Harnessing Hierarchical Label Distribution Variations in Test Agnostic Long-tail Recognition

Zhiyong Yang[1] **Qianqian Xu**[2] **Zitai Wang**[3 4] **Sicong Li**[3 4] **Boyu Han**[2 1]
Shilong Bao[3 4] **Xiaochun Cao**[5] **Qingming Huang**[1 2 6]

Abstract

This paper explores test-agnostic long-tail recognition, a challenging long-tail task where the test label distributions are unknown and arbitrarily imbalanced. We argue that the variation in these distributions can be broken down hierarchically into global and local levels. The global ones reflect a broad range of diversity, while the local ones typically arise from milder changes, often focused on a particular neighbor. Traditional methods predominantly use a Mixture-of-Expert (MoE) approach, targeting a few fixed test label distributions that exhibit substantial global variations. However, the local variations are left unconsidered. To address this issue, we propose a new MoE strategy, DirMixE, which assigns experts to different Dirichlet meta-distributions of the label distribution, each targeting a specific aspect of local variations. Additionally, the diversity among these Dirichlet meta-distributions inherently captures global variations. This dual-level approach also leads to a more stable objective function, allowing us to sample different test distributions better to quantify the mean and variance of performance outcomes. Theoretically, we show that our proposed objective benefits from enhanced generalization by virtue of the variance-based regularization. Comprehensive experiments across multiple benchmarks confirm the effectiveness of DirMixE. The code is available at https://github.com/scongl/DirMixE.

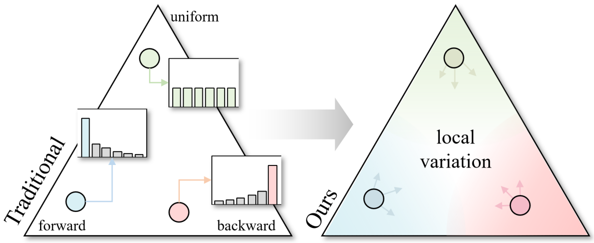


Figure 1. Traditional methods can only capture *global* variations of label distribution. By contrast, our DirMixE learns from both *global* and *local* variations, covering more test distributions.

1. Introduction

Traditional machine learning (ML) methods are usually designed for data with a balanced distribution (Krizhevsky et al., 2009; Deng et al., 2009). In contrast, most real-world data exhibits a long-tailed nature. Specifically, a few classes have a large number of samples (*i.e.*, head classes), while many others have far fewer (*i.e.*, tail classes) (Zhang et al., 2023b). This pattern produces profound studies in many tasks, including species classification (Horn et al., 2018; Miao et al., 2021), face recognition (Zhang et al., 2017; Zhong et al., 2019), medical image diagnosis (Galdran et al., 2021), social image understanding (Tang et al., 2016; Li et al., 2018), semantic segmentation (Zhong et al., 2023; Zhang et al., 2020), and object detection (Lin et al., 2017).

Hitherto, considerable methods have been developed to address this issue seeing its widespread nature. To evaluate model performance with long-tail distributions, most existing studies train their models on long-tailed data and test them on a balanced dataset (Cui et al., 2019; Cao et al., 2019; Kang et al., 2020; Menon et al., 2021; Kini et al., 2021b; Cui et al., 2021; Rangwani et al., 2022; Alshammari et al., 2022; Wang et al., 2023). This protocol poses an implicit assumption: the test distribution is balanced and fixed. However, this assumption is often invalid, as the test distribution is typically unknown and can change over time.

To reduce the bias in test distribution, Zhang et al. (2022) recently proposed a more practical setting called test-agnostic long-tailed recognition. This setting aims at ensuring model

[1] School of Computer Science and Tech., University of Chinese Academy of Sciences [2]Key Lab. of Intelligent Information Processing, Institute of Computing Tech., CAS [3]Institute of Information Engineering, CAS [4]School of Cyber Security, University of Chinese Academy of Sciences [5]School of Cyber Science and Tech., Shenzhen Campus of Sun Yat-sen University [6]BDKM, University of Chinese Academy of Sciences. Correspondence to: Qianqian Xu <xuqianqian@ict.ac.cn>, Qingming Huang <qmhuang@ucas.ac.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

efficacy across *heterogeneous* test distributions, each exhibiting varying levels of imbalance (ranging from long-tail and nearly uniform to inversely long-tail distributions). To address this challenge, they propose a novel mixture of experts approach called SADE, crafting distinct experts for each type of distribution. A typical feature of SADE is that each expert therein is assigned to *a fixed* label distribution. We argue that the random variation of the unseen test label distribution could be decomposed into **global** and **local** variations. The global variations capture the heterogeneous diversity across different distributions (say, the significant gap between long-tail and inverse long-tail distribution). The local variations capture the milder changes mostly concentrated within a neighbor. Back to SADE, the fixed assignment scheme could capture global variations by choosing sufficiently diverse target distributions for different experts. However, the local variations of label distributions can hardly be discovered by fixed points. In this sense, how to comprehensively utilize the hierarchical random variations remains an open problem. To address this issue, we propose a new mixture-of-experts strategy with the following contributions.

First, we explore how to better model the randomness of the unknown test label distributions in a hierarchical manner. To this end, we propose a probabilistic framework where the label distributions are presumed to be sampled from a meta-distribution. To capture the **global** variations, we formulate the meta-distribution as a mixture form of heterogeneous distributions. Herein, each component is a Dirichlet distribution capturing the **local** variations of a specific factor disentangled from the meta-distribution. Furthermore, we allocate a specific expert for each component of the meta-distribution to ensemble expert models enjoying diverse skills. The overall effect is illustrated in Fig.1.

On top of this, we take a step further to explore how to evaluate the *overall performance on the meta-distribution* and learn from it. We argue that the average performance is not stable enough for the complex random variations of the test label distributions. In this sense, we develop a Monte Carlo method for estimating the mean and variance of the averaged training loss obtained across different label distributions. From the optimization perspective, minimizing the variance term might prevent the model from performing better than the average level. We thus construct a semi-variance regularization as a surrogate for the empirical variance, selectively penalizing performances below the average level. The attained estimation forms the basis for our objective function in training the expert ensemble.

Finally, we dive into the *theoretical underpinnings for the generalization ability of our method.* We can attain a sharper upper bound of generalization error thanks to the variance-based regularization scheme. Moreover, it also indicates that Monte Carlo sampling offers better generalization than the previous method developed on fixed test distributions.

2. Related Work

This paper will provide a brief overview of several areas closely related to our study.

Loss Modification. The main problem with traditional methods is their poor performance on tail classes. A direct solution is modifying the loss function to enhance performance for these classes. Increasing the weights of tail-class losses is a common tactic (Morik et al., 1999; Cui et al., 2019) along this course. While seemingly reasonable, it can lead to unstable optimization (Cui et al., 2019; Cao et al., 2019; Wang et al., 2023) and ruin the overall performance. Cao et al. (2019) suggests using weighted terms mainly in the later stages of training. To further address the optimization issues caused by unequal weights, another line of research turns to adjust the logits through class-specific operations. For instance, Tan et al. (2020) observes that positive samples of one class can be negative for others, causing discouraging gradients for tail samples. To address this issue they proposed the equalized loss, adding class-dependent operations to ignore these gradients. LDAM imposes a larger margin penalty on tail-class logits for stronger regularization (Cao et al., 2019). The LA loss (Menon et al., 2021) uses additive adjustments on the logits to maintain the consistency of balanced error. Later, Kini et al. (2021a) combines additive and multiplicative terms (Ye et al., 2020) to form a unified approach called the VS loss. Most recently, Wang et al. (2023) provides a detailed generalization analysis of these loss-modification methods using a local contraction lemma.

Experts Ensembling. In our paper, we address the long-tail issue using a mixture of experts strategy. This approach is a specific form of ensemble learning that combines multiple models to enhance overall model generalization. Early long-tail ensemble learning methods (Zhou et al., 2020; Guo & Wang, 2021) use two network branches to handle long-tail and uniform distributions separately. During training, these branches' outputs are dynamically merged, shifting the focus progressively from head to tail classes. Li et al. (2020) finds that balanced datasets often outperform long-tail ones. As a result, they split the long-tail dataset into several more balanced subsets and designed networks with multiple branches. Since model diversity is crucial in ensemble learning, recent developments have turned their focus to improving experts with different skills. For instance, in ACE (Cai et al., 2021), three experts are created to learn from subsets containing all classes, middle+tail classes, and tail classes, respectively. RIDE (Wang et al., 2021) uses a KL-divergence based loss to encourage diversity among experts. SADE (Zhang et al., 2022) introduces a concept called test-agnostic long-tail recognition, which demands generalization across various test distributions. This is achieved by combining diverse experts trained with different logit adjustments using self-supervision. Finally, BalPoE (Aimar et al., 2023) suggests a balanced product of experts approach, blending multiple logit-adjusted experts using a product rule to suit the chosen test distribution.

In this paper, we construct a distributional mixture of experts model for the test-agnostic long-tail recognition task. Specifically, instead of using fixed test distributions, we adopt a Dirichlet mixture distribution to simultaneously model the global and local variations of the unseen label distribution. Moreover, our method also enjoys a sharp generalization bound.

3. Problem Formulation

In this study, we focus on the challenge of training a good model in the presence of a **long-tail** data distribution. We define our training data as $\mathcal{S} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$, sampled from distribution $\mathcal{D}$, with a total of N instances. Here, $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ represents the raw input feature of dimensionality d for instance i, and $y_i \in 1, 2, \cdots, C$ denotes the corresponding label in a C-class classification problem. Under this setting, if we decompose the joint data distribution as $\mathcal{D} = \mathbb{P}_{tr}[\boldsymbol{x}|y] \cdot \mathbb{P}_{tr}[y]$, then the label distribution is often highly skewed in the sense that $\max_i \mathbb{P}_{tr}[y = i] / \min_j \mathbb{P}_{tr}[y = j] >> 1$. This imbalance necessitates the design of specialized learning strategies to ensure adequate performance in the tail classes.

Test Agnostic Long-tail Recognition. Contrary to the assumption of a uniform test label distribution, we adopt the test agnostic long-tail recognition in (Zhang et al., 2022). This approach evaluates model performance across **multiple** test sets, each subjects to a different label distribution. From a distributional standpoint, we can frame this as a label shift problem, where i) $\mathbb{P}_{tr}[\boldsymbol{x}|y] = \mathbb{P}_{te}[\boldsymbol{x}|y]$ and ii) $\mathbb{P}_{tr}[y] \neq \mathbb{P}_{te}[y]$ generally. To model the uncertainty of the test label distributions, we assume that the labels are sampled in the following hierarchical process:

1) Due to label shift problem, the test label distribution $\mathbb{P}_{te}[y]$ should be more than just a constant function. In this paper, we assume that $\mathbb{P}_{te}[y]$ is rather sampled from a **meta-distribution** $\mathcal{E}$ over the simplex Δ^C, *i.e.*, $\mathbb{P}_{te}[y] \sim \mathcal{E}$. Moreover, the selected meta-distribution should: i) reflect the diversity of the label distributions with significant different degrees of imbalance (global variations); ii) reflect the local variations of the label distributions to make more stable predictions (local variations).
2) The test data is then sampled from the observed test distribution: $(\boldsymbol{x}, y) \sim \mathbb{P}_{te}[\boldsymbol{x}, y] = \mathbb{P}_{tr}[\boldsymbol{x}|y] \cdot \mathbb{P}_{te}[y]$.

Goal. We aim to develop a well-trained model that demonstrates good performance across the entire spectrum of test label distributions within $\mathcal{E}$. Moreover, if we denote the performance of a model f on test data $\mathcal{D}_{te}$ by a loss function ℓ, the goal can be expressed as:

$$\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathbb{P}_{te}[\boldsymbol{x},y]} \ell\left(y, f(\boldsymbol{x})\right) \text{ is small for most } \mathbb{P}_{te}[y] \sim \mathcal{E}.$$

4. Methodology

4.1. Preliminaries

To initiate our discussion, we revisit the approach to addressing the label shift problem in scenarios where the test label distribution $\mathbb{P}_{te}[y]$ is fixed but differs from the training distribution.

Following the convention of the classification problem, we want to learn a scoring function $\{f_\theta^{(1)}, \cdots, f_\theta^{(C)}\} : \mathbb{R}^d \to \Delta^C$ by minimizing the CE loss over the training data:

$$\ell_{CE}(f_\theta(\boldsymbol{x}), y) = -\log\left(\mathsf{softmax}\left(f_\theta^{(y)}(\boldsymbol{x})\right)\right).$$

With the adjusted logits (Menon et al., 2021), if the softmax function is calibrated, we have $\mathsf{softmax}(f_\theta^{(y)}(\boldsymbol{x})) \propto \mathbb{P}_{tr}[y|\boldsymbol{x}]$. We can then achieve the Fisher consistency in the sense that:

$$\hat{y} = \mathop{\mathrm{argmax}}_{i\in\{1,\cdots,C\}} f_\theta^{(i)}(\boldsymbol{x}) \approx \mathop{\mathrm{argmax}}_{i\in\{1,\cdots,C\}} \mathbb{P}_{tr}[y = i|\boldsymbol{x}]$$

However, due to the label shift problem, the prediction rule suffers a distributional bias since $\mathbb{P}_{tr}[y|\boldsymbol{x}] \neq \mathbb{P}_{te}[y|\boldsymbol{x}]$. To address this issue, (Hong et al., 2021) proposes to employ the logit adjustment method (Menon et al., 2021) to modify the loss function. Specifically, one can use **the following loss function denoted as** $\ell_{\mathsf{LA}}(f_\theta(\boldsymbol{x}), y; \mathbb{P}_{te})$:

$$\ell_{CE}\left(f_\theta^{(y)}(\boldsymbol{x}) - \log\left(\frac{\mathbb{P}_{te}[y]}{\mathbb{P}_{tr}[y]}\right)\right). \tag{1}$$

To see how the new logarithm factor eliminates the bias, we can again check the probability correspondence when calibration of the softmax score is achieved:

$$\mathsf{softmax}(f_\theta^{(y)}(\boldsymbol{x})) \propto \mathbb{P}_{tr}[y|\boldsymbol{x}] \cdot \frac{\mathbb{P}_{te}[y]}{\mathbb{P}_{tr}[y]} \propto \mathbb{P}_{te}[y|\boldsymbol{x}]$$

This implies that when the model is effectively trained, it is possible to re-establish the Bayes rule.

4.2. DirMixE: Learning with a Dirichlet Mixture of the Experts

We are ready to formulate our method, called DirMixE, for unknown label distributions $\mathbb{P}_{te}[y]$. Seeing the complicated random variation, *minimizing the mean error is not enough* *for a stable prediction*. In this sense, we aim to optimize the model performance from a distributional perspective. In other words, we want to ensure the average error is low for most possible choices of $\mathbb{P}_{te}[y]$. To do this, we control the average error's first (mean) and second (variance) moment to be small. Mathematically, we try to minimize the following objective function:

$$\underset{\mathbb{P}_{te}\sim\mathcal{E}}{\mathbb{E}}\left[\mathcal{R}(f;\mathbb{P}_{te})\right] + \lambda \cdot \underset{\mathbb{P}_{te}\sim\mathcal{E}}{\mathbb{V}}\left[\mathcal{R}(f;\mathbb{P}_{te})\right] \tag{2}$$

where $\mathcal{R}(f;\mathbb{P}_{te}) = \mathbb{E}_{(\boldsymbol{x},y)\in\mathbb{P}_{tr}}\, \ell_{\mathsf{LA}}(f_\theta(\boldsymbol{x}), y; \mathbb{P}_{te})$ is the expected logit adjustment loss (see Eq.1) given the label distribution $\mathbb{P}_{te}$, and λ is a trade-off coefficient.

It is hard to simultaneously minimize all the training losses over different distributions by learning a single model. To address this issue, we propose an ensemble learning strategy to deal with the heterogeneity among test distributions with different experts. We will elucidate the technical aspects of our approach by addressing three key questions: 1) How to model the distribution $\mathcal{E}$ to capture the test distribution of interest effectively, 2) How to assign different experts to distinct test distributions, and 3) How to approximate the objective function using sampling methods and empirical data.

Targeted $\mathcal{E}$. The target $\mathcal{E}$ is designed to capture the global and local variations of the label distribution, where we resort to the hierarchical structure of **mixture distributions**. Specifically, we define K as the number of components in this mixture distribution, using a discrete distribution to select a component i with probability p_i. Given the commonality of the Dirichlet distribution for sampling discrete probabilities (label distributions are discrete probabilities), we define the i-th component itself as a Dirichlet distribution with parameters $\boldsymbol{\alpha}^{(i)} = [\alpha_1^{(i)}, \cdots, \alpha_C^{(i)}]$. This mixture distribution framework facilitates the hierarchical sampling of $\mathbb{P}_{te}$, offering a structured approach to model the variation of test label distributions:

$$\mathbb{P}_{te}|i \;\sim\; \mathsf{Dir}\left(\alpha_1^{(i)}, \cdots, \alpha_C^{(i)}\right) \tag{3}$$

$$i|\boldsymbol{p} \;\sim\; \mathsf{Discrete}\left(p_1, \cdots, p_K\right) \tag{4}$$

In the meta-distributions, the diversity across different distributions captures the global variations. Moreover, within each component, the Dirichlet distribution makes sure that the local variations concentrated on the mean can be effectively characterized. In this sense, we can better utilize the hierarchy of the variations by virtue of the meta-distribution.

