jammalamadaka & SenGupta, 2001):

$$h(\rho, \theta; m, j) = \frac{1}{2\pi} \left(1 + 2\sum_{n=1}^{\infty} \rho_m^{n^a} \cos n(\theta_m - \mu_m) \right) \quad (3)$$

where $n \in \mathbb{N}$, $\rho_m \in [0, 1)$ denotes concentration parameter of m-th data in j-th class, μ_m denotes the center of j-th class and $a \in (0, 2]$. When $a = 1$, Equation 18 returns the wrapped Cauchy distribution and for $a = 2$, we get the wrapped Normal distribution (Jammalamadaka & SenGupta, 2001). The bigger ρ_m is, the more compact the wrapped Normal kernel is. Since $h(\rho, \theta; m, j)$ computes the probability θ_m belongs to j-th class with the optimized classifier weights, hence, for the correct class to be recognized based on Equation 1, $\mu_m \to 0$.

Note that in our proposed method, we approximate the uncertainty of each θ_m as wrapped Normal distribution parameterized by ρ_m and μ_m instead of modeling the

$f(\theta; j)$ directly (Hayat et al., 2019; Kobayashi, 2021). Such difference is shown in Figure 1(b) and (c).

Class-wise probability distribution. Subsequently, mixed distribution $f(\theta; j)$ can be obtained by summing all the $h(\rho, \theta; m, j)$ in j-th class:

$$f_{\text{mixed}}(\theta; j) = \frac{1}{M_j} \sum_{m=1}^{M_j} h(\rho, \theta; m, j) \quad (4)$$

where M_j is the total number of samples in j-th class. $f_{\text{mixed}}(\theta; j)$ describes the mixture of M_j wrapped Normal distributions centered around zero. Such an idea is used in the non-parametric estimation of a probability density function, such as kernel density estimation (KDE) (Rosenblatt, 1956; Parzen, 1962). However different from KDE, $\boldsymbol{\rho}_j$, a vector comprised of all ρ_m in j-th class, can be different in values, representing the heterogeneity of each data.

Theorem 1. *Let $f_{mixed}(\theta; j)$ be a mixed distribution formed by summing several wrapped Normal distributions $h(\rho, \theta; m, j)$ (Equation 18). $h(\rho, \theta; m, j)$ is centered at μ_m. μ_m follows Normal distribution $\mathcal{N}(0, \sigma)$ centered at zero, where $\sigma \to 0$. Then $f_{mixed}(\theta; j)$ can be approximated as:*

$$f_{mixed}(\theta; j) \sim \frac{1}{2\pi M_j} \sum_{m=1}^{M_j} \left(1 + 2 \sum_{n=1}^{\infty} \rho_m^{n^2} \cos n\theta_m \right) \quad (5)$$

Corollary 1.1. *Let $f_{mixed}(\theta; j)$ be a mixed distribution formed by mixing several wrapped Normal distributions (Equation 18 and Equation 5), then $f_{mixed}(\theta; j)$ is a wrapped distribution with cosine moments, α_{mixed}, given by*

$$\alpha_{mixed}^{\{n\}} = \frac{1}{M_j} \sum_{m=1}^{M_j} \alpha_m^{\{n\}} \quad (6)$$

where $\alpha_m^{\{n\}}$ is the n-th cosine trigonometric moment of $h(\rho, \theta; m, j)$.

We present the detailed proof of Theorem 1 and Corollary 1.1 in *Appendix A*. Theorem 1 essentially shows that when summing the wrapped Normal distributed kernel basis with a small perturbation away from zero, the result can be approximated as a sum of wrapped Normal distributions centered at zeros. Corollary 1.1 demonstrates that cosine moments of the mixed distribution can be obtained by averaging the cosine moments of each distribution. We note that the cosine moments of mixed distribution from two wrapped Normal distributions centered at zeros have been proven by Bailey, et. al. (Bailey & Codling, 2020). We here prove that it can be generalized to several functions that are not centered at zero under certain conditions.

3.2. Angular Feature Probability Approximation via Wrapped Cauchy Distribution.

It is vital to find the optimal presentation of $f_{\text{mixed}}(\theta; j)$. One straightforward solution is to use non-parametric approaches (Rosenblatt, 1956; Parzen, 1962). However, those methods usually require large computational costs for large dataset (Holmström, 2000). In our case, those methods also requires each ρ_m to be calculated separately. Therefore, we approximate $f_{\text{mixed}}(\theta; j)$ with parametric distribution, denoting $f(\theta, \rho; j)$. According to Theorem 1, $f(\theta, \rho; j)$ should also be an SWS distribution. Among the two predominant SWS distributions (Wrapped Cauchy distribution vs Wrapped Normal distribution), wrapped Cauchy distribution can fit Equation 5 better than the wrapped Normal distribution.

Theorem 2. *Let $f_{mixed}(\theta; j)$ be a mixed distribution formed by mixed several wrapped Normal distributions $h(\rho, \theta; m, j)$ of j-th class, centred around zero, defined in Equation 4. $f(\theta, \rho; j)$ is the approximated distribution with a choice of wrapped Normal $f_{WN}(\theta, \rho; j)$ and wrapped Cauchy $f_{WC}(\theta, \rho; j)$. Let $\rho_{j, min}$ of $f(\theta, \rho; j)$ minimize the least square error between $f(\theta, \rho; j)$ and mixed distribution $f_{mixed}(\theta; j)$ of j-th class:*

$$\Delta_{\rho_j, WC \text{ or } WN} = ||f_{WN \text{ or } WC}(\rho, \theta; j)$$
$$- \frac{1}{2\pi M_j} \sum_{m=1}^{M_j} \left(1 + 2 \sum_{n=1}^{\infty} \rho_m^{n^2} \cos n\theta_m \right) ||^2 \quad (7)$$

$$\rho_{min} = \arg\min_{\rho} \Delta_{\rho} \quad (8)$$

Then the least square error of optimal $\rho_{j, min}$ of j-th class is correlated with standard deviation of $(\sum_{m_c=1}^{M_c} \rho_m^{n^2})^{\frac{1}{n}}$ or $(\sum_{m_c=1}^{M_c} \rho_m^{n^2})^{\frac{1}{n^2}}$ with respect to $n \in [1, \infty)$

$$\Delta_{\rho_{j,min}, WN} \propto SD_{n=1} \left(\sum_{m=1}^{M_j} \rho_m^{n^2} \right)^{\frac{1}{n}}$$

$$\text{or } \Delta_{\rho_{j,min}, WC} \propto SD_{n=1} \left(\sum_{m=1}^{M_j} \rho_m^{n^2} \right)^{\frac{1}{n^2}} \quad (9)$$

Theorem 3. *Let ρ_m of individual $h(\rho, \theta; m, j)$ distribute uniformly across its defined domain $[0, 1)$, Δ_{WN} and Δ_{WC} defined in Equation 29, Then $\Delta_{\rho_{j,min}, WC} < \Delta_{\rho_{j,min}, WN}$*

The detailed proofs of Theorem 2 and Theorem 3 are provided in *Appendix A*. These proofs substantiate that a mixed distribution constituted by SWS distributions aligns better with the wrapped Cauchy distribution than with the wrapped Normal distribution. Although the proof is analytical, it is based on the numerical assumption that ρ_m in the j-th class is evenly distributed in $[0, 1)$ (Theorem 3).
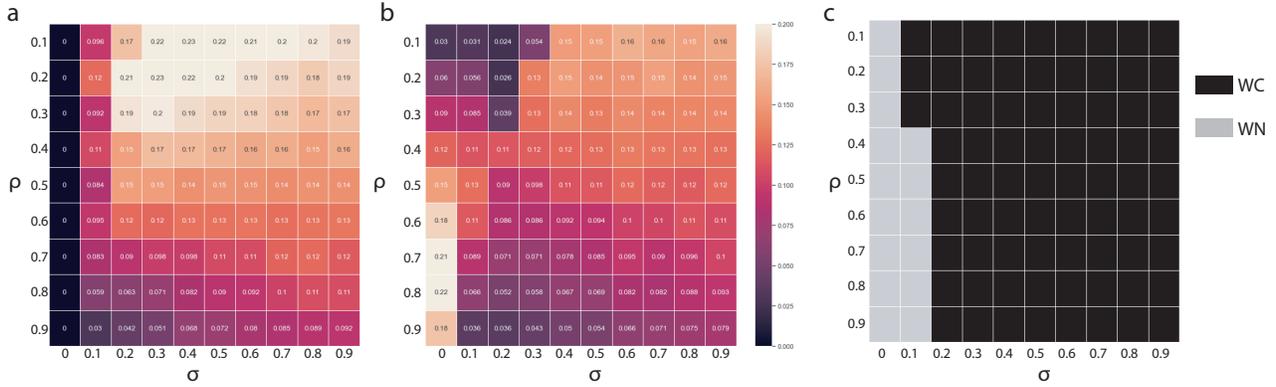
Figure 2: Heatmap of $\Delta_{\rho_{\min},\text{wn}}$ (a) and $\Delta_{\rho_{\min},\text{wc}}$ (b) with respect to ρ and σ. (c) Binary heatmap showing whether wrapped Cauchy (WC: black) or wrapped Normal (WN: gray) is preferred for simulated mixed distribution.

We also provide a numerical simulation for more general situations where $\rho_m \sim \mathcal{N}(\mu_\rho, \sigma_\rho)$. Given that $\rho \in [0, 1)$, we simulate μ_ρ and σ_ρ with the value of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Any ρ values outside the $[0, 1)$ domain are clipped to 0 and 1, respectively. Figure 2 shows the results, indicating a preference for the Wrapped Normal distribution when σ_ρ is small ($\sigma_\rho \leq 0.1$); otherwise, the wrapped Cauchy distribution is preferred. This implies that unless the mixed distribution $f_{\text{mixed}}(\theta; j)$ comprises wrapped Normal distributions with similar concentration parameters, the wrapped Cauchy distribution, due to its heavy tail, provides a better approximation for $f_{\text{mixed}}(\theta; j)$.

Assuming that ρ_m follows a uniform distribution is an idealized assumption that simplifies the learning process. In practice, the actual distribution of ρ_m might be more intricate. However, as our simulation of Gaussian-distributed ρ_m demonstrates, there is a trend: the greater the diversity of ρ_m values, the more advantageous the Cauchy distribution becomes as an approximation over Gaussian, given the heavy tail of the Cauchy distribution.

3.3. Large Margin ρ and optimization

It is important to obtain the optimal ρ. According to the geometric series, Equation 18 can be written in an alternative form with element-wise calculation(Jammalamadaka & SenGupta, 2001):

$$f(\boldsymbol{\rho}, \theta) = \frac{1 - \boldsymbol{\rho}^2}{2\pi(1 + \boldsymbol{\rho}^2 - 2\boldsymbol{\rho}\cos\theta))} \quad (10)$$

where $\boldsymbol{\rho}$ is the vector containing $\rho_{j \in [1,C]}$ from all classes with the total number of C. We note that this alternative form of Equation 18 is presented for the ease of calculating the margin between classes.

Large margin via WCDAS. Several studies have demonstrated that the large margin-based softmax approach can lead to better performance both in balanced (Deng et al.,

2019; Liu et al., 2017; 2016) and imbalanced datasets (Ren et al., 2020; Hayat et al., 2019; Cao et al., 2019). We here prove that WCDAS can perform equivalently as those methods under a certain domain of $\boldsymbol{\rho}$. However, we note that not all $\boldsymbol{\rho}$ in WCDAS contribute to a large margin. Intuitively, only high $\boldsymbol{\rho}$ leads to tighter clustering. We here provide the boundary of $\boldsymbol{\rho}$ that will lead to large inter-class margins.

Theorem 4. *Let ρ_j be the concentration parameter of wrapped Cauchy distribution f_{wc} of the j-th class. $\mathbf{x}$ is the normalized presentation feature and $\mathbf{w}$ is the normalized weights of the classifier layer. Let θ_j and θ_k be the angle between $\mathbf{x}$ and $\mathbf{w}$ of j-th and k-th class respectively, where $\mathbf{x}$ is from class j. When $\rho_j \in (0.42332, 1)$, then $\|f_{WC}(\theta_j) - f_{WC}(\theta_k)\| > \|\cos\theta_j - \cos\theta_k\|$ for any θ_j and θ_k when $\cos\theta_j > \cos\theta_k$. The margin can be expressed as*

$$\|f_{WC}(\theta_j) - f_{WC}(\theta_k)\| = \frac{\rho_j + \rho_j^2}{\pi(1 - \rho_j)^3} \|\cos\theta_j - \cos\theta_k\| \quad (11)$$

The detailed derivation is shown in *Appendix A*. Theorem 4 shows that within such a domain, WCDAS yields a larger margin compared with $\cos\theta$. It is worth mentioning that such behavior holds with any θ_j and θ_k. It is also shown that the larger ρ_j is, the larger the margin is (*Appendix Figure 6*). However, we note that our paper cannot prove that the margin is label-aware because of the gradient-based optimization. Therefore, the behaviors of ρ during optimization require numerical studies (see Section 4.3).

Optimization We can calculate its gradient with respect to ρ:

$$\frac{\partial f(\boldsymbol{\rho}, \theta)}{\partial \boldsymbol{\rho}} = \frac{-2\boldsymbol{\rho} + (1 + \boldsymbol{\rho}^2)\cos\theta}{\pi(1 + \boldsymbol{\rho}^2 - 2\boldsymbol{\rho}\cos\theta)^2} \quad (12)$$

Through direct visualization of Equation 12 (Figure 3), we notice two characteristics of our method: (1) when θ is away from 0, ρ decreases, i. e., $\frac{\partial f(\boldsymbol{\rho}, \theta)}{\partial \boldsymbol{\rho}} < 0$; when θ is

5
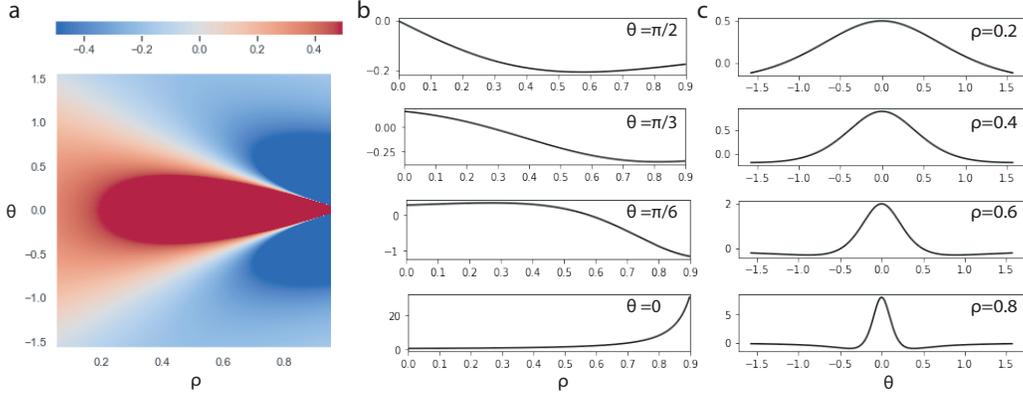
Figure 3: Gradient plot of $\frac{\partial f(\boldsymbol{\rho},\theta;j)}{\partial \boldsymbol{\rho}}$ with respect to ρ and θ(a). The cross sections plotted along ρ (b) and θ (c).

around 0, ρ increases, i. e., $\frac{\partial f(\boldsymbol{\rho},\theta)}{\partial \boldsymbol{\rho}} > 0$. (2) The gradient $\frac{\partial f(\boldsymbol{\rho},\theta)}{\partial \boldsymbol{\rho}}$ also increases when θ is around 0. Through the former characteristic, $\boldsymbol{\rho}$ is able to regulate the margin from the classifier layer. In contrast, the second characteristic can destabilize the whole network, since the value of $\boldsymbol{\rho}$ can also go beyond the defined domain. To address this issue, we define $\boldsymbol{w_\rho} \in (-\infty, \infty)$ so that $\boldsymbol{\rho}$ follows the behavior of sigmoid function with respect to $\boldsymbol{w_\rho}$, which approximates $\boldsymbol{\rho} \in [0, 1)$:

$$\boldsymbol{\rho} = \frac{1}{1+e^{-\boldsymbol{w_\rho}}}, \boldsymbol{w_\rho} \in \mathbb{R}^\mathbb{C} \qquad (13)$$

In summary, both the classifier and the feature extractor update the gradient. While the classifier is updated using our proposed method in Algorithm 1, the feature extractor (or encoder) is trained in a conventional manner.