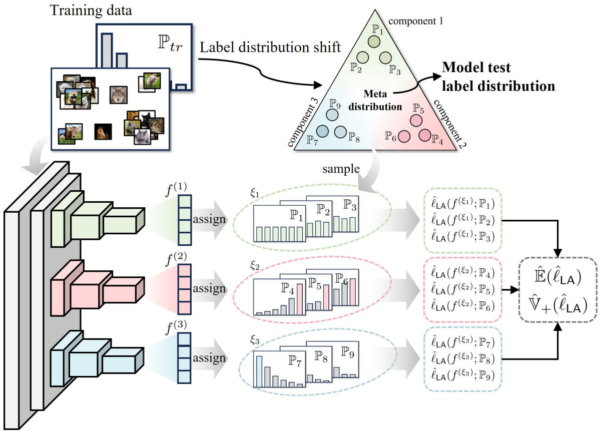

Training data
$\mathbb{P}_{tr}$
Label distribution shift
component 1
$\mathbb{P}_1$
$\mathbb{P}_2$
$\mathbb{P}_3$
Model test label distribution
Meta distribution
component 3
component 2
$\mathbb{P}_9$
$\mathbb{P}_5$
$\mathbb{P}_7$
$\mathbb{P}_8$
$\mathbb{P}_6$
$\mathbb{P}_4$
sample
$f^{(1)}$
$f^{(2)}$
$f^{(3)}$
assign
assign
assign
ξ_1
ξ_2
ξ_3
$\hat{\ell}_{\mathsf{LA}}(f^{(\xi_1)};\mathbb{P}_1)$
$\hat{\ell}_{\mathsf{LA}}(f^{(\xi_1)};\mathbb{P}_2)$
$\hat{\ell}_{\mathsf{LA}}(f^{(\xi_1)};\mathbb{P}_3)$
$\hat{\ell}_{\mathsf{LA}}(f^{(\xi_2)};\mathbb{P}_4)$
$\hat{\ell}_{\mathsf{LA}}(f^{(\xi_2)};\mathbb{P}_5)$
$\hat{\ell}_{\mathsf{LA}}(f^{(\xi_2)};\mathbb{P}_6)$
$\hat{\ell}_{\mathsf{LA}}(f^{(\xi_3)};\mathbb{P}_7)$
$\hat{\ell}_{\mathsf{LA}}(f^{(\xi_3)};\mathbb{P}_8)$
$\hat{\ell}_{\mathsf{LA}}(f^{(\xi_3)};\mathbb{P}_9)$
$\hat{\mathbb{E}}(\hat{\ell}_{\mathsf{LA}})$
$\hat{\mathbb{V}}_+(\hat{\ell}_{\mathsf{LA}})$

Figure 2. Illustration of the training process, where a hierarchical sampling process is employed to esitmate the empirical risk.

Note that prior research (Aimar et al., 2023; Hong et al., 2021; Zhang et al., 2022) can be regarded as a special case of the proposed approach, wherein the Dirichlet distribution is substituted with a fixed point (Dirac delta distribution). However, as we will demonstrate in the subsequent section, this approach limits the generalization capacity of the resulting model. This limitation arises because such fixed distribution schemes inadequately represent the randomness inherent in the test label distributions. **Mixture of Experts Strategy.** The overall model is expressed in a compositional structure

$$f_\theta^{(i)}(\cdot) = g^{(i)} \circ \psi(\cdot),\; i = 1, 2, \cdots, K,$$

where $f^{(i)}$ is the model allocated for the i-th component, $g^{(i)}$ reflects the distributional-specific feature, ψ reflects the invariant part and is shared across components.

In the upcoming discussions, we will see how $\mathcal{E}$ and the MoE strategy will help us construct an estimation of the loss distribution.

Empirical Approximation of the Objective Function. To approximate the population-level objective function (Eq.2), we utilize a Monte Carlo method to estimate the expectation and variance over $\mathcal{E}$. This is achieved by generating empirical data for test label distributions. Each time, we sample a test label distribution $\mathbb{P}$ and a component ξ from the hierarchical model as outlined in equations (3)-(4). The generated data set is represented as $\mathcal{P} = \{\mathbb{P}_j, \xi_j\}_{j=1}^M$. In this process, each expert $f_\theta^{(\xi_j)}$ is assigned to the sampled pair $(\mathbb{P}_j, \xi_j)$. The performance of these experts is then evaluated based on the following average loss:

$$\hat{\ell}_{\mathsf{LA}_j} = \frac{1}{N}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \ell_{\mathsf{LA}}\left(f_\theta^{(\xi_j)}(\boldsymbol{x}), y; \mathbb{P}_j\right).$$

From all of the above, we get the Monte Carlo estimation of the population mean and variance in (2) as:

$$\hat{\mathbb{E}}(\ell_{\mathsf{LA}}) = \frac{1}{M}\sum_{j=1}^{M}\hat{\ell}_{\mathsf{LA}_j},\; \hat{\mathbb{V}}(\ell_{\mathsf{LA}}) = \frac{1}{M}\sum_{j=1}^{M}\left(\hat{\ell}_{\mathsf{LA}_j} - \hat{\mathbb{E}}(\ell_{\mathsf{LA}})\right)^2.$$

Practically, minimizing the variance term tends to punish $\hat{\ell}_{\mathsf{LA}_j}$ being smaller than its empirical mean. This over-regularization can impede the training process, particularly for test distributions that are relatively easier to learn. To circumvent this challenge, we replace the variance penalty with a semi-variance term, which only penalizes $\hat{\ell}_{\mathsf{LA}_j}$ that are larger than the mean.

$$\hat{\mathbb{V}}_+(\ell_{\mathsf{LA}}) = \frac{1}{M}\sum_{j=1}^{M}\left(\left(\hat{\ell}_{\mathsf{LA}_j} - \hat{\mathbb{E}}(\ell_{\mathsf{LA}})\right)_+\right)^2.$$

Please see Appendix A for the sampling process in details.

Putting all together, we come to the final objective function:

$$\min_{g^{(1)},\cdots,g^{(K)},\ \psi} \hat{\mathbb{E}}(\ell_{\mathsf{LA}}) + \lambda\cdot\hat{\mathbb{V}}_+(\ell_{\mathsf{LA}}).$$

During testing, we adopt the self-supervision strategy as proposed in SADE (Zhang et al., 2022) to learn a set of model averaging weights $\omega_{te}^{(1)},\ \omega_{te}^{(2)},\ \cdots,\ \omega_{te}^{(K)}$. The prediction for a test data point is then determined by combining these weights, leading to a weighted average of the outputs from the individual models expressed as follows:

$$f_{te}(\boldsymbol{x}) = \sum_{i=1}^{K}\omega_{te}^{(i)}\cdot f^{(i)}(\boldsymbol{x}). \tag{5}$$

5. Theoretical Analysis

For the sake of simplicity, we will adopt **asymptotic notations** to perform magnitude comparison. Specifically, $f(n) \lesssim g(n)$ denotes $\exists C > 0, f(n) \leq C\cdot g(n)$, and $f(n) \gtrsim g(n)$ *vice versa*. $f(n) \asymp g(n)$ means $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$ hold simultaneously. All the proofs are deferred to Appendix B and D.

5.1. Generalization Ability of DirMixE

To show the generalization ability of our proposed method, we attempt to answer two questions theoretically. The first question is how well the proposed training method can generalize to unseen label distributions. In the following Thm., we answer this question by presenting an upper bound of $\mathbb{E}'(\ell_{\mathsf{LA}})$, the expected loss over all test label distributions sampled from $\mathcal{E}'$ (which is not necessarily identical to $\mathcal{E}$) and $\mathcal{D}$. We assume that $\mathbb{V} \asymp \mathbb{V}_+$, which will be justified in Sec.5.2.

Theorem 1 (Generalization Bound, Informal). *Let $\mathcal{E}$ be the true meta-distribution and $\mathcal{P}$ be an observed empirical distribution sampled by the Monte Carlo method. Moreover, the training data $\mathcal{S}$ are sampled in an i.i.d manner from its true distribution $\mathcal{D}$. Assume that $\mathbb{V}(\ell_{\mathsf{LA}}) \asymp \mathbb{V}_+(\ell_{\mathsf{LA}})$, then for any possible meta-distribution $\mathcal{E}'$ defined on the probability simplex $\mathbb{S}^{c-1}$ we denote $\mathbb{E}'(\ell_{\mathsf{LA}})$ as the corresponding version of the loss expectation, then the following inequality holds uniformly for all $f \in \mathcal{F}$ with probability at least $1-\delta$ over the randomness of the training data $\mathcal{S}$, and the sampled label distribution $\mathcal{P}$:*

$$\begin{aligned}\mathbb{E}^{'}(\ell_{\mathsf{LA}}) \lesssim{}& \underbrace{\hat{\mathbb{E}}(\ell_{\mathsf{LA}}) + \sqrt{\frac{\mathbb{C}_1}{M}\cdot \underset{+}{\hat{\mathbb{V}}}(\ell_{\mathsf{LA}})}}_{(1)} + \underbrace{\frac{\|\mathcal{E}-\mathcal{E}'\|_\infty}{C!}}_{(2)} \\ & + \underbrace{\mathbb{C}_2\cdot\left(M^{-3/4} + M^{-1/2}N^{-1/4} + N^{-1/2}\right)}_{(3)},\end{aligned}$$

where $\mathbb{C}_1, \mathbb{C}_2$ are complexity terms of the hypothesis space $\mathcal{F}$.

The bound can be decomposed into three terms:

Empirical Error Term: The first component (1) in the theorem represents an empirical error term that can be optimized during training. The implication of this bound is that employing a semi-variance regularization approach enables achieving a **sharper bound** compared to the traditional complexity of $O(N^{-1/2} + M^{-1/2})$. Our proposed method accomplishes this by optimizing $\hat{\mathbb{E}}(\ell_{\mathsf{LA}}) + \lambda\cdot\hat{\mathbb{V}}(\ell_{\mathsf{LA}})$, which serves as a smooth approximation of the empirical error term (1).

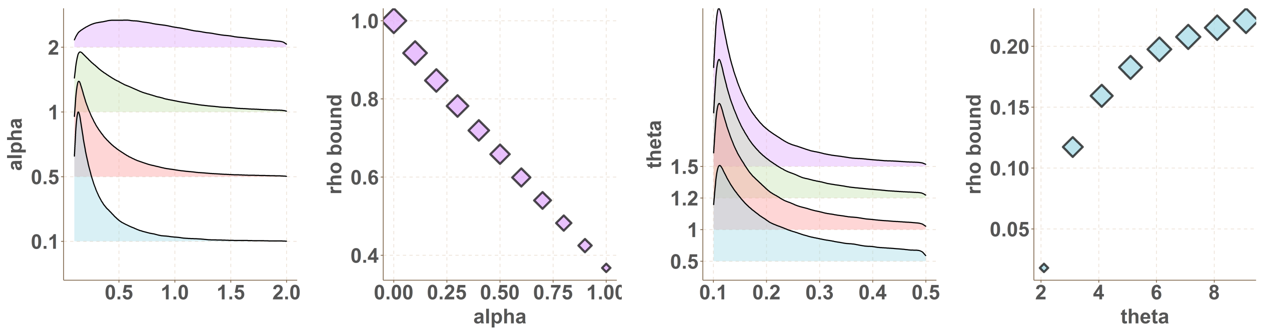
Approximation Error from Meta-Distribution Shift: The second term (2) provides an upper bound on the stochastic error that may arise due to a potential shift from the meta-distribution $\mathcal{E}$ to another distribution $\mathcal{E}'$. Following the insights from (Nguyen et al., 2020; 2023), employing sufficiently large components in the Dirichlet mixture can reduce the difference $||\mathcal{E}-\mathcal{E}'||_\infty$ to nearly zero.

Hierarchical Stochastic Error: The third term (3) quantifies the hierarchical stochastic error involved in using the sample sets $(\mathcal{S}, \mathcal{P})$ instead of the true distributions $(\mathcal{D}, \mathcal{E})$. As highlighted, these results enjoy a sharper residual error compared to conventional ones.

Furthermore, the theorem suggests that in cases where the test label distribution is not fixed, it is essential to sample a sufficiently large number of test label distributions (a large M) to mitigate the stochastic error components $M^{-3/4} + M^{-1/2}N^{-1/4} + N^{-1/2}$. This also underlines the advantage of DirMixE over existing Mixture of Experts (MoE) schemes, such as those referenced in (Aimar et al., 2023; Hong et al., 2021; Zhang et al., 2022), where M is typically fixed as $O(1)$ by using a fixed distribution assignment.

The second question is how well we can merge the experts during test time. To explore the answer, we check the performance of the self-supervised model averaging result f_{te} in (5).

Theorem 2 (Model Averaging Error). *Under the same*



(a) Density of Gamma w.r.t α (b) ρ Bound for the Gamma (c) Density of Pareto w.r.t θ (d) ρ Bound for the Pareto

Figure 3. **The Upper Bound for ρ for different distributions.** (a) shows the distribution density of the Gamma distribution with β fixed as 0.5 and α varied. The corresponding distribution has a tailed shape when $\alpha < 1$. (b) shows the upper bound ρ for $\alpha < 1$. One can observe that ρ is roughly greater than 0.4 in this range. (c) shows the distribution density of the Pareto distribution with ℓ_m fixed as 0.1 and θ varied. (d) shows the ρ value for $2 < \theta \leq 10$, we can see that ρ is slightly greater 0.2 for large θ. In all these observed cases, the assumption $\mathbb{V}_+ \asymp \mathbb{V}$ is admissible.

setting as Thm.1, the following results hold:

$$\mathbb{E}(\ell_{\mathsf{LA}}(f_{te})) \lesssim \sum_{i\in[K]} \mathbb{E}(\omega_i)\cdot\mathbb{E}(\ell_{\mathsf{LA}}(f^{(i)})) + \sum_{i\in[K]} \mathbb{COV}(\omega_i, \ell_{\mathsf{LA}}(f^{(i)})),$$

where $\mathbb{COV}$ is the covariance operator and all the expectations are taken over the joint distribution of $\mathcal{E}\otimes\mathcal{D}$.

The proof of Thm.2 directly follows (Zhang et al., 2023a) and thus is omitted.

The resulting upper bounds consist of two parts. The first part could be regarded as a weighted loss, which can be guaranteed to have a similar concentration as in Thm.2. The second term represents the covariance between the model weights and the corresponding losses. A well-designed model aggregation method should ideally assign lower weights to models that yield higher losses. In this study, we utilize the self-supervised method described in (Zhang et al., 2022), which effectively maximizes the mutual information between predictions and ground truth under certain conditions. Consequently, we anticipate a negative correlation between model weight and loss (negative covariance), significantly reducing model averaging error.

5.2. Upper Bounding Variance with Semi-Variance

In DirMixE, we replace the variance in empirical estimation with the semi-variance of loss to enhance optimization. As shown in Thm.1, the stochastic error associated with this approach remains small if $\mathbb{V}(\ell_{\mathsf{LA}}) \asymp \mathbb{V}_+(\ell_{\mathsf{LA}})$. Given that semi-variance is inherently smaller than variance, our analysis primarily focuses on validating the condition $\mathbb{V}(\ell_{\mathsf{LA}}) \lesssim \mathbb{V}_+(\ell_{\mathsf{LA}})$. Note that, for a well-trained model, there should be a negative correlation between the probability density and the magnitude of loss. Based on this assumption, we show that $\mathbb{V}(\ell_{\mathsf{LA}}) \lesssim \mathbb{V}_+(\ell_{\mathsf{LA}})$ for certain types of distributions satisfying the negative correlation, ranging from light-tail (such as the exponential distribution) to heavy-tail ones (such as the Pareto distribution) distributions. For clarity and simplicity, we will henceforth use ℓ **as shorthand for** ℓ_{LA} in this subsection.

Theorem 3. *Let $\rho = \frac{\mathbb{V}_+[\ell]}{\mathbb{V}[\ell]}$, then we have the following results for different distributions:*

a) *If ℓ subjects to an **exponential** distribution, i.e the p.d.f $p(\ell) \propto \exp(-c\ell)$, then we have: $\rho \geq \exp(-1)$.*

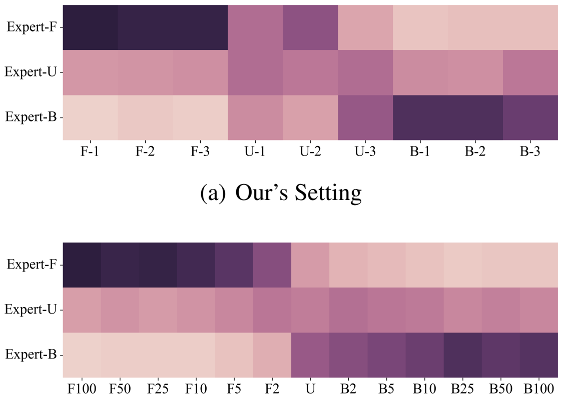
b) *If ℓ subjects a **Gamma** distribution with parameters α, β, i.e, the p.d.f $p(\ell) \propto \ell^{\alpha-1}\cdot\exp(-\beta\cdot\ell)$, then we have: $\rho \geq 1-\alpha\cdot\frac{\Gamma^{\uparrow}(\alpha,\alpha)}{\Gamma(\alpha)}$, where $\Gamma^{\uparrow}$ is the lower incomplete gamma : $\Gamma^{\uparrow}(s,x) = \frac{1}{\Gamma(\alpha)}\cdot\int_0^x t^{s-1}\cdot\exp(-t)dt$.*

c) *If ℓ subjects to a **Pareto** distribution with parameter θ, i.e the p.d.f $p(\ell) \propto (\ell/\ell_m)^{-\theta}$ for some $\ell_m > 0$, then we have:*

$$\rho = (\theta-1)^2\cdot\left[1-\phi(\theta)^{2-\theta}\right] + \theta\cdot(\theta-2)\cdot\left[2\cdot\phi(\theta)^{1-\theta}-\phi(\theta)^{-\theta}-1\right]$$

where $\phi(\theta) = \theta/(\theta-1)$.