Algorithm 1 Wrapped Cauchy Distributed Angular Softmax

1: **Input:** Epoch number E, feature representation $\boldsymbol{x}$, weights in classifier $\mathbf{w}$, scale s.
2: **Initialize:** w_ρ
3: **while** $e < E$ **do**
4:     **while** in Minibatch **do**
5:         $\boldsymbol{\rho} = \frac{1}{1+e^{-w_\rho}}$
6:         $\cos\theta = \frac{\mathbf{x}^\top \mathbf{w}_j}{\|\mathbf{x}\|\|\mathbf{w}\|}$
7:         $f(\boldsymbol{\rho},\theta) = \frac{1-\boldsymbol{\rho}^2}{2\pi(1+\boldsymbol{\rho}^2-2\boldsymbol{\rho}\cos\theta))}$
8:         Compute Softmax: $\frac{e^{f(\rho,\theta;j)}}{\sum_{c=1}^{C} e^{f(\rho,\theta;c)}}$
9:         Compute the cross entropy loss L
10:         Update $\mathbf{w}_\rho, \mathbf{w}$ based on gradients $\frac{\partial L}{\partial \mathbf{w}_\rho}, \frac{\partial L}{\partial \mathbf{w}}$
11:     **end while**
12:     $e \leftarrow e + 1$
13: **end while**

---

4. Empirical Experiments

4.1. Experimental setup

We perform extensive ablation experiments on different aspects of our method (Section 4.2 and 4.3). We also compared our approach with SOTA softmax-based methods (Section 4.4) using four large-scale long-tailed datasets: CIFAR10-LT/100-LT (Krizhevsky, 2009), ImageNet-LT (Liu et al., 2019; Deng et al., 2009) and iNaturalist 2018 (Van Horn et al., 2018). Among those datasets, CIFAR10-LT, CIFAR100-LT and ImageNet-LT are truncated from their balanced counterpart, following exponential decay across classes (Liu et al., 2019) (see detail descriptions in *Appendix B.1*).

Implementation. All models are trained using SGD optimizer with momentum 0.9, weight decay 10^{-4}. The learning rate decays by a cosine scheduler. Unless specified, we use 90 training epochs. Other hyper-parameters are listed in *Appendix Table 5*. The standard data augmentation is applied to input images. According to (Kang et al., 2020), we apply a decoupled representation learning and classifier learning: The whole network is first trained via an instance-balanced sampler (Kang et al., 2020). Only the classifier is further trained over 30 epochs sampled by a class-balanced sampler (Kang et al., 2020) or meta sampler (Ren et al., 2020). We apply WCDAS to both feature learning and classifier learning.

4.2. Wrapped Normal vs Wrapped Cauchy, Class-wise ρ vs Single ρ

In this numerical experiment, we further validate Theorem 3 utilizing ImageNet-LT. For a fair comparison, Angular Softmax (Equation 1) is used as a baseline instead of the conventional softmax function. Note that we implement von Mises–Fisher distribution to approximate wrapped Normal distribution (WNDAS). Table 1 shows that despite that both

ρ	one w_ρ for all classes ($w_\rho \in \mathbb{R}$)				class-wise $\boldsymbol{w_\rho}$ ($\boldsymbol{w_\rho} \in \mathbb{R}^{\mathbb{C}}$)			
Method	Many	Medium	Few	All	Many	Medium	Few	All
Angular Softmax	52.8	33.9	15.7	38.7	-	-	-	-
WNDAS	55.0	38.2	20.4	42.1	54.9	38.6	20.2	42.3
WCDAS	**56.2**	**40.4**	**21.7**	**43.8**	56.2	40.9	24.1	44.5

Table 1: Top 1 accuracy for ImageNet-LT (ResNet-10 (He et al., 2016)) with wrapped Normal distributed angular softmax (WNDAS) and WCDAS using one $w_\rho \in \mathbb{R}$ or class-wise $\boldsymbol{w_\rho} \in \mathbb{R}^{\mathbb{C}}$. The result validates Theorem 3

WNDAS and WCDAS display evident improvement from the baseline counterpart, WCDAS consistently performs better than WNDAS. Additionally, we test scenario when setting one w_ρ for all classes ($w_\rho \in \mathbb{R}$) or class-wise $\boldsymbol{w_\rho}$ ($\boldsymbol{w_\rho} \in \mathbb{R}^{\mathbb{C}}$). Our result proves that class-wise $\boldsymbol{w_\rho}$ shows superior performance. Intuitively, such results demonstrate that classes in the long-tailed training require different margins for better accuracy, consistent with previous observations (Cao et al., 2019; Ren et al., 2020).

4.3. w_ρ optimization.

Init.	Many	Medium	Few	All
2.0	55.6	40.5	23.2	43.9
1.0	57.3	40.5	21.4	44.3
0	56.0	40.7	23.5	44.2
He	56.2	41.1	22.6	44.3
Xa.	56.3	40.7	22.5	44.2
-1.0	56.2	40.9	24.1	44.5
-2.0	56.3	40.5	23.1	44.0

Table 2: Top 1 accuracy for ImageNet-LT (ResNet-10 (He et al., 2016)) with various w_ρ initialization (Init.) values. He (He et al., 2015) and Xavier (Xa.) (Glorot & Bengio, 2010). Initial learning rate: 0.4

Robustness of w_ρ Initialization. The initialization of parameters is a critical element in the optimization of deep networks, having significant impact on the quality of the final model. Given that our method introduces a new trainable parameter, $\boldsymbol{w_\rho}$, we performed empirical evaluations to assess its robustness under different initialization strategies. We observed some variance in the final outcomes depending on the initialization values used (*Appendix Table 6*). This discrepancy, however, could be mitigated by either extending the number of training epochs (*Appendix Table 6*) or increasing the learning rate (Table 4.3). This suggests that shorter training periods or smaller learning rates may not be adequate for our approach. We also experimented with the He (He et al., 2015) and Xavier (Glorot & Bengio, 2010) initialization strategies, both of which are zero-centered. The results indicated that the final model was less sensitive to these initialization methods (Table 4.3).

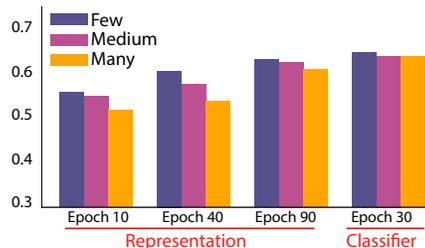**Visualizing ρ During Optimization.** For a closer look

Figure 4: Bar graph of ρ with respect to three sets of class at different stages of training: 10th, 40th, 90th epoch at representation learning and 30th epoch at classifier learning. Three sets of class include few (<20), medium (20-100) and many (>100). Class-balanced sampler are used in classifier learning.
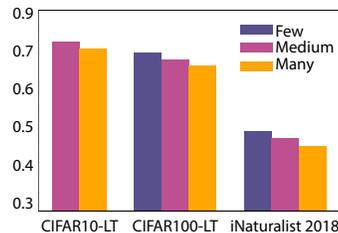

Figure 5: Bar graph of ρ values on CIFAR100-TL/10-TL and iNaturalist 2018.

at the optimization process, we graphically display the values of $\boldsymbol{\rho}$ during the two-stage decoupled learning phase, specifically for three class sets: few, medium, and many. With ImageNet-LT as an example (Figure 4), we observe that $\boldsymbol{\rho}$ increases with each epoch, suggesting that the wrapped Cauchy distribution becomes increasingly tight. During representation learning, different class frequencies correspond to different values of ρ. On average, the 'Few' class exhibits a larger ρ while the 'Many' class shows a smaller ρ (*Appendix Figure 7*). Larger ρ values lead to greater margins during training (Theorem 4).