According to the theorem, we present practical observation on the Gamma and Pareto distribution in Fig.3-(a). The gamma distributions are tail-shaped when $\alpha < 1$. In this sense, we plot the theoretical lower bound ρ for $\alpha < 1$ in Fig.3-(b). The results show that $\mathbb{V}_+ / \mathbb{V} > 0.4$ in most



(a) Our's Setting

(b) SADE's Setting

Figure 4. **Weight Assignment in the Self-supervised Aggregations on CIFAR-100. F,U,B** represent the forward, uniform and backward distributions, see Appendix F for more.

cases. Similar Fig.3-(c) shows that the Pareto distributions are generally tail-shaped. We then plot ρ for $2 < \theta < 10$ in Fig.3-(d). The results show that the $\mathbb{V}_+ / \mathbb{V}$ roughly resides in $[0.1, 0.3]$ when $\theta \geq 3$. Moreover, for the exponential distribution, we have a universal upper bound for ρ as $\exp(-1)$, which is greater than 0.2. In this sense, the proposed claim $\mathbb{V}_+ \asymp \mathbb{V}$ holds for a wide span of tail-shaped distributions.

In practical terms, $\mathbb{P}_{te}$ is sampled from a mixture distribution. Consequently, it is probable that the loss function is subject to a mixture distribution rather than a single distribution. The following theorem shows that ρ for a mixture distribution can also be lower bounded by the mean of the respective ρ_i values of its component distributions.

Theorem 4. *Let ℓ be sampled from a* ***mixture*** *distribution such that $p(\ell) = \sum_{k=1}^{K} \omega_k \cdot p_k(\ell)$, where $p_k(\ell)$ is the p.d.f. for the component k. Furthermore, for each component, we denote: $\mathbb{E}_k[\ell] = \mu_k, \mathbb{V}_k[\ell] = \sigma_k^2$. Under the assumption that 1) $\mathbb{P}\left[\ell \neq \mathbb{E}[\ell]\right]$ with probability one w.r.t the mixture distribution, 2) $\mathbb{V}[\ell]/\sigma_i^2 \geq \max\left\{\frac{\mathbb{V}_-[\ell]_{i,j}}{\mathbb{V}_-[\ell]_i}, 1\right\}$, then ρ can be upper bounded by the average of ρ_i in the sense that:*

$$\rho \leq \sum_i \omega_i \cdot \rho_i,$$

where $\mathbb{V}_-[\ell] = \mathbb{V}[\ell] - \mathbb{V}_+[\ell]$; $\rho_i, \mathbb{V}_-[\ell]_i$ are the corresponding versions of $\rho, \mathbb{V}_-[\ell]$ on the i-th component, and

$$\mathbb{V}_-[\ell]_{i,j} = \int_0^{\infty} \left((\ell - \mu_i)_-\right)^2 \cdot p_j(\ell) \cdot d\ell.$$

This implies that controlling ρ is feasible if we can manage each ρ_i. Ultimately, this leads to the conclusion that $\mathbb{V}(\ell_{\mathsf{LA}}) \asymp \mathbb{V}_+(\ell_{\mathsf{LA}})$ holds under mild conditions.

6. Experiments

In this section, we conduct a series of empirical studies to demonstrate the effectiveness of our proposed algorithm. **Due to space limitations, please refer to Appendix E and F for more details and experiments.**

6.1. The Choice of meta-distribution and Experts

We briefly introduce the choice of (3) and (4) to define the mixture distribution. Detailed implementations are shown in the appendix. Drawing inspiration from the skill-diverse expert learning approach in prior art, we employ a three-component mixture model for (3) to encapsulate three critical skills: forward component aligns with the training label distribution, indicative of performance in the head classes; uniform component corresponds to the local variation around uniform distribution; backward component represents the local variation around an inverse long-tail distribution of the training set, signifying performance in the tail distribution. Please refer to Appendix E.4 for more details.

6.2. Experiment Protocols

Evaluation Protocols. We evaluate the performance of various methods across multiple test datasets. Specifically, we employ **two regimes** to generate test datasets: **(a) Ours Setting.** We generate test data by sampling from the perturbed version of the meta-distribution in Sec.E. Subsequently, for each Dirichlet distribution, we sample **three label distributions** for testing on three forward/uniform/backward LT distributions, respectively (As in Tab.2). **(b) SADE's Setting**: (Zhang et al., 2022). Following SADE, the test datasets usually fall into one of three distribution types: (forward) *long-tail, uniform, and backward long-tail*, each defined by a different imbalance degree ρ. **Please see Appendix E for more details.**

6.3. Overall Performance

Tab.1-3 compare the overall performance on CIFAR-10, CIFAR-100, ImageNet for Our's setting, while those for sade's setting are shown in Appendix F. Moreover, we also include the results for `iNaturalist` in Appdendix.F. For fairness, we only compare the performance using MoE scheme and do not use mixup for all the competitors. Moreover, we adopt the self-supervised aggregation method of SADE to align test-time operations for all the MoE-based Models (SADE, BalPoE, DirMixE) except RIDE. The rationale for the exception is that all the experts in RIDE are designed for the same distribution, and there is already a routing strategy to choose the experts. We have the following observations on the results:

Table 1. CIFAR-10-LT (**Ours Setting**)

Method	Forward-LT			Uniform			Backward-LT			Mean
	1	2	3	1	2	3	1	2	3	
LDAM	89.66	90.50	90.30	74.27	74.39	74.77	60.65	59.89	60.42	74.98$_{(\pm 12.19)}$
LA	89.74	88.98	90.51	79.02	79.07	79.27	72.90	71.67	72.51	80.41$_{(\pm\ 7.17)}$
VS	85.12	84.71	85.24	80.69	80.50	80.85	82.10	81.21	80.95	82.37$_{(\pm\ 1.92)}$
LADE	87.31	87.11	87.79	79.69	79.72	80.20	76.79	75.77	76.47	81.21$_{(\pm\ 4.62)}$
DDC	87.07	87.07	87.24	81.62	81.44	82.03	80.15	79.04	79.47	82.79$_{(\pm\ 3.20)}$
RIDE	86.47	85.63	87.33	81.92	81.89	81.98	81.34	81.25	81.19	83.22$_{(\pm\ 2.35)}$
SADE	90.26	89.94	91.05	83.14	82.71	83.38	88.89	88.29	**89.48**	87.46$_{(\pm\ 3.19)}$
BalPoE	**91.30**	**91.54**	**92.72**	81.58	81.78	81.89	78.97	77.28	77.87	83.88$_{(\pm\ 5.86)}$
DirMixE	90.46	89.90	91.30	**83.24**	**82.98**	**83.71**	**89.39**	**88.78**	88.40	**87.57**$_{(\pm\ 3.12)}$

Table 2. Performance Comparison on CIFAR-100-LT (**Ours Setting**)

Method	Forward-LT			Uniform			Backward-LT			Mean
	1	2	3	1	2	3	1	2	3	
LDAM	64.17	63.59	66.43	38.16	37.35	38.07	17.32	15.04	16.95	39.68$_{(\pm 19.78)}$
LA	58.71	58.77	61.39	44.93	45.35	43.75	31.94	29.49	30.96	45.03$_{(\pm 11.81)}$
VS	58.20	58.30	59.99	39.32	41.24	42.13	27.37	24.53	28.14	42.14$_{(\pm 13.21)}$
LADE	57.42	57.76	61.06	43.35	42.92	43.05	32.40	30.34	33.47	44.64$_{(\pm 11.01)}$
DDC	59.26	56.89	60.90	45.05	44.86	45.59	32.59	32.05	34.41	45.73$_{(\pm 10.68)}$
RIDE	63.34	62.72	65.68	45.30	45.35	47.81	30.82	27.61	30.75	46.60$_{(\pm 14.02)}$
SADE	66.13	66.00	68.31	46.89	48.05	45.59	41.99	42.91	40.06	51.77$_{(\pm 10.90)}$
BalPoE	**67.75**	**67.80**	**69.98**	45.05	46.86	**48.81**	29.80	26.32	31.17	48.17$_{(\pm 16.19)}$
DirMixE	66.85	66.40	69.44	**47.99**	**49.41**	44.21	**44.41**	**47.01**	**44.35**	**53.34**$_{(\pm 10.22)}$

a) **Overall empirical trends**: 1) Our method shows similar performance to state-of-the-art (SOTA) methods like BalPoE and SADE for **Forward-LT** and **Uniform**. 2) However, in **Backward-LT**, our method's improvements are much more substantial. For instance, in Tab.2 (CIFAR-100, our setting), the performance gain varies from 2.4 to 4.1. In Tab.3 (ImageNet-LT, our setting), it ranges from 2.0 to 2.4. Thanks to these gains in `Backward-LT`, our method consistently achieves the best average performances, demonstrating its effectiveness.

b) **Performance differences across datasets**: CIFAR series performances (CIFAR-10, CIFAR-100) are comparable, likely due to their similar data distributions and scales. In contrast, the ImageNet dataset, with its distinct data distribution and scale, shows slightly different trends. Yet, the results still follow the pattern noted in a). Even though our method slightly lags behind SOTA methods on the ImageNet dataset for `Uniform`, the performance differences are mostly less than 0.5. Conversely, the improvement in `Backward` distributions is more significant, ensuring our method maintains the best average performances.

c) **Explaining the differences**: The distinct distribution of the ImageNet dataset compared to CIFAR-10 and CIFAR-100 may account for these discrepancies. The greater variation between different label distributions means the model must focus more on `Backward` distributions, slightly compromising performance in `Uniform` and `Forward` distributions to lead in overall performance.

please see Appdendix F.6 for comparisions with other baselines.

6.4. Experts Assignment

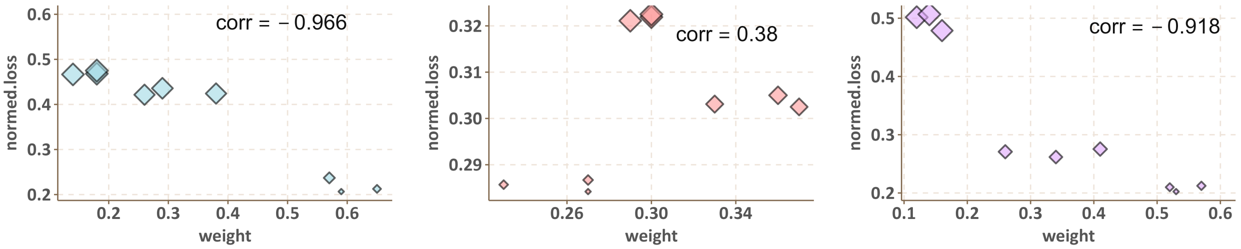
In this part, we validate the ability of the test-time self-supervised aggregation by visualizing the weight assignments for different label distributions in Fig.4. The forward and backward experts always tend to have a significant weight for their corresponding distributions. Uniform distributions tend to utilize all three experts. This is because tail and head classes are equally crucial for uniform distribution.

6.5. Correlation between Weights and Losses

Fig.5 shows the correlation between expert weights of DirMixE during the test phase and their corresponding loss. We normalize the losses to align the magnitude of the loss on different distributions, where the normalized loss of the expert i is $\ell_i / \sum_{i=1}^{3} \ell_i$. The results show a strong negative correlation on the forward and backward experts, and a

Table 3. ImageNet-LT (**Ours Setting**)

Method	Forward-LT			Uniform			Backward-LT			Mean
	1	2	3	1	2	3	1	2	3	
LDAM	61.74	62.22	61.49	47.51	47.37	48.63	32.67	32.54	32.00	47.35$_{(\pm 12.02)}$
LA	60.94	60.32	59.78	50.82	50.86	50.86	40.39	40.16	40.14	50.47$_{(\pm\ 8.22)}$
VS	61.14	59.60	59.00	52.04	52.22	53.18	44.03	44.47	43.02	52.08$_{(\pm\ 6.60)}$
LADE	63.58	62.29	61.92	53.48	52.38	53.31	41.53	42.31	41.25	52.45$_{(\pm\ 8.56)}$
DDC	59.42	59.45	58.30	51.36	51.83	51.51	42.50	44.09	43.47	51.33$_{(\pm\ 6.43)}$
RIDE	65.32	64.28	63.49	55.18	55.02	55.98	43.49	44.52	43.16	54.49$_{(\pm\ 8.47)}$
SADE	69.64	69.77	**70.35**	**58.51**	**58.96**	**58.69**	53.54	53.13	53.82	60.71$_{(\pm\ 6.86)}$
BalPoE	69.82	69.32	70.26	58.26	58.78	58.47	52.08	51.46	52.36	60.09$_{(\pm\ 7.37)}$
DirMixE	**70.13**	**70.88**	70.29	58.38	58.85	58.02	**55.59**	**55.09**	**56.25**	**61.50**$_{(\pm\ 6.43)}$



(a) Expert 1 (Forward)

(b) Expert 2 (Uniform)

(c) Expert 3 (Backward)

Figure 5. **The Correlation between Expert Weights and Loss.**

much weaker positive correlation for uniform ones. This is because uniform distributions do not have a significant bias on head/tail classes, producing a relatively stable average performance across different distributions. Above all, in most cases, we can observe negative correlations between loss and expert weight. According to Thm.2, the negative correlation tends to reduce the generalization error of the test-time aggregation scheme, validating the reasonability of observed performance advantage.

6.6. The Semi-Variance/Variance Ratio

Recall Thm.1, we adopt the assumption that $\mathbb{V}(\ell_{\mathsf{LA}}) \asymp V_+(\ell_{\mathsf{LA}})$. To validate this assumption, we calculate the semi-variance/variance ratio (ρ) in CIFAR-10 and CIFAR-100. We find that $\rho = 0.503, 0.509$, respectively for CIFAR-10 and 100, which obviously aligns with the assumption.

7. Conclusion

We consider the hierarchy of the global and local variations of the test label distributions in test agnostic long-tail recognition. To this end, we propose a Dirichlet MoE method named DirMixE. The label distributions are sampled from a meta-distribution, characterized by a mixture of Dirichlet distribution. In the proposed MoE strategy, each expert is assigned to a local Dirichlet distribution for a specific skill. The global and local variations are then captured by inter- and intra-component variations of the meta-distribution. This also leverages a stable objective function minimizing the mean and semi-variance of the loss with the help of Monte Carlo method. When $\mathbb{V}_+(\ell_{\mathsf{LA}}) \asymp \mathbb{V}(\ell_{\mathsf{LA}})$, we show that the proposed objective function enjoys an sharper bound by semi-variance regularization. Finally, extensive experiments demonstrate the efficacy of DirMixE.

Impact Statement

This work aims at general issues for the long-tail classification problems. If applied to fairness-sensitive applications, it might be helpful to improve the fairness for the minority classes (say specific group of people, minority species, *etc*).

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, in part by National Natural Science Foundation of China: 62236008, U21B2038, U23B2051, U2001202, 61931008, 62122075, 61976202, 62206264, and 92370102, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDB0680000, in part by the Innovation Funding of ICT, CAS under Grant No.E000000.

References

Aimar, E. S., Jonnarth, A., Felsberg, M., and Kuhlmann, M. Balanced product of calibrated experts for long-tailed recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19967–19977, 2023.

Alshammari, S., Wang, Y., Ramanan, D., and Kong, S. Long-tailed recognition via weight balancing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition,*, pp. 6887–6897, 2022.

Cai, J., Wang, Y., and Hwang, J. ACE: ally complementary experts for solving long-tailed recognition in one-shot. In *IEEE/CVF International Conference on Computer Vision*, pp. 112–121, 2021.

Cao, K., Wei, C., Gaidon, A., Aréchiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Annual Conference on Neural Information Processing Systems*, pp. 1565–1576, 2019.

Cui, J., Zhong, Z., Liu, S., Yu, B., and Jia, J. Parametric contrastive learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 695–704, 2021.

Cui, J., Zhong, Z., Tian, Z., Liu, S., Yu, B., and Jia, J. Generalized parametric contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2023. doi: 10.1109/TPAMI.2023.3278694.

Cui, Y., Jia, M., Lin, T., Song, Y., and Belongie, S. J. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Duchi, J. and Namkoong, H. Variance-based regularization with convex objectives. *ArXiv*, 2016.

Galdran, A., Carneiro, G., and Ballester, M. Á. G. Balanced-mixup for highly imbalanced medical image classification. In *Medical Image Computing and Computer Assisted Intervention*, pp. 323–333, 2021.

Guo, H. and Wang, S. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15089–15098, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6622–6632, 2021.

Horn, G. V., Aodha, O. M., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. J. The inaturalist species classification and detection dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 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 *International Conference on Learning Representations*, 2020.

Kini, G. R., Paraskevas, O., Oymak, S., and Thrampoulidis, C. Label-imbalanced and group-sensitive classification under overparameterization. In *Annual Conference on Neural Information Processing Systems*, pp. 18970–18983, 2021a.

Kini, G. R., Paraskevas, O., Oymak, S., and Thrampoulidis, C. Label-imbalanced and group-sensitive classification under overparameterization. In *NIPS*, volume 34, pp. 18970–18983, 2021b.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. pp. 1–60, 2009.

Li, Y., Wang, T., Kang, B., Tang, S., Wang, C., Li, J., and Feng, J. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10988–10997, 2020.

Li, Z., Tang, J., and Mei, T. Deep collaborative embedding for social image understanding. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2070–2083, 2018.

Lin, T., Goyal, P., Girshick, R. B., He, K., and Dollár, P. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pp. 2999–3007, 2017.

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

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample-variance penalization. In *Conference on Learning Theory*, 2009.

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.

Miao, Z., Liu, Z., Gaynor, K. M., Palmer, M. S., Yu, S. X., and Getz, W. M. Iterative human and automated identification of wildlife images. *Nat. Mach. Intell.*, 3(10): 885–895, 2021.

Morik, K., Brockhausen, P., and Joachims, T. Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring. In *International Conference on Machine Learning*, pp. 268–277, 1999.

Nguyen, T., Chamroukhi, F., Nguyen, H. D., and McLachlan, G. J. Approximation of probability density functions via location-scale finite mixtures in lebesgue spaces. *Communications in Statistics-Theory and Methods*, 52(14): 5048–5059, 2023.

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1):1750861, 2020.

Rangwani, H., Aithal, S. K., Mishra, M., and R., V. B. Escaping saddle points for effective generalization on class-imbalanced data. In *Annual Conference on Neural Information Processing Systems*, pp. 22791–22805, 2022.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., and Yan, J. Equalization loss for long-tailed object recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11659–11668, 2020.

Tang, J., Shu, X., Qi, G.-J., Li, Z., Wang, M., Yan, S., and Jain, R. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1662–1674, 2016.

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

Wang, Z., Xu, Q., Yang, Z., He, Y., Cao, X., and Huang, Q. A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. In *Annual Conference on Neural Information Processing Systems*, 2023.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *CVPR*, pp. 1492–1500, 2017.

Ye, H., Chen, H., Zhan, D., and Chao, W. Identifying and compensating for feature deviation in imbalanced deep learning. *CoRR*, abs/2001.01385, 2020.

Zhang, D., Zhang, H., Tang, J., Hua, X.-S., and Sun, Q. Causal intervention for weakly-supervised semantic segmentation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *NeurIPS*, volume 33, pp. 655–666. Curran Associates, Inc., 2020.

Zhang, Q., Wu, H., Zhang, C., Hu, Q., Fu, H., Zhou, J. T., and Peng, X. Provable dynamic fusion for low-quality multimodal data. *ICML*, 2023a.

Zhang, X., Fang, Z., Wen, Y., Li, Z., and Qiao, Y. Range loss for deep face recognition with long-tailed training data. In *IEEE International Conference on Computer Vision*, pp. 5419–5428, 2017.

Zhang, Y., Hooi, B., Hong, L., and Feng, J. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *Annual Conference on Neural Information Processing Systems*, pp. 34077–34090, 2022.

Zhang, Y., Kang, B., Hooi, B., Yan, S., and Feng, J. Deep long-tailed learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):10795–10816, 2023b.

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

Zhong, Z., Cui, J., Yang, Y., Wu, X., Qi, X., Zhang, X., and Jia, J. Understanding imbalanced semantic segmentation through neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19550–19560, 2023.

Zhou, B., Cui, Q., Wei, X., and Chen, Z. BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9716–9725, 2020.

Zhu, J., Wang, Z., Chen, J., Chen, Y. P., and Jiang, Y. Balanced contrastive learning for long-tailed visual recognition. In *CVPR*, pp. 6898–6907.

Contents

A. Detailed Explanation of the Sampling process

In our study, we aim to ensure good overall performance across the entire meta distribution $\mathcal{E}$ of label distributions. Since our access is limited to training data with a long-tailed distribution, we employ the Logit Adjustment (LA) loss to tailor the model for specific test-label distributions using only training data. Our objectives are to 1) construct a meta-distribution $\mathcal{E}$ for test-label distributions, 2) assess the loss for different test-label distributions in $\mathcal{E}$, and 3) maintain minimal change in the mean and variance of the loss over $\mathcal{E}$. For 1), we model $\mathcal{E}$ as a Dirichlet mixture distribution, representing global and local variations through the randomness across and within its components, respectively. The Monte Carlo method is vital for implementing 2) and 3), detailed further in below.

We have to deal with both the training data distribution and a meta-distribution for test-label distribution, leading to a hierarchical process for the calculation:

Layer 1: For a given test-label distribution $\mathbb{P}_{te}$, we measure a model's adaptability to this new distribution through the expected LA loss over training data parametrized by $\mathbb{P}_{te}$:

$$\hat{\ell}_{\mathsf{LA}}(f, \mathbb{P}_{te}) \approx \frac{1}{N} \cdot \sum_{i=1}^{N} \ell_{\mathsf{LA}}(f_\theta(x_i), y_i; \mathbb{P}_{te}),$$

where x_i, y_i represent the i-th instance's feature and label in the training data.

Layer 2: Across various $\mathbb{P}_{te}$s, we calculate the mean and variance of $\ell_{\mathsf{LA}}(f, \mathbb{P}_{te})$ over meta-distribution $\mathcal{E}$, represented as

$$\mathbb{E}_{\mathbb{P}_{te}\sim\mathcal{E}}\left[\hat{\ell}_{\mathsf{LA}}(f; \mathbb{P}_{te})\right],\ \mathbb{V}_{\mathbb{P}_{te}\sim\mathcal{E}}\left[\hat{\ell}_{\mathsf{LA}}(f; \mathbb{P}_{te})\right],$$

respectively.

To minimize both the mean and variance of the loss over $\mathcal{E}$, we employ the following objective:

$$\mathbb{E}_{\mathbb{P}_{te}\sim\mathcal{E}}\left[\hat{\ell}_{\mathsf{LA}}(f; \mathbb{P}_{te})\right] + \lambda \cdot \mathbb{V}_{\mathbb{P}_{te}\sim\mathcal{E}}\left[\hat{\ell}_{\mathsf{LA}}(f; \mathbb{P}_{te})\right],$$

where λ is a balancing coefficient.

Given the model f as a neural network, we can't obtain the mean and variance in the closed-form. Instead, we have to approximate the mean and variance for layer 2 using the Monte Carlo method, sampling a finite number of test distributions $\mathbb{P}_{te}$ from $\mathcal{E}$ and then do the estimation.

Our training pipeline, based on the Monte Carlo method, includes:

- a) Sampling a set of test-label distributions from the Dirichlet mixture distribution $\{(\mathbb{P}_j, \xi_j)\}_{i=1}^{M}$, where $\mathbb{P}_j$ is the sampled test-label-distribution, and ξ_j indicates its component. - b) Evaluating the empirical LA loss for each $\mathbb{P}_j$ on the training data, adjusting the label distribution accordingly, then computing the mean and (semi-)variance of these losses. - c) Training the model using backpropagation (BP).

Next, we outline each step of our process.

A.1. Step a)

In the main paper, we defined the meta distribution $\mathcal{E}$ as a Dirichlet mixture distribution:

$$\mathbb{P}_{te}|\xi \sim \mathsf{Dir}\left(\alpha^{(\xi)}\right),$$

$$\xi|\boldsymbol{p} \sim \mathsf{Discrete}\left(p_1, \cdots, p_K\right).$$

To sample a test distribution $\mathbb{P}_{te}$, we first select a Dirichlet component ξ from K options based on the discrete probability $(p_1, \cdots, p_K)$. We then sample $\mathbb{P}_{te}$ from the ξ-th component $\mathsf{Dir}\left(\alpha^{(\xi)}\right)$. This process repeats to generate a set of distributions $\{\mathbb{P}_j, \xi_j\}_{i=1}^{M}$.

A.2. Step b)

For each pair $(\mathbb{P}_j, \xi_j)$, we then calculate the loss. In our Mixture of Experts (MOE) strategy, the expert ξ_j is assigned to its corresponding Dirichlet distribution component, and we train $f_\theta^{(\xi_j)}$ for this task. With this assignment,, we get the empirical LA loss average on the training data: $\hat{\ell}_{\mathsf{LA}}(f^{(\xi_j)}; \mathbb{P}_j)$ for this specific test-label distribution. Repeating the calculate for each $(\mathbb{P}_j, \xi_j)$, we can then obtain the empirical mean and variance of $\hat{\ell}_{\mathsf{LA}}(f^{(\xi_j)}; \mathbb{P}_j)$ as:

$$\hat{\mathbb{E}}(\hat{\ell}_{\mathsf{LA}}) = \frac{1}{M} \sum_{j=1}^{M} \hat{\ell}_{\mathsf{LA}}(f^{(\xi_j)}; \mathbb{P}_j),$$

$$\hat{\mathbb{V}}(\hat{\ell}_{\mathsf{LA}}) = \frac{1}{M} \sum_{j=1}^{M} \left(\hat{\ell}_{\mathsf{LA}}(f^{(\xi_j)}; \mathbb{P}_j) - \hat{\mathbb{E}}\left(\hat{\ell}_{\mathsf{LA}}(f^{(\xi_j)}; \mathbb{P}_j) \right) \right)^2.$$

Moreover, we find that the variance regularization tends to punish loss functions smaller than its mean. We use semi-variance regularization, which only penalizes loss deviations larger than the mean, as a surrogate to avoid penalizing the smaller-than-mean losses:

$$\hat{\mathbb{V}}_+(\hat{\ell}_{\mathsf{LA}}) = \frac{1}{M} \sum_{j=1}^{M} \left(\left(\hat{\ell}_{\mathsf{LA}}(f^{(\xi_j)}; \mathbb{P}_j) - \hat{\mathbb{E}}\left(\hat{\ell}_{\mathsf{LA}}(f^{(\xi_j)}; \mathbb{P}_j) \right) \right)_+ \right)^2.$$

A.3. Step c)

We perform backpropagation (BP) based on the loss and train the MOE model.

Putting altogther, we summarize this procedure as the following algorithm.

Algorithm 1 Training Algorithm

```
Require: Batch size $bs$, Dirichlet components $\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \cdots, \boldsymbol{\alpha}^{(K)}$
Ensure: Trained models $f_\theta^{(1)}, f_\theta^{(2)}, \cdots, f_\theta^{(K)}$
for $j \leftarrow 1$ to $M$ do
    $\xi_j \leftarrow$ RandomInteger$(1, K)$ with probability $(p_1, \cdots, p_K)$
    Sample $\mathbb{P}_j$ from Dirichlet Distribution $Dir(\boldsymbol{\alpha}^{(\xi_j)})$
    Add sampled pair $(\mathbb{P}_j, \xi_j)$ to the test-label-distribution set $\mathcal{P}$.
end for
while not converged do
    Sample training data Batch $B$
    Sample a subset $\mathcal{P}'$ from $\mathcal{P}$
    for each $(\mathbb{P}_j, \xi_j) \in \mathcal{P}'$ do
        Calculate $\hat{\ell}_{LA_j} = \frac{1}{bs} \sum_{(\boldsymbol{x},y) \in B_i} \hat{\ell}_{LA}(f_\theta^{(\xi_j)}(\boldsymbol{x}), y; \mathbb{P}_j)$
    end for
    Calculate the mean $\hat{\mathbb{E}}(\ell_{\mathsf{LA}})$ and semi-variance $\hat{\mathbb{V}}_+(\ell_{\mathsf{LA}})$
    Calculate the objective function $L \leftarrow \hat{\mathbb{E}}(\ell_{\mathsf{LA}}) + \lambda \cdot \hat{\mathbb{V}}_+(\ell_{\mathsf{LA}})$
    Perform SGD with respect to $L$ to update the model
end while
return $f_\theta^{(1)}, f_\theta^{(2)}, \cdots, f_\theta^{(K)}$
```

B. Proof for the Upper Bound of ρ

B.1. Proof for Thm.1

Proof. 1): For exponential distribution, we have:

$$p_\lambda(\ell) = \lambda \cdot \exp(-\lambda \cdot \ell), \ \ \underset{\lambda}{\mathbb{E}}[\ell] = \frac{1}{\lambda}, \mathbb{V}[\ell]_\lambda = \frac{1}{\lambda^2}.$$

In this sense, we have:

$$\begin{aligned} \mathbb{V}[\ell] - \mathbb{V}_+[\ell] &= \int_0^{1/\lambda} (\ell - \frac{1}{\lambda})^2 \cdot p_\lambda(\ell) \cdot d\ell \\ &\le \frac{1}{\lambda^2} \cdot \mathbb{P}\left[\ell < \frac{1}{\lambda}\right] \\ &\le \frac{1 - \exp(-1)}{\lambda^2} \end{aligned}$$

Hence:

$$\frac{\mathbb{V}_+[\ell]}{\mathbb{V}[\ell]} \ge 1 - (1 - \exp(-1)) = \exp(-1)$$

2): For the gamma distribution, we have:

$$p_{\alpha,\beta}(\ell) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \ell^{\alpha-1} \cdot \exp(-\beta \cdot \ell), \ \mathbb{E}[\ell] = \frac{\alpha}{\beta}, \ \mathbb{V}[\ell] = \frac{\alpha}{\beta^2}.$$

Similarly, we have:

$$\begin{aligned} \mathbb{V}[\ell] - \mathbb{V}_+[\ell] &= \int_0^{\alpha/\beta} (\ell - \alpha/\beta)^2 \cdot p_{\alpha,\beta}(\ell) \cdot d\ell \\ &\le \left(\frac{\alpha}{\beta}\right)^2 \cdot \frac{1}{\Gamma(\alpha)} \cdot \int_0^{\alpha/\beta} (\beta \cdot \ell)^{\alpha-1} \cdot \exp(-\beta \cdot \ell) \cdot d(\beta \cdot \ell) \\ &= \left(\frac{\alpha}{\beta}\right)^2 \cdot \frac{1}{\Gamma(\alpha)} \cdot \int_0^{\alpha} (\ell)^{\alpha-1} \cdot \exp(-\ell) \cdot d(\ell) \\ &= \left(\frac{\alpha}{\beta}\right)^2 \cdot \frac{\Gamma^\uparrow(\alpha,\alpha)}{\Gamma(\alpha)} \end{aligned}$$

Then we have:

$$\frac{\mathbb{V}_+[\ell]}{\mathbb{V}[\ell]} \ge 1 - \alpha \cdot \frac{\Gamma^\uparrow(\alpha,\alpha)}{\Gamma(\alpha)}$$

3): For a Pareto distribution: we have

$$p_\theta(\ell) = \begin{cases} \frac{\theta \ell_m^\theta}{\ell^{\theta+1}}, & \ell \ge \ell_m, \\ 0, & \ell \le \ell_m \end{cases}, \ \mathbb{E}[\ell] = \begin{cases} \infty, & \theta \le 1, \\ \frac{\theta}{\theta-1}\ell_m, & \theta > 1 \end{cases}, \mathbb{V}[\ell] = \begin{cases} \infty, & \theta \le 2, \\ \theta/(\theta-1)^2 \cdot (\theta-2) \cdot \ell_m^2, & \theta > 2 \end{cases}$$

In this sense, we have:

$$\begin{aligned} \mathbb{V}_+[\ell] &= \int_{\ell_m}^{\frac{\theta}{\theta-1}\cdot\ell_m} \left(\ell - \frac{\theta}{\theta-1} \cdot \ell_m\right)^2 \cdot \frac{\theta \ell_m^\theta}{\ell^{\theta+1}} d\ell \\ &= \int_{\ell_m}^{\frac{\theta}{\theta-1}\cdot\ell_m} \frac{\theta \ell_m^\theta}{\ell^{\theta-1}} d\ell \ - \ 2 \cdot \int_{\ell_m}^{\frac{\theta}{\theta-1}\cdot\ell_m} \frac{\theta \cdot \ell_m^{\theta+1}}{(\theta-1) \cdot \ell^{\theta-1}} d\ell \ + \ \left(\frac{\theta}{\theta-1} \cdot \ell_m\right)^2 \cdot \int_{\ell_m}^{\frac{\theta}{\theta-1}\cdot\ell_m} \frac{\theta \ell_m^\theta}{\ell^{\theta+1}} d\ell \end{aligned}$$

$$\int_{\ell_m}^{\frac{\theta}{\theta-1}\cdot\ell_m} \frac{\theta\ell_m^{\theta}}{\ell^{\theta-1}} d\ell \;=\; \frac{\theta}{\theta-2}\cdot\ell_m^2\cdot\left[1-\left(\frac{\theta}{\theta-1}\right)^{2-\theta}\right]$$

$$2\cdot\int_{\ell_m}^{\frac{\theta}{\theta-1}\cdot\ell_m} \frac{\theta\cdot\ell_m^{\theta+1}}{(\theta-1)\cdot\ell^{\theta-1}}\cdot d\ell \;=\; 2\cdot\left(\frac{\theta}{\theta-1}\cdot\ell_m\right)^2\cdot\left[1-\left(\frac{\theta}{\theta-1}\right)^{1-\theta}\right]$$

$$\left(\frac{\theta}{\theta-1}\cdot\ell_m\right)^2\cdot\int_{\ell_m}^{\frac{\theta}{\theta-1}\cdot\ell_m} \frac{\theta\ell_m^{\theta}}{\ell^{\theta+1}} d\ell \;=\; \left(\frac{\theta}{\theta-1}\cdot\ell_m\right)^2\cdot\left[1-\left(\frac{\theta}{\theta-1}\right)^{\theta}\right]$$

Above all, we come to the conclusion:

$$\frac{\mathbb{V}_+[\ell]}{\mathbb{V}[\ell]} = (\theta-1)^2\cdot\left[1-\left(\frac{\theta}{\theta-1}\right)^{2-\theta}\right] + \theta\cdot(\theta-2)\cdot\left[2\cdot\left(\frac{\theta}{\theta-1}\right)^{1-\theta} - \left(\frac{\theta-1}{\theta}\right)^{\theta} - 1\right]$$

□

B.2. Proof for Thm.2

Lemma 1. *The function $f(x) = ((x)_-)^2$ is a convex for $x \neq 0$.*

Proof. Denote $f_1(x) = x^2, f_2(x) = (x)_-$, then we have:

$$f''(x) = f_2''(x)\cdot f_1'(f_2(x)) + f_1''(f_2(x))\cdot(f_2'(x))^2 = \begin{cases} 0, & x > 0 \\ 2, & x < 0. \end{cases}$$

Obviously, we have $f''(x) \geq 0$. The proof is thus finished.