Prior research has established that both tighter feature clustering (Kobayashi, 2021) and larger margins (Cao et al., 2019; Ren et al., 2020) enhance classification results, especially for tail classes (Cao et al., 2019). Our findings are consistent with these studies (Cao et al., 2019; Kobayashi, 2021; Ren et al., 2020). The frequency-dependent disparity in ρ decreases in classifier learning due to the use of the class-balanced sampler (Kang et al., 2020). It's also notable

Dataset	CIFAR-100-LT			CIFAR-10-LT		
Imbalance factor	200	100	10	200	100	10
Focal loss (Lin et al., 2017)	40.2 ± 0.5	43.8 ± 0.1	60.0 ± 0.6	71.8 ± 2.1	77.1 ± 0.2	90.3 ± 0.2
LDAM loss (Cao et al., 2019)	41.3 ± 0.4	46.1 ± 0.1	62.1 ± 0.3	73.6 ± 0.1	78.9 ± 0.9	90.3 ± 0.1
cRT (Kang et al., 2020)	44.5 ± 0.1	50.0 ± 0.2	63.3 ± 0.1	76.6 ± 0.2	82.0 ± 0.2	91.0 ± 0.0
LWS (Kang et al., 2020)	45.3 ± 0.1	50.5 ± 0.1	63.4 ± 0.1	78.1 ± 0.0	83.7 ± 0.0	91.1 ± 0.0
BALMS (Ren et al., 2020)	45.5 ± 0.0	50.8 ± 0.0	63.0 ± 0.0	81.5 ± 0.0	84.9 ± 0.0	91.3 ± 0.0
Angular-based Softmax						
Angular Softmax	44.2 ± 0.5	49.7 ± 0.6	64.1 ± 0.2	80.9 ± 0.2	83.8 ± 0.2	91.4 ± 0.1
L-Softmax (Liu et al., 2016)	46.2 ± 0.2	51.3 ± 0.2	64.8 ± 0.1	79.9 ± 0.4	85.0 ± 0.2	91.8 ± 0.1
AM-Softmax (Deng et al., 2019)	45.4 ± 0.4	50.1 ± 0.1	63.9 ± 0.2	77.5 ± 0.4	81.6 ± 0.5	90.9 ± 0.7
t-vMF Similarity (Kobayashi, 2021)	46.2 ± 0.2	50.3 ± 0.5	64.7 ± 0.2	80.9 ± 0.3	83.8 ± 0.3	91.2 ± 0.3
WCDAS (ours)	$\mathbf{49.3} \pm 0.1$	$\mathbf{52.5} \pm 0.1$	$\mathbf{65.8} \pm 0.1$	$\mathbf{81.7} \pm 0.1$	$\mathbf{86.4} \pm 0.3$	$\mathbf{92.4} \pm 0.2$

Table 3: Top 1 accuracy (mean $\pm$ SD) for CIFAR-10/100-LT training with ResNet32 (He et al., 2016). Results of Angular Softmax (Eq. 1), L-Softmax, AM-Softmax and t-vMF Similarity are reproduced with optimal hyper-parameters reported in their original papers. WCDAS generally outperforms SOTA methods.

Dataset	ImageNet-LT				iNaturalist 2018			
	Many	Medium	Few	All	Many	Medium	Few	All
OLTR (Liu et al., 2019)	43.4	35.0	18.5	35.5	65.7	66.3	63.4	65.2
Center loss (Wen et al., 2016)	53.0	35.1	15.6	39.1	71.7	66.0	60.4	64.3
cRT (Kang et al., 2020)	49.9	37.5	23.0	40.3	70.9	67.0	66.4	67.3
LWS (Kang et al., 2020)	48.0	37.5	22.9	39.6	69.0	68.2	66.6	67.7
BALMS (Ren et al., 2020)	48.0	38.3	22.9	39.9	66.8	67.4	67.9	68.1
Angular-based Softmax								
Angular Softmax (Eq. 1)	52.8	33.9	15.7	38.7	71.8	65.3	61.4	65.0
L-Softmax (Liu et al., 2016)	54.0	35.1	15.4	39.1	72.7	66.1	60.1	64.5
AM-Softmax (Deng et al., 2019)	54.2	36.0	16.7	40.3	73.1	67.3	61.9	65.9
t-vMF Similarity (Kobayashi, 2021)	55.4	39.9	22.5	43.5	75.1	**72.2**	69.7	71.0
WCDAS (class-balanced)	**56.2**	40.9	24.1	**44.5**	75.5	72.3	69.8	**71.8**
WCDAS (meta)	53.8	**41.7**	**25.3**	44.1	71.4	**72.3**	**70.5**	70.8

Table 4: Top 1 accuracy for ImageNet-LT (ResNet10 (He et al., 2016)) and iNaturalist 2018 (ResNet50 (He et al., 2016)). Results are reproduced with the same settings of our method (*Appendix Table 5*). Comparison of original results are provided in *Appendix Table 8* together with more SOTA methods included.

that class-dependent ρ values can be observed across all the tested datasets (Figure 5). Moreover, our method shows that even with different initial positions, the aforementioned pattern holds true and ρ tends to converge to similar values (*Appendix Figure 7*), demonstrating stability during training.

4.4. Comparing with SOTAs

We performed an extensive comparison of our method with state-of-the-art (SOTA) softmax-based methods designed for long-tail recognition on CIFAR-10/100-LT (Table 3), ImageNet-LT (Table 4), and iNaturalist 2018 (Table 4). In addition, we included several leading angular-based softmax approaches for comparison, adhering to the same decoupled two-step training procedures. The class-balanced sampler was used for classifier learning in these methods. To ensure a fair comparison, our method also utilized the same sampler.

A more detailed discussion about the choice of sampler is provided in *Appendix B.4*.

Given that WCDAS requires a larger learning rate (0.4) for ImageNet-LT and iNaturalist 2018, we sought to exclude the possibility that the superior results of our model could be attributed to the larger learning rate. To do this, we present two tables: one with SOTA methods reproduced using a learning rate of 0.4 (Table 4), and the other featuring results directly obtained from the original papers (*Appendix Table 8*). Both Table 4 and *Appendix Table 8* indicate that our method achieves better accuracy than other competing softmax-based methods.

Furthermore, the improvement is particularly noticeable in the tail class, which consists of fewer samples. For instance, our method improved the accuracy from 22.5 to 25.3 for ImageNet-LT, without compromising the head

class. Previous works often sacrificed other classes in the process of improving accuracy (Kang et al., 2020; Ren et al., 2020). This improvement was even more evident on CIFAR-100/10-LT, likely because fewer samples per class are more susceptible to noise, an aspect our method accounts for.

5. Conclusion

We've introduced the WCDAS approach for long-tail visual recognition tasks. Generally, WCDAS outperforms state-of-the-art (SOTA) softmax-based methods across all four datasets. The symmetric-wrapped stable (SWS) family incorporates a wide variety of distributions, each with its unique properties (Jammalamadaka & SenGupta, 2001). Our work expands the understanding of their utility in various contexts and challenges established methods, such as vMF. Four distinct *advantages* distinguish our method from previous works (Kobayashi, 2021; Cao et al., 2019) and contribute to its superior performance: (1) WCDAS accommodates out-of-distribution "imperfect" data due to its heavy tail, while still ensuring compact intra-class feature clustering. (2) WCDAS operates like a large margin angular softmax when ρ is large. As ρ increases during training, our loss function aligns with the classification task's cross-entropy loss. We provide a visualization of the loss surface with respect to ρ and θ. (3) Our empirical study shows that tail classes have larger ρ, leading to more compact clusters and larger margins (Theorem 4). Previous studies have confirmed the significant performance benefits of these factors (Cao et al., 2019). (4) Our method, unlike previous user-defined parameter approaches (Kobayashi, 2021), achieves optimal performance with trainable ρ.

However, WCDAS has *limitations*: it may necessitate a different learning rate or number of epochs compared to other methods, indicating a need for parameter re-tuning. Although WCDAS displays label-aware behaviors for all tested datasets, our paper does not offer a theoretical proof for this.

Looking to the future, WCDAS can serve as the softmax function replacement in deep learning models, improving other deep learning methods, such as mixture-of-experts (Wang et al., 2021b; Zhang et al., 2021b), and contrastive learning-based methods (Cui et al., 2021; Wang et al., 2021a). Additionally, WCDAS could potentially be applicable to long-tail video recognition (Zhang et al., 2021a) and long-tail object detection (Feng et al., 2021) with minimal or no adjustments. However, further validation is necessary for these domains.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2021.05.008. URL https://www.sciencedirect.com/science/article/pii/S1566253521001081.

Bailey, J. and Codling, E. Emergence of the wrapped cauchy distribution in mixed directional data. *AStA Advances in Statistical Analysis*, 105, 10 2020. doi: 10.1007/s10182-020-00380-7.

Bridle, J. S. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Proceedings of the 2nd International Conference on Neural Information Processing Systems*, NIPS'89, pp. 211–217, Cambridge, MA, USA, 1989. MIT Press.

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.

Cao, K., Chen, Y., Lu, J., Arechiga, N., Gaidon, A., and Ma, T. Heteroskedastic and imbalanced deep learning with adaptive regularization. In *International Conference on Learning Representations*, 2021.

Chang, J., Lan, Z., Cheng, C., and Wei, Y. Data uncertainty learning in face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5709–5718, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00575. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00575.

Cui, J., Zhong, Z., Liu, S., Yu, B., and Jia, J. Parametric contrastive learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 695–704, 2021. doi: 10.1109/ICCV48922.2021.00075.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *CVPR*, 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, 2009. doi: 10.1109/CVPR.2009.5206848.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition (CVPR), pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

Feng, C., Zhong, Y., and Huang, W. Exploring classification equilibrium in long-tailed object detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3397–3406, 2021.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010. URL https://proceedings.mlr.press/v9/glorot10a.html.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Han, H., Wang, W.-Y., and Mao, B.-H. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In Huang, D.-S., Zhang, X.-P., and Huang, G.-B. (eds.), *Advances in Intelligent Computing*, pp. 878–887, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31902-3.

Hayat, M., Khan, S., Zamir, S. W., Shen, J., and Shao, L. Gaussian affinity for max-margin class imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Holmström, L. The accuracy and the computational complexity of a multivariate binned kernel density estimator. *J. Multivar. Anal.*, 72(2):264–309, feb 2000. ISSN 0047-259X. doi: 10.1006/jmva.1999.1863. URL https://doi.org/10.1006/jmva.1999.1863.

Hong, F., Yao, J., Zhou, Z., Zhang, Y., and Wang, Y. Long-tailed partial label learning via dynamic rebalancing. In *ICLR*, 2023.

Huang, C., Li, Y., Loy, C. C., and Tang, X. Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5375–5384, 2016. doi: 10.1109/CVPR.2016.580.

Huang, C., Li, Y., Loy, C. C., and Tang, X. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2781–2794, 2020. doi: 10.1109/TPAMI.2019.2914680.

Jammalamadaka, S. R. and SenGupta, A. *Topics in Circular Statistics*, volume 5. World Scientific Publishing Co. Pte. Ltd., 2001. ISBN 9789812779267. doi: 10.1142/9789812779267. URL http://dx.doi.org/10.1142/9789812779267.

Jingru Tan, Changbao Wang, B. L. Q. L. W. O. C. Y. J. Y. Equalization loss for long-tailed object recognition. In *ArXiv:2003.05176*, 2020.

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.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kobayashi, T. t-vmf similarity for regularizing in-class feature distribution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Kubát, M. and Matwin, S. Addressing the curse of imbalanced training sets: One-sided selection. In *ICML*, 1997.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollar, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Liu, W., Wen, Y., Yu, Z., and Yang, M. Large-margin softmax loss for convolutional neural networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 507–516, 2016.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Parzen, E. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33 (3):1065–1076, 1962. ISSN 00034851. URL http://www.jstor.org/stable/2237880.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

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.

Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832 – 837, 1956. doi: 10.1214/aoms/1177728190. URL https://doi.org/10.1214/aoms/1177728190.

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. In *NeurIPS*, 2019.

Tong Wu, Ziwei Liu, Q. H. Y. W. and Lin, D. Adversarial robustness under long-tailed distribution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018. doi: 10.1109/CVPR.2018.00914.

Wang, P., Han, K., Wei, X.-S., Zhang, L., and Wang, L. Contrastive learning based hybrid networks for long-tailed image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Wang, X., Lian, L., Miao, Z., Liu, Z., and Yu, S. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021b.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. A discriminative feature learning approach for deep face recognition. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *Computer Vision – ECCV 2016*, pp. 499–515, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46478-7.

Yang, S., Liu, L., and Xu, M. Free lunch for few-shot learning: Distribution calibration. In *International Conference on Learning Representations (ICLR)*, 2021.

Ye, H.-J., Chen, H.-Y., Zhan, D.-C., and Chao, W.-L. Identifying and compensating for feature deviation in imbalanced deep learning. *ArXiv*, abs/2001.01385, 2020.

Zhang, X., Wu, Z., Weng, Z., Fu, H., Chen, J., Jiang, Y.-G., and Davis, L. S. Videolt: Large-scale long-tailed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7960–7969, October 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, 2023.

A. Proofs and Derivations

A.1. Proof to Theorem 1

$$h(\rho, \theta; m, j) = \frac{1}{2\pi} \left(1 + 2 \sum_{n=1}^{\infty} \rho_m^{n^2} \cos n(\theta_m - \mu_m) \right), n \in \mathbb{N} \tag{14}$$

$$h(\rho, \theta; m, j) = \frac{1}{2\pi} \left(1 + 2 \sum_{n=1}^{\infty} \rho_m^{n^2} (\sin n\mu_m \sin n\theta_m + \cos n\mu_m \cos n\theta_m) \right) \tag{15}$$

Given that $\mu_m \to 0$, then $\sin n\mu_m$ can be approximated as $n\mu_m$ and $\cos n\mu_m$ can be approximated as 1. Therefore:

$$h(\rho, \theta; m, j) = \frac{1}{2\pi} \left(1 + 2 \sum_{n=1}^{\infty} \rho_m^{n^2} (n\mu_j \sin n\theta_m + \cos n\theta_m) \right) \tag{16}$$

Therefore, mixed distribution $f_{\text{mixed}}(\theta; j)$ can be written as:

$$f_{\text{mixed}}(\theta; j) = \frac{1}{M_j} \left(\underbrace{\frac{1}{2\pi} \left(1 + 2 \sum_{n=1}^{\infty} \rho_1^{n^2} (n\mu_1 \sin n\theta_1 + \cos n\theta_1) \right) + \cdots + \frac{1}{2\pi} \left(1 + 2 \sum_{n=1}^{\infty} \rho_{M_j}^{n^2} (n\mu_{M_j} \sin n\theta_{M_j} + \cos n\theta_{M_j}) \right)}_{M_j} \right) \tag{17}$$

$$f_{\text{mixed}}(\theta; j) = \frac{1}{2\pi} + \frac{1}{M_j} \underbrace{\left(2 \sum_{n=1}^{\infty} (n\rho_1^{n^2} \mu_1 \sin n\theta_1 + \cdots + n\rho_{M_j}^{n^2} \mu_{M_j} \sin n\theta_{M_j}) \right)}_{M_j} \tag{18}$$

$$+ \frac{1}{M_j} \underbrace{\left(2 \sum_{n=1}^{\infty} (\rho_1^{n^2} \cos n\theta_1 + \cdots + \rho_{M_j}^{n^2} \cos n\theta_{M_j}) \right)}_{M_j} \tag{19}$$

Given that μ follows $\mathcal{N}(0, \sigma)$, Therefore we can further approximate:

$$\frac{1}{M_j} \underbrace{\left(2 \sum_{n=1}^{\infty} (n\rho_1^{n^2} \mu_1 \sin n\theta_1 + \cdots + n\rho_{M_j}^{n^2} \mu_{M_j} \sin n\theta_{M_j}) \right)}_{M_j} \to 0 \tag{20}$$

Subsequently, we obtain:

$$f_{\text{mixed}}(\theta; j) = \frac{1}{2\pi M_j} \sum_{m=1}^{M_j} \left(1 + 2 \sum_{n=1}^{\infty} \rho_m^{n^2} \cos n\theta_m \right) \tag{21}$$

A.2. Proof to Corollary 1.1

$$f_{\text{mixed}}(\theta; j) = \frac{1}{2\pi M_j} \sum_{m=1}^{M_j} \left(1 + 2 \sum_{n=1}^{\infty} \rho_m^{n^2} \cos n\theta_m \right) \tag{22}$$

$$= \frac{1}{2\pi} \left(1 + 2 \sum_{n=1}^{\infty} \frac{1}{M_j} \sum_{m=1}^{M_j} \rho_m^{n^2} \cos n\theta_m \right) \tag{23}$$

Therefore, we get:

$$f_{\text{mixed}}(\theta; j) = \frac{1}{2\pi} \left(1 + 2 \sum_{n=1}^{\infty} \alpha_{\text{mixed}}^{\{n\}} \cos n\theta_m \right) \tag{24}$$

$$\tag{25}$$

where α_{mixed} is the cosine moment of the mixed distribution:

$$\alpha_{\text{mixed}}^{\{n\}} = \frac{1}{M_j} \sum_{m=1}^{M_j} \rho_m^{n^2} = \frac{1}{M_j} \sum_{m=1}^{M_j} \alpha_m^{\{n\}} \tag{26}$$

A.3. Proof to Theorem 2

$$\Delta_\rho = \frac{1}{\pi} \sum_{n=1}^{\infty} \left((\alpha_{\text{fit}}^{\{n\}} - \alpha_{\text{mixed}}^{\{n\}}) \cos n\theta \right)^2 \tag{27}$$

For any θ and n, to minimize Δ_ρ, it is equivalently as minimizing $\sum_{n=1}^{\infty} (\alpha_{\text{fit}}^{\{n\}} - \alpha_{\text{mixed}}^{\{n\}})^2$:

$$\Delta_\rho = \frac{1}{\pi} \sum_{n=1}^{\infty} (\alpha_{\text{fit}}^{\{n\}} - \alpha_{\text{mixed}}^{\{n\}})^2 \tag{28}$$