□

Proof. We consider a mixture distribution of K-components with a p.d.f function

$$p_m(\ell) = \sum_{k=1}^{K} \omega_k \cdot p_k(\ell),$$

where $p_k(\ell)$ is the p.d.f. for the k-th component, and ω_k is probability to observe the k-th component. In this sense, we have:

$$\mathbb{E}[\ell] = \sum_{k=1}^{K} \omega_k\cdot\mu_k,\;\; \mathbb{V}[\ell] \overset{(a)}{\geq} \sum_{k=1}^{K} \omega_k\cdot\sigma_k^2.$$

where μ_k, σ_k^2 are the corresponding means and variance given the component k. Here (a) follows from the total law of variance.

In this sense, we can bound the semi-variance as:

$$
\begin{aligned}
\mathbb{V}_-[\ell] &= \int_0^\infty ((\ell - \mathbb{E}[\ell])_-)^2 \cdot p_m(\ell) \cdot d\ell \\
&= \int_0^\infty \left((\ell - \sum_{k=1}^{k} \omega_k \cdot \mu_k)_- \right)^2 \cdot p_m(\ell) \cdot d\ell \\
&\overset{(*)}{\leq} \sum_{k=1}^{k} \omega_k \cdot \int_0^\infty ((\ell - \mu_k)_-)^2 \cdot p_m(\ell) \cdot d\ell \\
&= \sum_{i,j} \omega_i \cdot \omega_j \cdot \int_0^\infty ((\ell - \mu_i)_-)^2 \cdot p_j(\ell) \cdot d\ell
\end{aligned}
$$

$$
\begin{aligned}
1 - \rho \leq & \frac{1}{\mathbb{V}[\ell]} \cdot \left(\sum_{i,j} \omega_i \cdot \omega_j \cdot \int_0^\infty ((\ell - \mu_i)_-)^2 \cdot p_j(\ell) \cdot d\ell \right) \\
\overset{(**)}{\leq} & \sum_{i,j} \left(\omega_i \cdot \omega_j \cdot \int_0^\infty ((\ell - \mu_i)_-)^2 \cdot p_i(\ell) \cdot d\ell \cdot \frac{1}{\sigma_i^2} \right) \\
\leq & \sum_i \omega_i \cdot \frac{\mathbb{V}_-[\ell]_i}{\mathbb{V}[\ell]_i}
\end{aligned}
$$

$(*)$ is from the fact that $((\cdot)_-)^2$ is a convex function when $x \neq 0$, and that $(**)$ is from the assumption that:

$$
\frac{\mathbb{V}_-[\ell]_{i,j}}{\mathbb{V}_-[\ell]_i} \leq \frac{\mathbb{V}[\ell]}{\sigma_i^2}, \ \frac{\mathbb{V}[\ell]}{\sigma_i^2} > 1
$$

which further implies that

$$
\frac{1}{\mathbb{V}[\ell]} \cdot \int_0^\infty ((\ell - \mu_i)_-)^2 \cdot p_j(\ell) \cdot d\ell \leq \frac{1}{\sigma_i^2} \cdot \int_0^\infty ((\ell - \mu_i)_-)^2 \cdot p_i(\ell) \cdot d\ell
$$

It thus becomes clear that:

$$
\begin{aligned}
\rho &= 1 - (1 - \rho) \\
&\overset{(***)}{\geq} 1 - \sum_i \omega_i \cdot \frac{\mathbb{V}_-[\ell]_i}{\mathbb{V}[\ell]_i} \\
&= \sum_i \omega_i - \sum_i \omega_i \cdot \frac{\mathbb{V}_-[\ell]_i}{\mathbb{V}[\ell]_i} \\
&= \sum_i \omega_i \cdot \left(1 - \frac{\mathbb{V}_-[\ell]_i}{\mathbb{V}[\ell]_i} \right) \\
&= \sum_i \omega_i \cdot (1 - (1 - \rho_i)) \\
&= \sum_{i=1}^{K} \omega_i \cdot \rho_i
\end{aligned}
$$

Here $(***)$ is from the proven fact that $1 - \rho \ \leq \ \sum_i \omega_i \cdot \frac{\mathbb{V}_-[\ell]_i}{\mathbb{V}[\ell]_i}$

□

C. Explanations for the Theoretical Results

Basic Notations: In our approach, we use a natural training dataset consisting of images and their labels from the distribution $\mathcal{D}$, and a constructed set of potential test-label distributions from $\mathcal{E}$. We denote the training dataset as $\mathcal{S}$ and the constructed set as $\mathcal{P}$. Additionally, the training sample size is N and the constructed set size is M.

Our primary theoretical findings relate to how well our method generalizes to new, unseen test-label distributions. According to standard learning theory, we measure generalization error by the difference between a) the expected error across the joint distribution of test-label and training data, and b) the empirical average across the sampled $\mathcal{S}$ and $\mathcal{P}$. This difference results from stochastic errors in sampling and estimation. Reflecting on the previous question, the Monte Carlo process incorporates a hierarchical sampling approach, also leading to a hierarchical structure of stochastic errors, a significant challenge in our theoretical analysis. This explain this below.

C.1. Stochastic Errors

Inner Layer (training data): For a specific test label distribution $\mathbb{P}_i$ in $\mathcal{P}$, we define $\ell_{\mathcal{D},\mathcal{E},i}$ and $\ell_{\mathcal{S},\mathcal{E},i}$ as the expected loss on training distribution $\mathcal{D}$ and the empirical average loss on training data $\mathcal{S}$, respectively:

$$\ell_{\mathcal{D},\mathcal{E},i} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell_{\mathsf{LA}}(f^{(\xi_i)}(x), y; \mathbb{P}_i)\right]$$

$$\ell_{\mathcal{S},\mathcal{E},i} = \frac{1}{N}\sum_{j=1}^{N}\ell_{\mathsf{LA}}(f^{(\xi_i)}(x_j), y_j; \mathbb{P}_i)$$

These values estimate the error of the inner layer concerning the training data, specifically measuring the discrepancy when substituting expectation with empirical average for a given test-label distribution $\mathbb{P}_i$.

Outer Layer (test label distributions): We also consider the expectation over the meta-distribution $\mathcal{E}$ to assess the outer Monte Carlo sampling error.

In practical training, we rely on Monte Carlo estimation results and finite training data to compute the empirical average:

$$\hat{\mathbb{E}}_{\mathcal{E}}\left[\ell_{\mathcal{S},\mathcal{E},i}\right] = \frac{1}{NM}\sum_{i=1}^{M}\sum_{j=1}^{N}\ell_{\mathsf{LA}}(f^{(\xi_i)}(x_j), y_j; \mathbb{P}_i)$$

But Theoretically, the genearlization error should be measured by the expected loss on the joint distribution of test label distribution $\mathcal{E}'$ (which could differ from $\mathcal{E}$ used in training) and training data $\mathcal{D}$, expressed as:

$$\mathbb{E}_{\mathbb{P}\sim\mathcal{E}'}\left[\ell_{\mathcal{D},\mathcal{E}',i}\right] = \mathbb{E}_{\mathbb{P}\sim\mathcal{E}'}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell_{\mathsf{LA}}(f^{(\xi)}(x), y; \mathbb{P})\right]\right]$$

Finally, we discuss the upper bound of the generalization error and briefly introduce the proof's key idea.

Error Decomposition

Assume that our model is chosen from a hypothesis space $\mathcal{F}$ (for instance, CNNs within certain weight norm limits). The generalization ability of the entire hypothesis set is often measured by the worst-case performance gap:

$$\Delta = \sup_{f\in\mathcal{F}}\left[\mathbb{E}_{\mathbb{P}\sim\mathcal{E}'}\left[\ell_{\mathcal{D},\mathcal{E}',i}\right] - \hat{\mathbb{E}}_{\mathcal{E}}\left[\ell_{\mathcal{S},\mathcal{E},i}\right]\right].$$

Analyzing this error directly is challenging due to its hierarchical nature. Nevertheless, we can further derive an upper bound by summing three types of error:

i) **Meta-distribution Approximation Error**: This error arises from approximating the ideal meta-distribution $\mathcal{E}'$ with $\mathcal{E}$, expressed as:

$$\sup_{f\in\mathcal{F}} \left[\mathbb{E}_{\mathbb{P}\sim\mathcal{E}'}\left[\ell_{\mathcal{D},\mathcal{E}',i}\right] - \mathbb{E}_{\mathbb{P}\sim\mathcal{E}}\left[\ell_{\mathcal{D},\mathcal{E},i}\right]\right].$$

ii) **Label Dist**: This error stems from the approximation of the expected values over $\mathcal{E}$ with a Monte Carlo average, which can be shown as:

$$\sup_{f\in\mathcal{F}} \left[\mathbb{E}_{\mathbb{P}\sim\mathcal{E}}\left[\ell_{\mathcal{D},\mathcal{E},i}\right] - \frac{1}{M}\sum_{i=1}^{M}\ell_{\mathcal{D},\mathcal{E},i}\right].$$

iii) **Data Estimation Error**: This error is due to approximating the expectation over the training distribution $\mathcal{D}$ with the empirical average from the training data S:

$$\sup_{f\in\mathcal{F}} \left[\mathbb{E}_{\mathbb{P}\sim\mathcal{E}}\left[\frac{1}{M}\sum_{i=1}^{M}\ell_{\mathcal{D},\mathcal{E},i} - \frac{1}{M}\sum_{i=1}^{M}\ell_{\mathcal{S},\mathcal{E},i}\right]\right].$$

C.2. Key Idea of the Proof:

From the error decomposition i)-iii) above, we outline an overall limit for the generalization gap Δ.

- For i), we establish an upper bound using the loss function's boundedness, the probability simplex's volume, and the largest variation between two label distributions (measured by the inf-norm $\|\cdot\|_\infty$).

- For ii), the most complex part of this proof, we extend the Bernstein-type concentration bounds to upper bound the error with the empirical semi-variance from our objective function. This approach shows how the objective function directly enhances generalization.

- For iii), we translate this into the uniform convergence bound, where the covering number indicates the hypothesis class's complexity.

D. Proof for Overall Generalization Upper Bound

D.1. Basic Definitions of The Hypothesis Class

The Hypothesis Class. In this paper, we consider the Generalization ability based on the employed architecture. This means that we only consider models chosen from the following hypothesis class $\mathcal{F}$:

$$\mathcal{F} = \left\{\mathbb{R}^C \leftarrow \mathbb{R}^d : f_\theta^{(i)}(\cdot) = g^{(i)} \circ \psi(\cdot), i \in [K]., \ g^{(i)}, \psi \text{ are chosen from specific subclass of deep neural networks.}\right\}$$

Note that since $f_\theta^{(i)}$ are scoring functions for multi-class classification problems, they must be **vector-valued functions**. For the sake of simplicity, **we will use $f(x)$ under this context to denote the collection** $\{f_\theta^{(i)}\}_{i\in[K]}$. We use $\underline{f \in \mathcal{F}}$ to express choosing one such collections out of $\mathcal{F}$.

The Norm of Hypotheses. To measure the complexity of $\mathcal{F}$, we must define a norm on each hypothesis f. To do this, we adopt the overall infinity norm (over all K subbranches, all C classes, and all input features $x \in \mathcal{X}$):

$$\|f\|_\infty = \max_{i\in[K]} \max_{j\in\{1,2,\cdots,C\}} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|f_{\theta,j}^{(i)}(\boldsymbol{x})\right|.$$

Here $f_{\theta,j}^{(i)}(\boldsymbol{x})$ means the output score of the i-th branch and j-th (channel) for feature $\boldsymbol{x}$. It is easy to check that the new infinity norm is also a norm.

Measuring Complexity with Covering Number. Given a functional class $\mathcal{F}$, we can use the covering number $\mathcal{N}_\infty(\mathcal{F}, \epsilon, n)$ to measure its corresponding complexity with the following definition.

$$\mathcal{N}_\infty(\mathcal{F}, \epsilon, n) = \sup_{\boldsymbol{x}\in\mathcal{X}^n} \mathcal{N}(\epsilon, \mathcal{F}, ||\cdot||_\infty)$$

where $\mathcal{N}(\epsilon, \mathcal{F}, ||\cdot||_\infty)$ is the smallest number of infinity norm open balls, denoted as N_o, such that there exist N_o open balls that covers entire class $\mathcal{F}$. In other words, $\exists\, \mathcal{B}_i \subseteq \mathcal{F}$, such that $\mathcal{F} \subseteq \bigcup_{i=1}^{N_o} \mathcal{B}_i$, where $\mathcal{B}_i = \{f : \|f - f_i\|_\infty \le \epsilon\}$.

In the proof, we will the following lemmas to as basic tools.

D.2. Fundamental Inequalities

Lemma 2. *The volume of an c-dimensional probability simplex is* $c!$*. In other words, we have:*

$$V_c = \int\limits_{\sum_{i=1}^{c} x_i = 1} 1 \cdot dx_1 \cdot dx_2 \cdots dx_c = \frac{1}{c!}$$

Proof. We proof it by induction.

Base Case, c=1 Obviously, we have:

$$V_1 = \int_0^1 dx = 1.$$

Induction Supposes that $V_{i-1} = (i-1)!$, we have:

$$\begin{aligned}
V_i &= \int\limits_{\sum_{j=1}^{i} x_j = 1} 1 \cdot dx_1 \cdot dx_2 \cdots dx_i \\
&= \int_0^1 \left(\int\limits_{\sum_{j=1}^{i-1} x_j = 1 - x_i} 1 \cdot dx_1 \cdot dx_2 \cdots dx_{i-1} \right) dx_i \\
&\overset{u_j = x_j/(1-x_i), j=1,2,\cdots,i-1}{=} \int_0^1 (1-x_i)^{i-1} \cdot \left(\int\limits_{\sum_{j=1}^{i-1} u_j = 1} 1 \cdot du_1 \cdot du_2 \cdots du_{i-1} \right) dx_i \\
&= V_{i-1} \cdot \int_0^1 (1-x_i)^{i-1} dx_i \\
&= (1/i) \cdot V_{i-1} \\
&= 1/i!
\end{aligned}$$

The proof is then completed by expanding the induction recursively. □

Lemma 3. *When $g(\cdot)$ is Lipschitz continuous, the following holds:*

$$\|g(x) - g(\tilde{x})\|_\infty \le \sup \|\nabla_x g\|_p \cdot \|x - \tilde{x}\|_q, \tag{6}$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

Proof.

$$\begin{aligned}
|g(x) - g(\tilde{x})| &= \left| \int_0^1 \langle \nabla g(\tau x + (1-\tau)\tilde{x}), x - \tilde{x} \rangle \, d\tau \right| \\
&\le \sup_{x \in \mathcal{X}} \left[\|\nabla g\|_p \right] \cdot \left\| x - \tilde{x} \right\|_q
\end{aligned} \tag{7}$$

□

Lemma 4. *Let $\mathbb{P}_{tr}[y]$ be training label distribution and $\mathbb{P}_{te}[y]$ a test label distribution. We denote their ratio as $q_i = \frac{\mathbb{P}_{te}[i]}{\mathbb{P}_{tr}[i]}$, $i = 1, 2, \cdots, C$. Then we have the LA loss:*

$$\ell_{LA}\left(f_\theta^\xi(\boldsymbol{x}), j; \mathbb{P}_j\right) = \ell_{CE}\left(\mathsf{softmax}\left(f_{y,\theta}^\xi(\boldsymbol{x}) - \log\left(q_y\right)\right)\right)$$

is 2-Lip. continuous w.r.t. the defined infinity norm.