For wrapped Cauchy distribution, $\alpha_{\text{fit}}^{\{n\}} = \rho_{\text{WC}}^n, \forall \rho \in \mathbb{R}$. For wrapped Normal distribution, $\alpha_{\text{fit}}^{\{n\}} = \rho_{\text{WN}}^{n^2}, \forall \rho \in \mathbb{R}$. $\alpha_{\text{mixed}}^{\{n\}} = \frac{1}{M_j} \sum_{m=1}^{M_j} \rho_m^{n^2}$ according to Corollary 1.1. Without losing the generosity, we derive the case of wrapped Cauchy distribution as an example:

$$\Delta_{\rho_{\min}, \text{WC}} = \frac{1}{\pi} \sum_{n=1}^{\infty} (\rho_{\min}^n - \frac{1}{M_j} \sum_{m=1}^{M_j} \rho_m^{n^2})^2 = \frac{1}{\pi} \sum_{n=1}^{\infty} (\rho_{\min}^n - \alpha_{\text{mixed}}^{\{n\}})^2 \tag{29}$$

$(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}} = (\frac{1}{M_j} \sum_{m=1}^{M_j} \rho_m^{n^2})^{\frac{1}{n}}$ can be treated as the n-th component of a cluster. To minimize the error $\Delta_{\rho_{\min}, \text{wc}}$, $\rho_{\min}$ is the centroid of the cluster composed of n number of $(\frac{1}{M_j} \sum_{m=1}^{M_j} \rho_m^{n^2})^{\frac{1}{n}}$. Therefore,

$$\rho_{\min} = \mathbb{E}_{n \in [1,\infty)} (\frac{1}{M_j} \sum_{m=1}^{M_j} \rho_m^{n^2})^{\frac{1}{n}} = \mathbb{E}_{n \in [1,\infty)} (\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}} \tag{30}$$

Take Equation 30 back into Equation 29 substituting $\rho_{\min}^n$, we approximate:

$$\Delta_{\rho_{\min},\text{wc}} = \frac{1}{\pi} \sum_{n=1}^{\infty} \left((\mathbb{E}_{n \in [1,\infty)}(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}})^n - \alpha_{\text{mixed}}^{\{n\}} \right)^2$$

$$= \frac{1}{\pi} \sum_{n=1}^{\infty} \left((\mathbb{E}_{n \in [1,\infty)}(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}})^n - ((\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}})^n \right)^2 \tag{31}$$

According to Binomial Theorem:

$$x^n - a^n = (x - a)(a^{n-1} + xa^{n-2} + \cdots x^{n-2}a + x^{n-1}). \tag{32}$$

Let $\overline{a} = \mathbb{E}_{n \in [1,\infty)}(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}}$ and $a_n = (\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}}$, we simplify Equation 31 into:

$$\Delta_{\rho_{\min},\text{WC}} \propto \sum_{n=0}^{\infty} (\overline{a}^n - a_n^n)^2$$

$$\propto (\overline{a} - a_1)^2 + (\overline{a} - a_2)^2(\overline{a} + a_2)^2 + \cdots$$

$$+ (\overline{a} - a_{n-1})^2(a^{n-1} + \overline{a}a^{n-2} + \cdots \overline{a}^{n-2}a + \overline{a}^{n-1})^2. \tag{33}$$

We expand the Equation 31 based on Equation 33:

$$\Delta_{\rho_{\min},\text{WC}} = \frac{1}{\pi} \sum_{n=1}^{\infty} \left(\mathbb{E}_{n \in [1,\infty)}(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}} - (\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}} \right)^2 \left((\mathbb{E}_{n \in [1,\infty)}(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}})^{n-1} + \cdots + (\alpha_{\text{mixed}}^{\{n\}})^{\frac{n-1}{n}} \right)^2 \tag{34}$$

Because $\rho_m \in [0,1)$, $\alpha_{\text{mixed}}^{\{n\}} \in [0,1)$ and the value of $\alpha_{\text{mixed}}^{\{n\}}$ decrease as n increases. Therefore, higher order terms in Equation 34 can be neglected ($n>1$). Accordingly, we get:

$$\Delta_{\rho_{\min},\text{WC}} \sim \frac{1}{\pi} \left(\mathbb{E}_{n \in [1,\infty)}(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}} - (\alpha_{\text{mixed}}^{\{1\}}) \right)^2 + \mathcal{O}(n)$$

$$\sim \frac{1}{\pi} \left(\mathbb{E}_{n \in [1,\infty)}(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}} - \frac{1}{M_j} \sum_{m=1}^{M_j} \rho_m \right)^2 + \mathcal{O}(n)$$

$$\propto SD_{n=1}(\frac{1}{M_j} \sum_{m=1}^{M_j} \rho_m^{n^2})^{\frac{1}{n}} \tag{35}$$

Following a similar derivation, when $\alpha_{\text{fit}}^{\{n\}} = \rho^{n^2}$:

$$\Delta_{\rho_{\min},\text{WN}} \sim \frac{1}{\pi} \left(\mathbb{E}_{n \in [1,\infty)}(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n^2}} - (\alpha_{\text{mixed}}^{\{1\}}) \right)^2 + \mathcal{O}(n)$$

$$\sim \frac{1}{\pi} \left(\mathbb{E}_{n \in [1,\infty)}(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n^2}} - \frac{1}{M_j} \sum_{m=1}^{M_j} \rho_m \right)^2 + \mathcal{O}(n)$$

$$\propto SD_{n=1}(\frac{1}{M_j} \sum_{m=1}^{M_j} \rho_m^{n^2})^{\frac{1}{n^2}} \tag{36}$$

A.4. Proof to Theorem 3

Let ρ_m of individual $h(\rho, \theta; m, j)$ distribute uniformly across its defined domain $[0,1)$. Assuming that we have N number of $h(\rho, \theta; m, j)$ (i. e., $M_j = N$), then $\rho_1 = 1/N, \rho_1 = 2/N, \cdots, \rho_N = (N-1)/N$.

$$\alpha_{\text{mixed}}^{\{n\}} = \frac{1}{N} \sum_{m=1}^{M} \left(\frac{m}{N}\right)^{n^2} \tag{37}$$

Given Faulhaber's formula, which is:

$$\sum_{k=1}^{N} k^p = \frac{N^{p+1}}{p+1} + \frac{1}{2}N^p + \sum_{k=2}^{p} \frac{B_k}{k!} \frac{p!}{(p-k+1)!} N^{p-k+1} \tag{38}$$

The coefficients involve Bernoulli numbers B_j. For each n, we get:

For $n = 1$,

$$\alpha_{\text{mixed}}^{\{1\}} = \frac{1}{N} \sum_{k=1}^{N} \rho = \frac{1}{N} \sum_{k=1}^{N} \frac{k}{N} = (1 + \frac{1}{N})/2 \tag{39}$$

For $n = 2$,

$$\alpha_{\text{mixed}}^{\{2\}} = \frac{1}{N} \sum_{k=1}^{N} \rho^4 = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{k}{N}\right)^4 = \frac{1}{5} + \frac{1}{2}\frac{1}{N} + \frac{1}{3}\frac{1}{N^2} - \frac{1}{30}\frac{1}{N^4} \tag{40}$$

For $n = 3$,

$$\alpha_{\text{mixed}}^{\{3\}} = \frac{1}{N} \sum_{k=1}^{N} \rho^9 = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{k}{N}\right)^9 = \frac{1}{10} + \frac{1}{2}\frac{1}{N} + \frac{3}{4}\frac{1}{N^2} - \frac{7}{10}\frac{1}{N^4} + \frac{1}{2}\frac{1}{N^6} - \frac{3}{20}\frac{1}{N^8} \tag{41}$$

$$\cdots \tag{42}$$

For $n = n_0$,

$$\alpha_{\text{mixed}}^{\{n\}} = \frac{1}{N} \sum_{k=1}^{N} \rho^{n_0^2} = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{k}{N}\right)^{n_0^2} = \frac{1}{n_0^2 + 1} + \frac{1}{2N} + \frac{1}{N} \sum_{k=2}^{n_0^2} \frac{B_k}{k!} \frac{n_0^2!}{(n_0^2 - k + 1)!} \frac{1}{N^{k-1}} \tag{43}$$

Given $N > 1$, then $\alpha_{\text{mixed}}^{\{n\}} \in (0, 1)$ (Equation 39 - 43). Hence, $(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}}$ shows more "variance" than $(\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n^2}}$ given $\alpha_{\text{mixed}}^{\{n\}} \in (0, 1)$. Therefore, $(\mathbb{E}_{n \in [1,\infty)} (\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n}} - \alpha_{\text{mixed}}^{\{1\}})^2 < (\mathbb{E}_{n \in [1,\infty)} (\alpha_{\text{mixed}}^{\{n\}})^{\frac{1}{n^2}} - \alpha_{\text{mixed}}^{\{1\}})^2$. According to Equation 35 and Equation 36:

$$\Delta_{\rho_{\min}, \text{WC}} < \Delta_{\rho_{\min}, \text{WN}} \tag{44}$$

A.5. Proof to Theorem 4

$$\|f_{\text{wc}}(\theta_1) - f_{\text{wc}}(\theta_2)\| = \left\| \frac{1 - \rho^2}{2\pi(1 + \rho^2 - 2\rho\cos\theta_1)} - \frac{1 - \rho^2}{2\pi(1 + \rho^2 - 2\rho\cos\theta_2)} \right\| \tag{45}$$

$$= \frac{1 - \rho^2}{2\pi} \left\| \frac{2\rho\cos\theta_1 - \cos\theta_2}{(1 + \rho^2 - 2\rho\cos\theta_1)(1 + \rho^2 - 2\rho\cos\theta_2)} \right\| \tag{46}$$

$$\geq \frac{1 - \rho^2}{2\pi} \frac{2\rho\|\cos\theta_1 - \cos\theta_2\|}{(1 + \rho^2 - 2\rho)(1 + \rho^2 - 2\rho)} \tag{47}$$

Simplified the above equation as follows:

$$\frac{1-\rho^2}{2\pi}\frac{2\rho\|\cos\theta_1-\cos\theta_2\|}{(1+\rho^2-2\rho)(1+\rho^2-2\rho)} \tag{48}$$

$$=\frac{\rho+\rho^2}{\pi(1-\rho)^3}\|\cos\theta_1-\cos\theta_2\| \tag{49}$$

In order to show a larger margin, it needs to satisfy the following condition:

$$\frac{\rho+\rho^2}{\pi(1-\rho)^3}\|\cos\theta_1-\cos\theta_2\| \geq \|\cos\theta_1-\cos\theta_2\| \tag{50}$$

$$\frac{\rho+\rho^2}{\pi(1-\rho)^3} \geq 1 \tag{51}$$

Solving Equation 51, we get $\rho \geq 0.42332$. Additionally, notice from Equation 51 that the larger ρ is, the larger the margin is (Figure 6).
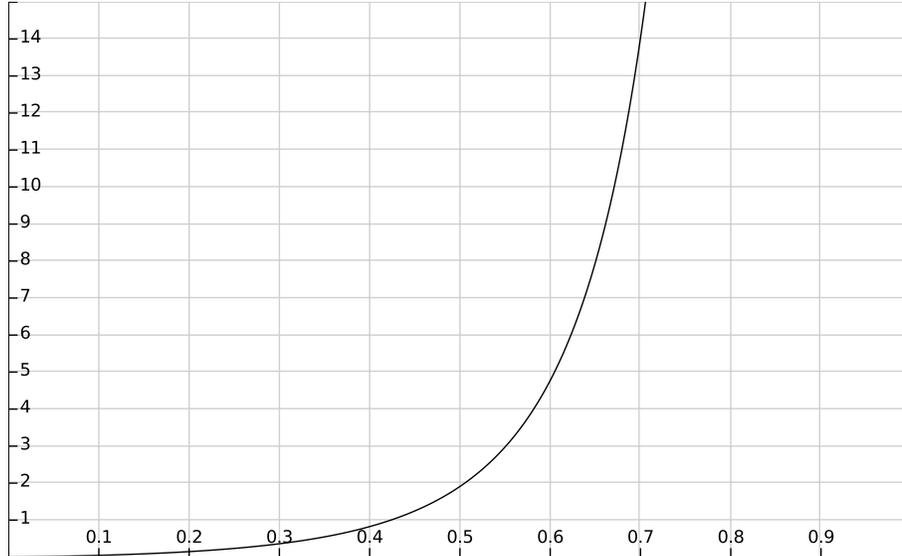

Figure 6: plot of $\frac{\rho+\rho^2}{\pi(1-\rho)^3}$ (Y-axis) with respect to ρ (X-axis).

B. Supplementary Results

B.1. Experiment settings

CIFAR10-LT and CIFAR100-LT: CIFAR10-LT and CIFAR100-LT contain 10 and 100 classes, respectively. Various imbalance factors (10-200) are evaluated. an imbalance factor β is calculated by $\beta = \frac{M_{max}}{M_{min}}$ where M_{max} and M_{min} are the numbers of training samples for the most and least frequent classes respectively. We employ the ResNet-32 backbone for these two datasets, similar to previous works. Given that CIFAR-LT 10/100 tends to show large variances in performance results, as stated in (Ren et al., 2020), we, therefore, report the mean and standard error from 3 independent replicas.

ImageNet-LT: It contains 1000 classes, and the number of images per class ranges from 1280 to 5 images with an imbalance factor of 256. ResNet-10 and ResNext-50 backbones are used for the experiments. ImageNet-LT is also used for various ablation studies.

iNaturalist 2018: It is a naturally imbalanced fine-grained dataset with 8,142 categories, following the long-tailed distribution. The number of images per class ranges from 1000 to 2, with an imbalance factor of 500. We use ResNet-50 as the backbone and apply the same training settings as for ImageNet-LT except batch size 512.

Evaluation Setup. After training on the long-tailed dataset, we evaluate the models on the corresponding balanced test/validation dataset and report top-1 accuracy. To give further insight, we report accuracy on three splits of the set of classes for ImageNet-LT and iNaturalist 2018: Many-shot (>100 images), Medium-shot (20-100 images), and Few-shot (<20 images), adopting from OLTR (Liu et al., 2019).

Hyperparameters for the best performance. Backbones and hyper-parameters of our method used for all datasets are listed in Table 5.

Datasets	Epochs	lr (representation/classifier)	Backbone	Init.	s
CIFAR100-LT	300	0.2/0.2	ResNet-32	0.	trainable (Kobayashi, 2021)
CIFAR10-LT	300	0.2/0.2	ResNet-32	0.	trainable(Kobayashi, 2021)
ImageNet-LT	90	0.4/0.2	ResNet-10	-1.	trainable(Kobayashi, 2021)
iNaturalist 2018	200	0.4/0.2	ResNet-50	1.	250

Table 5: Choice of hyper-parameter in all datasets. lr: Initial learning rate of GSD with cosine scheduler. Init: Initialization of w_ρ. Trainable s are implemented following (Kobayashi, 2021)

B.2. Impact of epoch number.

As Table 6 shows, when using a learning rate of 0.2, the overall performance of our method improves with more training epochs, indicating inadequate training. However, we note that such an improvement is attributed to the accuracy improvement of Class "Many". Meanwhile, the accuracy of Class "Few" decreases slightly with more training epochs. It is likely due to the fact that the model weighs more on high-frequency classes with longer training time. Therefore, we increase the learning rate while the same training epoch (Table 2 in *Main text*).

	90 epochs				150 epochs			
Initialization	Many	Medium	Few	All	Many	Medium	Few	All
2.0	55.3	40.4	22.9	43.7	56.9	40.6	22.1	44.3
1.0	55.1	40.2	23.0	43.7	56.9	40.6	22.0	44.3
0	55.0	40.1	22.4	43.3	56.3	39.8	21.7	43.9
-1.0	55.6	40.3	22.8	43.7	57.4	40.8	21.7	44.5
-2.0	55.4	40.3	22.3	43.5	56.3	40.5	22.2	44.0

Table 6: Top 1 accuracy for ImageNet-LT (ResNet-10) with various w_ρ initialization (Init.) values. Initial learning rate: 0.2.

B.3. Class-wise ρ optimization

Convergence of ρ. Regardless of initialization, ρ are able to converge to similar values (Figure 7), demonstrating our method is robust against initialization.

Sampling method	Many	Medium	Few	All
Class balanced sampling (Kang et al., 2020)	56.2	40.9	24.1	44.5
Meta-sampling (Ren et al., 2020) (lr = 0.005)	54.0	42.0	23.0	44.0
Meta-sampling (Ren et al., 2020) (lr = 0.01)	53.8	41.7	25.3	44.1
Meta-sampling (Ren et al., 2020) (lr = 0.05)	52.3	41.2	27.7	43.7

Table 7: Top 1 accuracy for ImageNet-LT (ResNet-10) with different sampler in classifier learning. We use 3 different learning rates in meta sampling.