Proof. According to Lem.3, if for any fixed ξ:

$$\sup_{(\boldsymbol{x},y)\in\mathcal{Z},f\in\mathcal{F}} \left\|\nabla_f \ell_{LA}\left(f_\theta^\xi(\boldsymbol{x}), j; \mathbb{P}_{tr}\right)\right\|_1 \leq 2 \tag{8}$$

Then we have:

$$\left|\ell_{LA}\left(f_\theta^\xi(\boldsymbol{x}), j; \mathbb{P}_j\right) - \ell_{LA}\left(\tilde{f}_\theta^\xi(\boldsymbol{x}), j; \mathbb{P}_j\right)\right| \leq 2\cdot\|f^\xi - \tilde{f}^\xi\|_\infty \leq 2\cdot \max_{i\in[K]} \|f^{(i)} - \tilde{f}^{(i)}\|_\infty = 2\|f-\tilde{f}\|_\infty$$

Hence, we only need to proof (8). To see this,

$$\begin{aligned}\left|\frac{\partial \ell_{LA}\left(f_\theta(\boldsymbol{x}), j; \mathbb{P}_{tr}\right)}{\partial f_\theta^{(j)}(\boldsymbol{x})}\right| &= \left|\frac{\partial\left(\log\left[\sum_i \exp(f^{(i)}(\boldsymbol{x}) - q_i) - (f_y - q_y)\right]\right)}{\partial f^{(j)}(\boldsymbol{x})}\right| \\ &= \left|\mathsf{softmax}\left(f_\theta^{(j)}(\boldsymbol{x}) - \log\left(q_j\right)\right) - I[j=y]\right|.\end{aligned}$$

Since we have:

$$\begin{aligned}\left\|\nabla_f \ell_{LA}\left(f_\theta(\boldsymbol{x}), y; \mathbb{P}_{tr}\right)\right\|_1 &= \sum_j \left|\mathsf{softmax}\left(f_\theta^{(j)}(\boldsymbol{x}) - \log\left(q_j\right)\right) - I[j=i]\right| \\ &= 2\cdot\left(1 - \mathsf{softmax}\left(f_\theta^{(y)}(\boldsymbol{x}) - \log\left(q_y\right)\right)\right) \\ &\leq 2.\end{aligned}$$

The proof is completed since $\boldsymbol{x}, y, f$ are arbitrarily chosen.

□

Lemma 5. *Given a meta-distribution $\mathcal{E}$ of label distributions, let $\mathcal{E}$ be the Dirichlet mixture distribution of the components defined in the main paper. Let*

$$\mathcal{P} \sim \mathcal{E}^M,\ \mathcal{P} = \{\mathbb{P}_1, \cdots, \mathbb{P}_M\}$$

For any function $\ell(f)$ with the property that $\ell(f) = \ell(f^j)$, for the j Dirichlet component, if for any fixed j, $\ell(f^j)$ is L-Lip. continuous w.r.t to the infinity norm, then $\mathbb{E}_\mathcal{E}[\ell(f)], \hat{\mathbb{E}}_\mathcal{E}[\ell(f)]$ is also L-Lip. continuous w.r.t to the defined general infinity norm.

Proof. We only prove the result for $\mathbb{E}_\mathcal{E}[\ell(f)]$, since the result for $\hat{\mathbb{E}}_\mathcal{E}[\ell(f)]$ can be proven similarly. Since

$$\mathop{\mathbb{E}}_{\mathcal{E}}[\ell(f)] = \sum_{i\in[K]} p_i \cdot \mathop{\mathbb{E}}_{\mathcal{E}|i}[\ell(f^i)],$$

we have:

$$|\mathop{\mathbb{E}}_{\mathcal{E}}[\ell(f)] - \mathop{\mathbb{E}}_{\mathcal{E}}[\ell(\tilde{f})]| \leq \sum_{i\in[K]} p_i \cdot |\mathop{\mathbb{E}}_{\mathcal{E}|i}[\ell(f^i)] - \mathop{\mathbb{E}}_{\mathcal{E}|i}[\ell(\tilde{f}^i)]| \leq \sum_{i\in[K]} L\cdot p_i \cdot \|f^i - \tilde{f}^i\|_\infty \leq L\cdot\|f-\tilde{f}\|_\infty$$

□

D.3. Basic (Uniform) Concentration Inequalities

Please refer to Sec.D.1 for basic definitions of hypothesis class $\mathcal{F}$, covering number $\mathcal{N}_\infty$, and the infinity norm $\|\cdot\|_\infty$.

Lemma 6. *(Duchi & Namkoong, 2016) Let $f:\mathbb{R}^N\to\mathbb{R}$ be convex and L-Lip. continuous w.r.t to ℓ_2 norm over $[a,b]^n$, and let $Z_1,\cdots,Z_N$ be independent random variables on $[a,b]$. Then for all $t\geq 0$*

$$\max\left\{\mathbb{P}\left[f(Z_{1:N})\geq\mathbb{E}\left[f(Z_{1:N})\right]\right]+t,\ \mathbb{P}\left[f(Z_{1:N})\leq\mathbb{E}\left[f(Z_{1:N})\right]-t\right]\right\}\leq\exp\left(-\frac{t^2}{2L^2(b-a)^2}\right)$$

Lemma 7 ((Maurer & Pontil, 2009))**.** *Let $\{\boldsymbol{x}_i,y_i\}_{i=1}^N$ be i.i.d samples from a data distribution, the loss function of a prediction is given by $\ell(\boldsymbol{x}_i,y_i)=\ell(f(\boldsymbol{x}_i),y_i)$ for a scoring function $f\in\mathcal{F}$. For sufficiently large N and sufficiently small δ, with probability at least $1-\delta$ over the randomness of the data sampling, we have:*

$$\mathbb{E}[\ell]\lesssim\hat{\mathbb{E}}[\ell]+\sqrt{\frac{\hat{\mathbb{V}}[\ell]\cdot\log\left(\mathcal{M}/\delta\right)}{N}}+\frac{B\cdot\log\left(\mathcal{M}/\delta\right)}{N}$$

holds uniformly for all $f\in\mathcal{F}$, with $\mathcal{M}=\mathcal{N}_\infty(\mathcal{F},1/N,2N)$.

Lemma 8. *Let $\{\boldsymbol{x}_i,y_i\}_{i=1}^N$ be i.i.d samples from a data distribution, the loss function of a prediction is given by $\ell(\boldsymbol{x}_i,y_i)=\ell(f(\boldsymbol{x}_i),y_i)\in[0,B]$ for a scoring function $f\in\mathcal{F}$. If the loss function is L-Lip. continuous w.r.t. the infinity norm $||f||_\infty=\max_{\boldsymbol{x}\in\mathcal{X}}|f(\boldsymbol{x})|$. Then the following inequality holds with probability at least $1-\delta$ uniformly for all $f\in\mathcal{F}$:*

$$\left|\mathbb{E}[\ell]-\hat{\mathbb{E}}[\ell]\right|\lesssim B\cdot\sqrt{\frac{\log(\mathcal{M}'/\delta)}{N}}$$

where $\mathcal{M}'=\mathcal{N}_\infty(\mathcal{F},1/2LN,N)$.

Proof. According to the basic property of the covering number, we can find a covering of $\mathcal{F}$ with a set of open ball $\{\mathcal{B}_1,\cdots,\mathcal{B}_{\mathcal{M}'}\}$. For each $\mathcal{B}_j$ the center is defined as f_j. In this sense, we have $\mathcal{B}_j=\{f\in\mathcal{F}:\|f-f_j\|_\infty\leq\epsilon/3L\}$

In this sense we have the following results according to the union bound:

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left[\left|\mathbb{E}[\ell]-\hat{\mathbb{E}}[\ell]\right|\right]\geq t+\epsilon\right]\leq\sum_{j=1}^{\mathcal{M}'}\mathbb{P}\left[\sup_{f\in\mathcal{B}_j}\left[\left|\mathbb{E}[\ell]-\hat{\mathbb{E}}[\ell]\right|\right]\geq t+\epsilon\right]$$

Fix one j, we must have:

$$\begin{aligned}\mathbb{P}\left[\sup_{f\in\mathcal{B}_j}\left[\left|\mathbb{E}[\ell]-\hat{\mathbb{E}}[\ell]\right|\right]\geq t+\epsilon\right]\leq&\mathbb{P}\left[\sup_{f\in\mathcal{B}_j}\left[\left|\mathbb{E}[\ell]-\mathbb{E}[\ell_j]\right|\right]\geq\frac{\epsilon}{2}\right]\\&+\mathbb{P}\left[\left|\hat{\mathbb{E}}[\ell_j]-\hat{\mathbb{E}}[\ell_j]\right|\geq t\right]+\mathbb{P}\left[\sup_{f\in\mathcal{B}_j}\left[\left|\hat{\mathbb{E}}[\ell]-\hat{\mathbb{E}}[\ell_j]\right|\right]\geq\frac{\epsilon}{2}\right]\end{aligned}\tag{9}$$

Since ℓ is L-Lip. w.r.t the infinity norm, we have:

$$|\ell-\ell_j|\leq L\cdot\|f-f_j\|_\infty\leq L\cdot\frac{\epsilon}{2L}\leq\frac{\epsilon}{2}.$$

Hence, $\sup_{f\in\mathcal{B}_j}\left[\left|\mathbb{E}[\ell]-\mathbb{E}[\ell_j]\right|\right]\geq\frac{\epsilon}{2}$, $\sup_{f\in\mathcal{B}_j}\left[\left|\hat{\mathbb{E}}[\ell]-\hat{\mathbb{E}}[\ell_j]\right|\right]\geq\frac{\epsilon}{2}$ has probability 0.

Moreover, according to the Hoeffding's inequality, we have:

$$\mathbb{P}\left[\left|\mathbb{E}[\ell_j]-\hat{\mathbb{E}}[\ell_j]\right|\geq t\right]\leq 2\cdot\exp\left(-\frac{2\cdot t^2}{B^2}\right)$$

Combing the arguments all above, we have:

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left[\left|\mathbb{E}[\ell]-\hat{\mathbb{E}}[\ell]\right|\right]\le\epsilon\right]\ge 1-\mathcal{N}_\infty(\mathcal{F},\epsilon/2L,N)\cdot\exp\left(-\frac{2t^2}{B^2}\right).$$

The proof is finished by setting $\delta=\exp\left(-\frac{t^2}{B^2}\right)$ and $\epsilon=1/N$.

□

Lemma 9. *Let $\{\boldsymbol{x}_i,y_i\}_{i=1}^N$ be i.i.d samples from a data distribution, the loss function of a prediction is given by $\ell(\boldsymbol{x}_i,y_i)=\ell(f(\boldsymbol{x}_i),y_i)\in[0,B]$ for a scoring function $f\in\mathcal{F}$. If ℓ is L-Lip. continuous, then for any L_g Lip. continuous function of f, we have the following inequality holds with probability at least $1-\delta$ uniformly for all $f\in\mathcal{F}$:*

$$\left|g\Big(\mathbb{E}[\ell]\Big)-g\left(\hat{\mathbb{E}}[\ell]\right)\right|\lesssim L_g\cdot B\cdot\sqrt{\frac{\log(\mathcal{M}'/\delta)}{N}}$$

where $\mathcal{M}'=\mathcal{N}_\infty(\mathcal{F},1/2LN,N)$.

Proof. Since g is L_g Lip., we have:

$$\sup_{f\in\mathcal{F}}\left[\left|g\Big(\mathbb{E}[\ell]\Big)-g\left(\hat{\mathbb{E}}[\ell]\right)\right|\right]\le L_g\sup_{f\in\mathcal{F}}\left[\left|\mathbb{E}[\ell]-\hat{\mathbb{E}}[\ell]\right|\right]$$

Thus we have:

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left[\left|g\Big(\mathbb{E}[\ell]\Big)-g\left(\hat{\mathbb{E}}[\ell]\right)\right|\right]\ge t+\epsilon\right]\le\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left[\left|\mathbb{E}[\ell]-\hat{\mathbb{E}}[\ell]\right|\right]\ge\frac{t+\epsilon}{L_g}\right]$$

Then the claim follows from Lem.8 using $\epsilon'=\frac{\epsilon}{L_g},\ t'=\frac{t}{L_g}$.

□

D.4. Proof of the Main Result

D.4.1. Notations in the Proof

The Hierarchy of Stochastic Error. Recall that we are using a stratified sampling process for the Dirichlet mixture distribution. Hence, we each component j of the mixture, we sample M label distributions, denoted as

$$\mathcal{P}\sim\mathcal{E}^M,\ \mathcal{P}=\{\mathbb{P}_1,\cdots,\mathbb{P}_M\}$$

This forms an empirical sample of the test label distributions. We denote corresponding overall estimation as $\hat{\mathbb{E}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}],\hat{\mathbb{V}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$. Moreover, we also have a fixed training dataset

$$\mathcal{S}\sim\mathcal{D}^N,\mathcal{S}=\{\boldsymbol{x}_i,y_i\}_{i=1}^N,$$

This forms an empirical estimation of the training data distribution. We mark the corresponding empirical (population resp.) estimations with subscript $\mathcal{S}$ ($\mathcal{D}$ resp.). To estimate the excess risk, we must estimate the hierarchical stochastic error coming from label distribution sampling and data sampling. To do this, we present the following error decompose regime.

In this proof, we will employ two groups of intermediate empirical estimations.

The first group involves quantities for a fixed test label distribution $\mathbb{P}_i\in\mathcal{P}$. Specifically, we denote $\ell_{\mathcal{D},\mathcal{E},i},\ell_{\mathcal{S},\mathcal{E},i}$ as the expected loss on distribution $\mathcal{D}$ (empirical average loss on the training data $\mathcal{D}$ resp.):

$$\begin{aligned}\ell_{\mathcal{D},\mathcal{E},i}&=\mathop{\mathbb{E}}_{\mathcal{D}}\left[\ell_{\mathsf{LA}}(f^{(\xi_i)}(\boldsymbol{x}),y;P_i)\right]\\ \ell_{\mathcal{S},\mathcal{E},i}&=\mathop{\hat{\mathbb{E}}}_{\mathcal{S}}\left[\ell_{\mathsf{LA}}(f^{(\xi_i)}(\boldsymbol{x}),y;P_i)\right]\end{aligned}\qquad(10)$$

Table 4. Some Important Notations Used in the Proof.

Notation	Description
Basic Quantities	
N	The number of data instances in the training data.
M	The number of sampled test distributions in the Monte Carlo process.
τ	A specific test label distribution.
$\mathcal{E}$	The meta-distribution of label distributions.
$\hat{\ell}_{\mathcal{S},\mathcal{E}}$	$\sum_{(\boldsymbol{x}_i y_i)\in\mathcal{S}} \ell_{\mathsf{LA}}(f_\theta(\boldsymbol{x}_i), y_i; \mathbb{P})$ The expected loss when the label distribution is fixed (for example $\mathbb{P}$).
$\ell_{\mathcal{D},\mathcal{E}}$	The empirical loss on training data $\mathcal{S}$ when the label distribution is fixed.
$\ell_{\mathcal{S},\mathcal{E},i}$	$\hat{\mathbb{E}}_{\mathcal{S}}\left[\ell_{\mathsf{LA}}(f^{(\xi_i)}(\boldsymbol{x}), y; P_i)\right]$ The empirical risk on training data $\mathcal{S}$ for $\mathcal{P}_i \in \mathcal{P}$
$\ell_{\mathcal{D},\mathcal{E},i}$	$\mathbb{E}_{\mathcal{D}}\left[\ell_{\mathsf{LA}}(f^{(\xi_i)}(\boldsymbol{x}), y; P_i)\right]$ The empirical risk on training data distribution $\mathcal{D}$ for $\mathcal{P}_i \in \mathcal{P}$
$\bar{\ell}_{\mathcal{P}}$	$\frac{1}{M}\sum_{m=1}^{M} \ell_{\mathsf{LA}}(f^{(\xi_m)}(\boldsymbol{x}), y; P_m)$ The empirical average over $\mathcal{P}$ given a fixed sample pair $(\boldsymbol{x}, y)$.
	Estimations based on the true meta-distribution $\mathcal{E}$
$\mathbb{E}_{\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]$	The expected $\hat{\ell}_{\mathcal{S},\mathcal{E}}$ over the meta-distribution of label distributions
$\mathbb{E}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$	The expected $\ell_{\mathcal{D},\mathcal{E}}$ over the meta-distribution of label distributions
$\mathbb{V}_{\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]$	The variance of $\hat{\ell}_{\mathcal{S},\mathcal{E}}$ over the meta-distribution of label distributions
$\mathbb{V}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$	The variance $\ell_{\mathcal{D},\mathcal{E}}$ over the meta-distribution of label distributions
$\mathbb{V}_{+,\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$	The semi-variance $\ell_{\mathcal{D},\mathcal{E}}$ over the meta-distribution of label distributions
	Estimations based on Empirical meta-distribution $\mathcal{P}$
$\hat{\mathbb{E}}_{\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]$	The empirical average of $\hat{\ell}_{\mathcal{S},\mathcal{E}}$ over the sampled label distributions in the Monte Carlo process.
$\hat{\mathbb{E}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$	The empirical average of $\ell_{\mathcal{D},\mathcal{E}}$ oover the sampled label distributions in the Monte Carlo process.
$\hat{\mathbb{V}}_{\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]$	The empirical variance of $\hat{\ell}_{\mathcal{S},\mathcal{E}}$ over the sampled label distributions in the Monte Carlo process.
$\hat{\mathbb{V}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$	The empirical variance $\ell_{\mathcal{D},\mathcal{E}}$ over the sampled label distributions in the Monte Carlo process.
$\hat{\mathbb{V}}_{+,\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]$	The empirical semi-variance of $\hat{\ell}_{\mathcal{S},\mathcal{E}}$ over the sampled label distributions in the Monte Carlo process.
$\hat{\mathbb{V}}_{+,\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$	The empirical semi-variance $\ell_{\mathcal{D},\mathcal{E}}$ over the sampled label distributions in the Monte Carlo process.
	Complexity Measures
$\mathcal{N}_\infty(\mathcal{F}, \epsilon, n)$	The covering number for a hypothesis class $\mathcal{F}$, with radius of the covering open ball chosen as ϵ.

The other groups of quantities regard the average loss under all sampled label distributions as a fixed loss and then evaluate their sample- and population-level means. Specifically, let

$$\bar{\ell}_{\mathcal{P}} = \frac{1}{M}\sum_{m=1}^{M} \ell_{\mathsf{LA}}(f^{(\xi_m)}(\boldsymbol{x}), y; P_m),$$

then we can rewrite $\hat{\mathbb{E}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}], \hat{\mathbb{E}}_{\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]$ as:

$$\begin{aligned}\hat{\underset{\mathcal{E}}{\mathbb{E}}}[\ell_{\mathcal{D},\mathcal{E}}] &= \underset{\mathcal{D}}{\mathbb{E}}\left[\bar{\ell}_{\mathcal{P}}\right]\\ \hat{\underset{\mathcal{E}}{\mathbb{E}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}] &= \hat{\underset{\mathcal{S}}{\mathbb{E}}}\left[\bar{\ell}_{\mathcal{P}}\right]\end{aligned} \tag{11}$$

D.4.2. The Formal Restate and Proof

Theorem 5 (Restate of Thm.1). *For label distributions, let $\mathcal{E}$ be the true meta-distribution and $\mathcal{P}$ be a observed empirical distribution for label distribution sampled by the Monte Carlo method. Moreover, the training data $\mathcal{S}$ are sampled in a i.i.d manner from its true distribution $\mathcal{D}$. Assume that $\mathcal{N}_\infty(\mathcal{F}, \epsilon, M) \le \left(\frac{r}{\epsilon}\right)^{\nu}$, $\ell(\cdot) \in [0, B]$, and $\mathbb{V}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] \asymp \mathbb{V}_{+,\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$ following the notations in Tab.4, all any possible meta-distribution $\mathcal{E}'$ defined on the probability simplex $\mathbb{S}^{c-1}$, the following* *inequality holds uniformly for all $f \in \mathcal{F}$ with probability at least $1-\delta$ over the randomness of $\mathcal{S}, \mathcal{P}$:*

$$\mathop{\mathbb{E}}_{\mathcal{E}'}[\ell_{\mathcal{D},\mathcal{E}'}] \lesssim \mathop{\hat{\mathbb{E}}}_{\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}] + \sqrt{\frac{\nu \cdot \hat{\mathbb{V}}_{+,\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}] \cdot \log\left(\zeta_1\right)}{M}} + \frac{B \cdot \|\mathcal{E}-\mathcal{E}'\|_\infty}{C!} + B \cdot \sqrt{\nu^{3/2} \cdot \frac{\log^{3/2}\left(\zeta_1\right)}{M\sqrt{M}}}$$

$$+ B \cdot \sqrt{\nu^{3/2} \cdot \frac{\sqrt{\log\left(\zeta_2\right)} \cdot \log\left(\zeta_1\right)}{M\sqrt{N}}} + \frac{B \cdot \nu \cdot \log\left(\zeta_1\right)}{M} + B \cdot \sqrt{\nu \cdot \frac{\log(\zeta_2)}{N}},$$

where

$$C_M = 3M+4, \;\; \zeta_1 = (C_M BM/\delta)^{(1/\nu)} \cdot r, \;\; \zeta_2 = (C_M N/\delta)^{(1/\nu)} \cdot r,$$

Proof.

NOTE: Please see Tab.4 and Sec.D.4.1 for all the necessary notations not explained herein.

The Error Decomposition. The overall uniform excess risk can be decomposed into three parts:

$$\sup_{f\in\mathcal{F}}\left[\mathop{\mathbb{E}}_{\mathcal{E}'}[\ell_{\mathcal{D},\mathcal{E}'}] - \mathop{\hat{\mathbb{E}}}_{\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right] \le \underbrace{\sup_{f\in\mathcal{F}}\left[\mathop{\mathbb{E}}_{\mathcal{E}'}[\ell_{\mathcal{D},\mathcal{E}'}] - \mathop{\mathbb{E}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]\right]}_{(1)} + \underbrace{\sup_{f\in\mathcal{F}}\left[\mathop{\mathbb{E}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] - \mathop{\hat{\mathbb{E}}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]\right]}_{(2)} + \underbrace{\sup_{f\in\mathcal{F}}\left[\mathop{\hat{\mathbb{E}}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] - \mathop{\hat{\mathbb{E}}}_{\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right]}_{(3)}$$

The first part is due to the meta-distribution shift from the known distribution τ to an unknown distribution τ'; the second part reflects the random error produced from using the Monte Carlo estimation; while the third part reflects the error coming from the empirical estimation of the population-level mean loss when the Monte Carlo result is given.

Since bounding (1) and (3) are relatively simple, the derivation will be presented following the order $(1) \to (3) \to (2)$.

The Bound for (1). Since $\ell \in [0, B]$, we have:

$$\begin{aligned}\sup_{f\in\mathcal{F}}\left[\mathop{\mathbb{E}}_{\mathcal{E}'}[\ell_{\mathcal{D},\mathcal{E}'}] - \mathop{\mathbb{E}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]\right] &\le B \cdot \int_{\mathbb{S}_{c-1}} |\mathcal{E}'(\mathbb{P}) - \mathcal{E}(\mathbb{P})| d\mathbb{P} \\ &\overset{(a)}{\le} \frac{B \cdot \|\mathcal{E}'(\mathbb{P}) - \mathcal{E}(\mathbb{P})\|_\infty}{C!}\end{aligned}$$

$\mathbb{S}_{c-1}$ is the probabilistic simplex on $\mathbb{R}^c$, $\mathbb{P}$ is a c-dimensional probability allocation sampled either from $\mathcal{E}$ or $\mathcal{E}'$. In the derivation, (a) is a direct result from Lem.2.

The Bound for (3). According to (11), we have the following result:

$$\sup_{f\in\mathcal{F}}\left[\mathop{\hat{\mathbb{E}}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] - \mathop{\hat{\mathbb{E}}}_{\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right] = \sup_{f\in\mathcal{F}}\left[\mathop{\mathbb{E}}_{\mathcal{D}}\left[\bar{\ell}_{\mathcal{P}}\right] - \mathop{\hat{\mathbb{E}}}_{\mathcal{S}}\left[\bar{\ell}_{\mathcal{P}}\right]\right]$$

Under this reformulation, we then derive an upper bound by concentration over the randomness of $\bar{\ell}_{\mathcal{P}}$. By applying Lem.8, we have the following result holds with probability at least $1-\delta/(C_M)$ for all $f \in \mathcal{F}$:

$$\mathop{\hat{\mathbb{E}}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] - \mathop{\hat{\mathbb{E}}}_{\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}] \lesssim B \cdot \sqrt{\frac{\log(C_M \cdot \mathcal{N}_1/\delta)}{N}},$$

where $\mathcal{N}_1 = \mathcal{N}_\infty(\mathcal{F}, 1/4N, N)$, where we used the fact of the Lip. constant in Lem.4-5 and the basic property of expectation.

The Bound for (2). Applying Lem.7 to the Monte Carlo process, we have the following result holds uniformly with probability at least $1-\delta/(C_M)$:

$$\mathop{\mathbb{E}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] \lesssim \mathop{\hat{\mathbb{E}}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] + \sqrt{\frac{\hat{\mathbb{V}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M}} + \frac{B \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M}$$

with $\mathcal{M} = \mathcal{N}_\infty(\mathcal{F}, \frac{1}{4M}, 2M)$. We present the rest of derivation in two steps

Step 1. We first upper bound $\hat{\mathbb{V}}_\mathcal{E}[\ell_{\mathcal{D},\mathcal{E}}]$ with $\mathbb{V}_{+,\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$ and get a resulting inequality.

First we show that $\hat{\mathbb{V}}_\mathcal{E}[\ell_{\mathcal{D},\mathcal{E}}]$ is 1-Lip w.r.t ℓ_2 norm defined on the empirical measure $\mathcal{P}$:

$$\|f - g\|_{2,M} = \sqrt{\sum_{i=1}^{M} \left(f(\mathbb{P}_M^{(j)}) - g(\mathbb{P}_M^{(j)}) \right)^2}.$$

To do this, we denote $\boldsymbol{\ell} = [\ell_{\mathcal{D},\mathcal{E},1}, \cdots, \ell_{\mathcal{D},\mathcal{E},M}]$, where

$$\ell_{\mathcal{D},\mathcal{E},i} = \mathop{\mathbb{E}}_{\mathcal{D}} \left[\ell_{\mathsf{LA}}(f^{(\xi_i)}(\boldsymbol{x}), y; P_i) \right]$$

is the expected loss given a fixed test label distribution P_i. Then, one can rewrite $\hat{\mathbb{V}}_\mathcal{E}[\ell_{\mathcal{D},\mathcal{E}}]$ as:

$$\begin{aligned} \hat{\mathop{\mathbb{V}}_{\mathcal{E}}}[\ell_{\mathcal{D},\mathcal{E}}] &= \frac{1}{M} \left(\boldsymbol{\ell}^\top \boldsymbol{\ell} - \frac{1}{M} \cdot \boldsymbol{\ell}^\top 1_{M\times M} \boldsymbol{\ell} \right) \\ &= \frac{1}{M} \left\| (\boldsymbol{I} - \frac{1}{m} 1_{M\times M}) \boldsymbol{\ell} \right\|^2 \end{aligned}$$

where $\boldsymbol{\ell}^\top 1_{M\times M}$ is an M by M all-one matrix. Here we used the property:

$$(\boldsymbol{I} - \frac{1}{m} 1_{M\times M})^\top (\boldsymbol{I} - \frac{1}{m} 1_{M\times M}) = (\boldsymbol{I} - \frac{1}{m} 1_{M\times M}).$$

$$\nabla_\ell \hat{\mathop{\mathbb{V}}_{\mathcal{E}}}[\ell_{\mathcal{D},\mathcal{E}}] = \frac{2}{M} \cdot (\boldsymbol{I} - \frac{1}{m} 1_{M\times M}) \cdot \boldsymbol{\ell}$$

According to Lem.3, we need to bound:

$$\begin{aligned} \left\| \nabla_\ell \hat{\mathop{\mathbb{V}}_{\mathcal{E}}}[\ell_{\mathcal{D},\mathcal{E}}] \right\|_{2,M} &= \frac{2}{M} \cdot \left\| (\boldsymbol{I} - \frac{1}{m} 1_{M\times M}) \cdot \boldsymbol{\ell} \right\|_2 \\ &\le \frac{2}{M} \cdot \left\| (\boldsymbol{I} - \frac{1}{m} 1_{M\times M}) \right\|_{sp} \cdot \|\boldsymbol{\ell}\|_2 \\ &\le \frac{2}{M} \cdot \left(1 - \frac{1}{M} \cdot \lambda_{min} \left(1_{M\times M} \right) \right) \cdot MB \\ &= 2 \cdot B, \end{aligned}$$

where for a real square matrix $\boldsymbol{A} \in \mathbb{R}^{M\times M}$, $||\boldsymbol{A}||_{sp} = \lambda_{max}(\boldsymbol{A})$ is the spectral norm. The last equality follows that $\lambda_{min} \left(1_{M\times M} \right) = 0$ since the rank of all-one matrix is 1. According to Lem.6, we have the following inequality holds for a fixed $f \in \mathcal{F}$ with probability at least $1 - \delta/(C_M)$:

$$\hat{\mathop{\mathbb{V}}_{\mathcal{E}}}[\ell_{\mathcal{D},\mathcal{E}}] \lesssim \mathop{\mathbb{V}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] + B^2 \cdot \sqrt{\frac{\log(C_M/\delta)}{M}}.$$

To construct a uniform bound, we derive the Lip. constant for $\hat{\mathbb{V}}_{\mathcal{E}|j}[\ell_{\mathcal{D},\mathcal{E}}]$ and $\mathbb{V}_{\mathcal{E}|j}[\ell_{\mathcal{D},\mathcal{E}}]$ w.r.t the change of f in terms of the infinity norm.

First we have:

$$|\hat{\mathop{\mathbb{V}}_{\mathcal{E}}}[\ell_{\mathcal{D},\mathcal{E}}] - \hat{\mathop{\mathbb{V}}_{\mathcal{E}}}[\tilde{\ell}_{\mathcal{D},\mathcal{E}}]| \le 2 \cdot B \left\| \ell_{\mathcal{D},\mathcal{E}} - \tilde{\ell}_{\mathcal{D},\mathcal{E}} \right\|_{2,M} \le 4 \cdot B \cdot \left\| f - \tilde{f} \right\|_\infty,$$

where $\tilde{\ell}_{\mathcal{D},\mathcal{E}}$ is the loss function evaluated on $\tilde{f}$. The second inequality used the fact that $\ell_{\mathcal{D},\mathcal{E}}$ is $4B$-Lip. continuous (a direct result of Lem.4 and the basic property of the expectation).

Moreover, since $\mathbb{V}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] = (1-\frac{1}{M}) \cdot \mathbb{E}[\hat{\mathbb{V}}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]]$. Thus, we have $\mathbb{V}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$ is also $4B$-Lip continuous. Above all, following a similar proof as Lem.8, we have:

$$\hat{\underset{\mathcal{E}}{\mathbb{V}}}[\ell_{\mathcal{D},\mathcal{E}}] \lesssim \underset{\mathcal{E}}{\mathbb{V}}[\ell_{\mathcal{D},\mathcal{E}}] + B^2 \cdot \sqrt{\frac{\log(C_M \cdot \mathcal{M}_1/\delta)}{M}}$$

holds uniformly over all $f \in \mathcal{F}$ with practicality at least $1-\delta/C_M$, where $\mathcal{M}_1 = \mathcal{N}_\infty(\mathcal{F}, 1/8BM, M)$.

Putting all together, we have the following result holds uniformly with probability at least $1 - 2\cdot\delta/(C_M)$:

$$\begin{aligned}
\underset{\mathcal{E}}{\mathbb{E}}[\ell_{\mathcal{D},\mathcal{E}}] &\lesssim \hat{\underset{\mathcal{E}}{\mathbb{E}}}[\ell_{\mathcal{D},\mathcal{E}}] + \sqrt{\frac{\mathbb{V}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M}} \\
&+ B \cdot \sqrt{\frac{\log\left(C_M \cdot \mathcal{M}_1/\delta\right) \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M\sqrt{M}}} + \frac{B \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M} \\
&\asymp \hat{\underset{\mathcal{E}}{\mathbb{E}}}[\ell_{\mathcal{D},\mathcal{E}}] + \sqrt{\frac{\mathbb{V}_{+,\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}] \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M}} \\
&+ B \cdot \sqrt{\frac{\log\left(C_M \cdot \mathcal{M}_1/\delta\right) \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M\sqrt{M}}} + \frac{B \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M}.
\end{aligned} \tag{12}$$

Step 2. We now derive an upper bound for $\mathbb{V}_{+,\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$ by $\hat{\mathbb{V}}_{+,\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]$.

First, we provide an upper bound of $\mathbb{V}_{+,\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$ by $\hat{\mathbb{V}}_{+,\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$:

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \left[\underset{+,\mathcal{E}}{\mathbb{V}}[\ell_{\mathcal{D},\mathcal{E}}] - \hat{\underset{+,\mathcal{E}}{\mathbb{V}}}[\ell_{\mathcal{D},\mathcal{E}}]\right] \leq &\sum_{i=1}^{M} \frac{1}{M} \cdot \sup_{f\in\mathcal{F}} \left[\left| \left(\left(\ell_{\mathcal{D},\mathcal{E},i} - \hat{\underset{\mathcal{E}}{\mathbb{E}}}[\ell_{\mathcal{D},\mathcal{E}}]\right)_+\right)^2 - \left(\left(\ell_{\mathcal{D},\mathcal{E},i} - \underset{\mathcal{E}}{\mathbb{E}}[\ell_{\mathcal{D},\mathcal{E}}]\right)_+\right)^2 \right|\right] \\
&+ \sup_{f\in\mathcal{F}} \left[\left| \sum_{i=1}^{M} \frac{1}{M} \left(\left(\ell_{\mathcal{D},\mathcal{E},i} - \underset{\mathcal{E}}{\mathbb{E}}[\ell_{\mathcal{D},\mathcal{E}}]\right)_+\right)^2 - \underset{\tau}{\mathbb{E}}\left[\left(\left(\ell_{\mathcal{D},\mathcal{E},i} - \underset{\mathcal{E}}{\mathbb{E}}[\ell_{\mathcal{D},\mathcal{E}}]\right)_+\right)^2\right] \right|\right]
\end{aligned} \tag{13}$$

According to Lem.9, we have the following holds with probability at least $1-\delta/(C_M)$:

$$\sup_{f\in\mathcal{F}} \left[\left| \left(\left(\ell_{\mathcal{D},\mathcal{E},i} - \hat{\underset{\mathcal{E}}{\mathbb{E}}}[\ell_{\mathcal{D},\mathcal{E}}]\right)_+\right)^2 - \left(\left(\ell_{\mathcal{D},\mathcal{E},i} - \underset{\mathcal{E}}{\mathbb{E}}[\ell_{\mathcal{D},\mathcal{E}}]\right)_+\right)^2 \right|\right] \lesssim B^2 \cdot \sqrt{\frac{\log(C_M \cdot \mathcal{N}_\infty(\mathcal{F}, 1/4M, M)/\delta)}{M}}.$$

Here we used the fact that $\left((x-y)_+\right)^2$ is $2 \cdot B$-Lip w.r.t y if $y \in [0, B]$.