B.4. Impact of the sampler in decoupled training

The sampler is demonstrated to be critical when training with an imbalanced dataset, especially in classifier learning. To assess which sampler yields better performance for WCDAS, we compare two predominant sampling approaches: class-balanced sampler and meta sampler. For a fair comparison, we conducted three experiments with a meta sampler using different learning rates. Table 7 shows that a class-balanced sampler consistently shows better results than a meta sampler
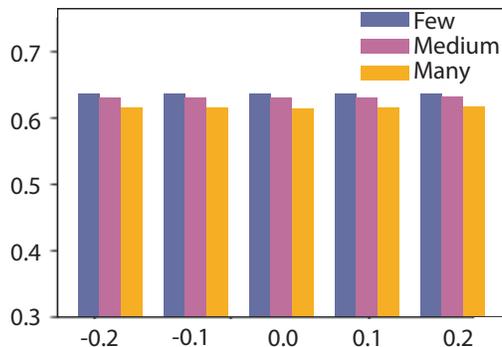
Figure 7: Bar graph of ρ values at 90th epoch with respect to different weight initialization values. Three sets of class are plotted include few (<20), medium (20-100) and many (>100).

when considering all classes. However, a meta-sampler provides a more balanced accuracy across classes with medium or few examples.

B.5. Comparison with selected SOTA methods using same settings: large learning rate.

Table 8 shows the comparison of our method with SOTA softmax-based methods. We note that the results are directly copied from the original paper. Our method shows superior performance. We also note that those methods show no improvement beyond the results from their original papers when applying a larger learning rate (*Table 4*), indicating that a learning rate of 0.2 is sufficient or optimal for those methods.

Dataset	Imagenet-LT				iNaturalist 2018			
	Many	Medium	Few	All	Many	Medium	Few	All
Focal loss (Lin et al., 2017)	36.4	29.9	16.0	30.5	-	-	-	61.1
OLTR (Liu et al., 2019)	43.2	35.1	18.5	35.6	65.9	66.3	63.6	65.4
Center loss (Wen et al., 2016)	53.1	35.0	15.6	39.2	71.5	66.0	61.8	65.8
cRT (Kang et al., 2020)	52.3	39.5	23.2	41.8	73.2	68.8	68.9	69.3
LWS (Kang et al., 2020)	-	-	-	41.4	71.5	71.3	69.7	70.7
LDAM loss (Cao et al., 2019)	-	-	-	36.1	-	-	-	64.6
τ-normalized (Kang et al., 2020)	51.9	38.3	22.5	40.6	71.1	68.9	69.3	69.3
BALMS (Ren et al., 2020)	50.3	39.5	**25.3**	41.8	-	-	-	-
Angular based Softmax								
L-Softmax (Liu et al., 2016)	53.7	35.1	16.4	39.5	71.2	66.3	60.9	64.7
AM-Softmax (Deng et al., 2019)	54.0	36.0	18.6	40.5	72.5	67.6	63.2	66.4
t-vMF Similarity (Kobayashi, 2021)	55.2	40.6	22.3	43.7	74.2	72.1	69.9	71.1
WCDAS (class-balanced)	**56.2**	40.9	24.1	**44.5**	**75.5**	**72.3**	69.8	**71.8**
WCDAS (meta)	53.8	**41.7**	**25.3**	44.1	71.4	**72.3**	**70.5**	70.8

Table 8: Top 1 accuracy for ImageNet-LT (ResNet10 (He et al., 2016)) and iNaturalist 2018 (ResNet50 (He et al., 2016)). Results are copied directly from the original papers.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2021.05.008. URL https://www.sciencedirect.com/science/article/pii/S1566253521001081.

Bailey, J. and Codling, E. Emergence of the wrapped cauchy distribution in mixed directional data. *AStA Advances in Statistical Analysis*, 105, 10 2020. doi: 10.1007/s10182-020-00380-7.

Bridle, J. S. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Proceedings of the 2nd International Conference on Neural Information Processing Systems*, NIPS'89, pp. 211–217, Cambridge, MA, USA, 1989. MIT Press.

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.

Cao, K., Chen, Y., Lu, J., Arechiga, N., Gaidon, A., and Ma, T. Heteroskedastic and imbalanced deep learning with adaptive regularization. In *International Conference on Learning Representations*, 2021.

Chang, J., Lan, Z., Cheng, C., and Wei, Y. Data uncertainty learning in face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5709–5718, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00575. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00575.

Cui, J., Zhong, Z., Liu, S., Yu, B., and Jia, J. Parametric contrastive learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 695–704, 2021. doi: 10.1109/ICCV48922.2021.00075.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *CVPR*, 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, 2009. doi: 10.1109/CVPR.2009.5206848.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

Feng, C., Zhong, Y., and Huang, W. Exploring classification equilibrium in long-tailed object detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3397–3406, 2021.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010. URL https://proceedings.mlr.press/v9/glorot10a.html.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Han, H., Wang, W.-Y., and Mao, B.-H. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In Huang, D.-S., Zhang, X.-P., and Huang, G.-B. (eds.), *Advances in Intelligent Computing*, pp. 878–887, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31902-3.

Hayat, M., Khan, S., Zamir, S. W., Shen, J., and Shao, L. Gaussian affinity for max-margin class imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Holmström, L. The accuracy and the computational complexity of a multivariate binned kernel density estimator. *J. Multivar. Anal.*, 72(2):264–309, feb 2000. ISSN 0047-259X. doi: 10.1006/jmva.1999.1863. URL https://doi.org/10.1006/jmva.1999.1863.

Hong, F., Yao, J., Zhou, Z., Zhang, Y., and Wang, Y. Long-tailed partial label learning via dynamic rebalancing. In *ICLR*, 2023.

Huang, C., Li, Y., Loy, C. C., and Tang, X. Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5375–5384, 2016. doi: 10.1109/CVPR.2016.580.

Huang, C., Li, Y., Loy, C. C., and Tang, X. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2781–2794, 2020. doi: 10.1109/TPAMI.2019.2914680.

Jammalamadaka, S. R. and SenGupta, A. *Topics in Circular Statistics*, volume 5. World Scientific Publishing Co. Pte. Ltd., 2001. ISBN 9789812779267. doi: 10.1142/9789812779267. URL http://dx.doi.org/10.1142/9789812779267.

Jingru Tan, Changbao Wang, B. L. Q. L. W. O. C. Y. J. Y. Equalization loss for long-tailed object recognition. In *ArXiv:2003.05176*, 2020.

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.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kobayashi, T. t-vmf similarity for regularizing in-class feature distribution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Kubát, M. and Matwin, S. Addressing the curse of imbalanced training sets: One-sided selection. In *ICML*, 1997.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollar, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Liu, W., Wen, Y., Yu, Z., and Yang, M. Large-margin softmax loss for convolutional neural networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 507–516, 2016.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Parzen, E. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3): 1065–1076, 1962. ISSN 00034851. URL http://www.jstor.org/stable/2237880.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

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.

Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27 (3):832 – 837, 1956. doi: 10.1214/aoms/1177728190. URL https://doi.org/10.1214/aoms/1177728190.

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. In *NeurIPS*, 2019.

Tong Wu, Ziwei Liu, Q. H. Y. W. and Lin, D. Adversarial robustness under long-tailed distribution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018. doi: 10.1109/CVPR.2018.00914.

Wang, P., Han, K., Wei, X.-S., Zhang, L., and Wang, L. Contrastive learning based hybrid networks for long-tailed image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Wang, X., Lian, L., Miao, Z., Liu, Z., and Yu, S. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021b.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. A discriminative feature learning approach for deep face recognition. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *Computer Vision – ECCV 2016*, pp. 499–515, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46478-7.

Yang, S., Liu, L., and Xu, M. Free lunch for few-shot learning: Distribution calibration. In *International Conference on Learning Representations (ICLR)*, 2021.

Ye, H.-J., Chen, H.-Y., Zhan, D.-C., and Chao, W.-L. Identifying and compensating for feature deviation in imbalanced deep learning. *ArXiv*, abs/2001.01385, 2020.

Zhang, X., Wu, Z., Weng, Z., Fu, H., Chen, J., Jiang, Y.-G., and Davis, L. S. Videolt: Large-scale long-tailed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7960–7969, October 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, 2023.