According to Lem.8, we have holds with probability at least $1-\delta/(C_M)$:

$$\sup_{f\in\mathcal{F}} \left[\left| \sum_{i=1}^{M} \frac{1}{M} \left(\left(\ell_{\mathcal{D},\mathcal{E},i} - \underset{\mathcal{E}}{\mathbb{E}}[\ell_{\mathcal{D},\mathcal{E}}]\right)_+\right)^2 - \underset{\mathcal{E}}{\mathbb{E}}\left[\left(\left(\ell_{\mathcal{D},\mathcal{E},i} - \underset{\mathcal{E}}{\mathbb{E}}[\ell_{\mathcal{D},\mathcal{E}}]\right)_+\right)^2\right] \right|\right] \lesssim B \cdot \sqrt{\frac{\log(C_M \cdot \mathcal{M}_2/\delta)}{M}},$$

where $\mathcal{M}_2 = \mathcal{N}_\infty(\mathcal{F}, 1/16BM, M)$ Here we used the fact that $\left(\left(\ell_{\mathcal{D},\mathcal{E},i} - \mathbb{E}_{\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]\right)_+\right)^2$ is $8 \cdot B$-Lip w.r.t to infinity norm in $\mathcal{F}$.

Putting all together with a union bound, we have with probability at least $1-\delta/(C_M)$:

$$\sup_{f\in\mathcal{F}} \left[\underset{+,\mathcal{E}}{\mathbb{V}}[\ell_{\mathcal{D},\mathcal{E}}] - \hat{\underset{+,\mathcal{E}}{\mathbb{V}}}[\ell_{\mathcal{D},\mathcal{E}}]\right] \lesssim B^2 \cdot \sqrt{\frac{\log(C_M \cdot \mathcal{M}_2/\delta)}{M}} + B \cdot \sqrt{\frac{\log(C_M \cdot \mathcal{M}_2/\delta)}{M}}$$

Next, we provide an upper bound of $\hat{\mathbb{V}}_{+,\mathcal{E}}[\ell_{\mathcal{D},\mathcal{E}}]$ by $\hat{\mathbb{V}}_{+,\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]$

$$\sup_{f\in\mathcal{F}}\left[\underset{+,\mathcal{E}}{\hat{\mathbb{V}}}[\ell_{\mathcal{D},\mathcal{E}}]-\underset{+,\mathcal{E}}{\hat{\mathbb{V}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right]\leq\sum_{i=1}^{M}\frac{1}{M}\cdot\Bigg(\underbrace{\sup_{f\in\mathcal{F}}\left[\left|\left(\left(\ell_{\mathcal{D},\mathcal{E},i}-\underset{\mathcal{E}}{\hat{\mathbb{E}}}[\ell_{\mathcal{D},\mathcal{E}}]\right)_+\right)^2-\left(\left(\ell_{\mathcal{D},\mathcal{E},i}-\underset{\mathcal{E}}{\hat{\mathbb{E}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right)_+\right)^2\right|\right]}_{(a)}$$
$$+\underbrace{\sup_{f\in\mathcal{F}}\left[\left|\left(\left(\ell_{\mathcal{D},\mathcal{E},i}-\underset{\mathcal{E}}{\hat{\mathbb{E}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right)_+\right)^2-\left(\left(\ell_{\mathcal{S},\mathcal{E},i}-\underset{\mathcal{E}}{\hat{\mathbb{E}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right)_+\right)^2\right|\right]}_{(b)}\Bigg)$$

For (a), we have the following results hold with probability at least $1-\delta/(C_M)$:

$$\begin{aligned}&\sup_{f\in\mathcal{F}}\left[\left|\left(\left(\ell_{\mathcal{D},\mathcal{E},i}-\underset{\mathcal{E}}{\hat{\mathbb{E}}}[\ell_{\mathcal{D},\mathcal{E}}]\right)_+\right)^2-\left(\left(\ell_{\mathcal{D},\mathcal{E},i}-\underset{\mathcal{E}}{\hat{\mathbb{E}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right)_+\right)^2\right|\right]\\&\leq 4B\cdot\sup_{f\in\mathcal{F}}\left[\left|\underset{\mathcal{E}}{\hat{\mathbb{E}}}[\ell_{\mathcal{D},\mathcal{E}}]-\underset{\mathcal{E}}{\hat{\mathbb{E}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right|\right]\\&\\&=4B\cdot\sup_{f\in\mathcal{F}}\left[\left|\underset{\mathcal{D}}{\mathbb{E}}\left[\bar{\ell}_{\mathcal{P}}\right]-\underset{\mathcal{S}}{\hat{\mathbb{E}}}\left[\bar{\ell}_{\mathcal{P}}\right]\right|\right]\\&\lesssim B^2\cdot\sqrt{\frac{\log(C_M\cdot\mathcal{N}_1/\delta)}{N}}\end{aligned}$$

For (b), we have the following results hold with probability at least $1-\delta/(C_M)$:

$$\begin{aligned}&\sup_{f\in\mathcal{F}}\left[\left|\left(\left(\ell_{\mathcal{D},\mathcal{E},i}-\underset{\mathcal{E}}{\hat{\mathbb{E}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right)_+\right)^2-\left(\left(\ell_{\mathcal{S},\mathcal{E},i}-\underset{\mathcal{E}}{\hat{\mathbb{E}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right)_+\right)^2\right|\right]\\&\leq 4B\cdot\sup_{f\in\mathcal{F}}\left[\left|\ell_{\mathcal{D},\mathcal{E},i}-\ell_{\mathcal{S},\mathcal{E},i}\right|\right]\\&\lesssim B^2\cdot\sqrt{\frac{\log(C_M\cdot\mathcal{N}_1/\delta)}{N}}\end{aligned}$$

Combining (a) and (b) with a union boubsnd, we have the following result holds with probability at least $1-2M\cdot\delta/(C_M)$

$$\sup_{f\in\mathcal{F}}\left[\underset{+,\mathcal{E}}{\hat{\mathbb{V}}}[\ell_{\mathcal{D},\mathcal{E}}]-\underset{+,\mathcal{E}}{\hat{\mathbb{V}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right]\lesssim B^2\cdot\sqrt{\frac{\log(C_M\cdot\mathcal{N}_1/\delta)}{N}}$$

all together, we have the following bound holds with probability at least $1-(3M+1)\cdot\delta/(C_M)$ for (13)

$$\sup_{f\in\mathcal{F}}\left[\underset{+,\mathcal{E}}{\mathbb{V}}[\ell_{\mathcal{D},\mathcal{E}}]-\underset{+,\mathcal{E}}{\hat{\mathbb{V}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\right]\lesssim B^2\cdot\sqrt{\frac{\log(C_M\cdot\mathcal{N}_1/\delta)}{N}}+B\cdot\sqrt{\frac{\log(C_M\cdot\mathcal{M}_2/\delta)}{M}}\tag{14}$$

We finish the bound for (2) by combining (12) and (14), which suggests that the following result holds with probability at least $1-(3M+3)\cdot\delta/(C_M)$

$$\underset{\mathcal{E}}{\mathbb{E}}[\ell_{\mathcal{D},\mathcal{E}}]\lesssim\underset{\mathcal{E}}{\hat{\mathbb{E}}}[\ell_{\mathcal{D},\mathcal{E}}]+\sqrt{\frac{\hat{\mathbb{V}}_{+,\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}]\cdot\log\left(C_M\cdot\mathcal{M}/\delta\right)}{M}}+B\cdot\sqrt{\frac{\log\left(C_M\cdot\mathcal{M}_2/\delta\right)\cdot\log\left(C_M\cdot\mathcal{M}/\delta\right)}{M\sqrt{M}}}$$
$$+B\cdot\sqrt{\frac{\log\left(C_M\cdot\mathcal{N}_1/\delta\right)\cdot\log\left(C_M\cdot\mathcal{M}/\delta\right)}{M\sqrt{N}}}+\frac{B\cdot\log\left(C_M\cdot\mathcal{M}/\delta\right)}{M}.$$

Final Step. Finally, by combining the bound for (1),(2),(3), we reach that the following result holds with probability at least $1-\delta$ uniformly for all $f \in \mathcal{F}$:

$$\begin{aligned}\underset{\mathcal{E}}{\mathbb{E}}[\ell_{\mathcal{D},\mathcal{E}}] \lesssim \underset{\mathcal{E}}{\hat{\mathbb{E}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}] &+ \frac{B \cdot \|\mathcal{E}-\mathcal{E}'\|_\infty}{C!} + \sqrt{\frac{\hat{\mathbb{V}}_{+,\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}] \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M}} + B \cdot \sqrt{\frac{\sqrt{\log\left(C_M \cdot \mathcal{M}_2/\delta\right)} \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M\sqrt{M}}} \\ &+ B \cdot \sqrt{\frac{\sqrt{\log\left(C_M \cdot \mathcal{N}_1/\delta\right)} \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M\sqrt{N}}} + \frac{B \cdot \log\left(C_M \cdot \mathcal{M}/\delta\right)}{M} + B \cdot \sqrt{\frac{\log(C_M \cdot \mathcal{N}_1/\delta)}{N}},\end{aligned}$$

Under the Assumption that

$$\mathcal{N}_\infty(\mathcal{F}, \epsilon, M) \le \left(\frac{r}{\epsilon}\right)^\nu,$$

we have the following result holds with probability at least $1-\delta$ uniformly over all $f \in \mathcal{F}$:

$$\begin{aligned}\underset{\mathcal{E}}{\mathbb{E}}[\ell_{\mathcal{D},\mathcal{E}}] \lesssim \underset{\mathcal{E}}{\hat{\mathbb{E}}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}] &+ \frac{B \cdot \|\mathcal{E}-\mathcal{E}'\|_\infty}{C!} + \sqrt{\frac{\nu \cdot \hat{\mathbb{V}}_{+,\mathcal{E}}[\hat{\ell}_{\mathcal{S},\mathcal{E}}] \cdot \log\left(\zeta_1\right)}{M}} + B \cdot \sqrt{\nu^{3/2} \cdot \frac{\sqrt{\log\left(\zeta_1\right)} \cdot \log\left(\zeta_1\right)}{M\sqrt{M}}} \\ &+ B \cdot \sqrt{\nu^{3/2} \cdot \frac{\sqrt{\log\left(\zeta_2\right)} \cdot \log\left(\zeta_1\right)}{M\sqrt{N}}} + \frac{B \cdot \nu \cdot \log\left(\zeta_1\right)}{M} + B \cdot \sqrt{\nu \cdot \frac{\log(\zeta_2)}{N}},\end{aligned}$$

where $\zeta_1 = (C_M BM/\delta)^{(1/\nu)} \cdot r,\ \zeta_2 = (C_M N/\delta)^{(1/\nu)} \cdot r.$ □

E. Additional Experiment Settings

E.1. Implentation Details.

Following (Zhang et al., 2022; Aimar et al., 2023), we employ ResNeXt-50 (Xie et al., 2017) and ResNet-32 (He et al., 2016) as the backbones for ImageNet-LT and CIFAR-LT datasets, respectively. In CIFAR-LT experiments, we train the model for 200 epochs using Stochastic Gradient Descent (SGD) (Robbins & Monro, 1951). The initial learning rate is set at 0.1, with 0.9 momentum rate and 128 batch size. Moreover, a step learning rate schedule is adopted, which reduces the learning rate by a factor of 10 at the 160-th and 180-th epoch, respectively. Regarding the ImageNet-LT dataset, the model is trained 180 epochs using SGD. Here, the initial learning rate is 0.025 with 0.9 momentum and 64 batch size. Then, the learning rate is adjusted through a cosine annealing schedule, which gradually declines from 0.025 to 0 over 180 epochs. Finally, for the sake of fair comparisons, we re-implement the above methods using their publicly available code and conduct experiments on the same device.

E.2. Datasets

Dataset Descriptions. We conduct experiments on three popular benchmark datasets for imbalanced learning: **(a) CIFAR-10-LT and CIFAR-100-LT** datasets. The original CIFAR-10 and CIFAR-100 datasets (Krizhevsky et al., 2009) have $50,000$ images for training and $10,000$ images for validation with 10 and 100 categories, respectively. Following (Cui et al., 2019), we use the long-tailed version of CIFAR 10 and CIFAR 100 datasets with imbalanced ratio $\rho = N_{\max}/N_{\min} = 100$. **(b) ImageNet-LT** dataset. We adopt the ImageNet-LT dataset proposed by (Liu et al., 2019), which is sampled from ImageNet (Deng et al., 2009) following the *Pareto* distribution. Briefly, it consists of $115.8K$ images from 1000 classes, with 1280 images in the most frequent class and 5 images in the minority.

E.3. Competitors

- **Label-Distribution-Aware Margin (LDAM)** (Cao et al., 2019) improves the performance of minority classes by encouraging larger margins for minority classes.

- **Logit Adjustment (LA)** (Menon et al., 2021) advances the conventional softmax cross-entropy by ensuring Fisher consistency in minimizing the balanced error.

- **Vector Scaling (VS)** (Kini et al., 2021b) proposes to leverage both multiplicative and additive logit adjustments to address label imbalance problems.
- **LAbel distribution DisEntangling (LADE)** (Hong et al., 2021) X regards the long-tailed learning as a label shift problem and aims to disentangle the source label distribution from the model prediction to match the target label distribution during training.
- **Data Dependent Contraction (DDC)** (Wang et al., 2023) designs a Deferred Re-Weighting scheme to boost performance for imbalanced learning, which is also compatible with the VS loss.
- **RoutIng Diverse Experts (RIDE)** (Wang et al., 2021) proposes a distribution-aware multiple expert routing system, which can efficiently reduce both model bias and variance for imbalanced learning.
- **Self-supervised Aggregation of Diverse Experts (SADE)** (Zhang et al., 2022) is a state-of-the-art test-agnostic long-tailed learning method, which trains multiple skill-diverse experts to tackle different class distributions and adopts self-supervision to aggregate the decisions of multiple experts to adapt unknown test class distributions.
- **Balanced Product of Experts (BalPoE)** (Aimar et al., 2023) is another state-of-the-art long-tailed learning algorithm that successfully extends logit adjustment to the mixture of experts.

E.4. The Choice of meta-distribution and Experts

We briefly introduce the choice of (3) and (4) to define the mixture distribution. Detailed implementations are shown in the appendix. Drawing inspiration from the skill-diverse expert learning approach in prior art, we employ a three-component mixture model for (3) to encapsulate three critical skills. Each component is determined by a specific choice of α, the parameter for Dirichlet distribution. The forward component $\boldsymbol{\alpha}^{(f)}$ aligns with the training label distribution, indicative of performance in the head classes. The uniform component $\boldsymbol{\alpha}^{(u)}$ corresponds to a uniform distribution, reflecting adherence to the conventional long-tail testing protocol. The backward component represents an inverse long-tail distribution of the training set (where head classes are transformed into tail classes and *vice versa*), signifying performance in the tail distribution. Furthermore, for (4), we chose a uniform distribution ($p_1 = p_2 = p_3 = 1/3$) in our model, considering the equal significance of all three skills. This prevents the unfair oversight of any particular skill.

E.5. Construction of Training meta-distribution $\mathcal{E}$

Consider the training class distribution represented by $P_1, P_2, ..., P_C$, where C denotes the number of classes. Without loss of generality, **we assume that the classes are sorted in a sense**: $P_1 \geq P_2 \geq ... \geq P_C$. Otherwise, we can change the numbering and obtain the same result.

The training meta-distribution $\mathcal{E}$ is a Dirichlet mixture distribution as Sec.E.4. It consists of three Dirichlet distribution components: the forward component $\alpha^{(f)}$, the uniform component $\alpha^{(u)}$, and the backward component $\alpha^{(b)}$.

The forward component parameter $\alpha^{(f)}$ is set element-wisely:

$$\alpha_i^{(f)} = S \cdot P_i,\ i = 1, 2, \ldots, C$$

where S is a predefined normalization factor to control the variance of the component. **Here we set it as** $S = 10000$ **for all datasets.** Since the mean of the Dirichlet distribution is exactly $\alpha^{(f)}$, this component represents local variations concentrated around the long-tail distribution aligned with the training data.

The backward component $\alpha^{(b)}$ is set element-wisely as:

$$\alpha_i^{(b)} = S \cdot P_{(C-i)}$$

where S is a predefined normalization factor to control the variance of the component. **Here we set it as** $S = 10000$ **for all datasets.** By connecting $\alpha_i^{(b)}$ with $P_{(C-i)}$, the long-tail distribution is reversed, where the head classes become tail classes, and *vice versa*. Since the mean of the Dirichlet distribution is exactly $\alpha^{(b)}$, this component represents local variations concentrated around the **inverse** long-tail distribution aligned with the training data.

The uniform component $\alpha^{(u)}$ is set element-wisely as:

$$\alpha_i^{(u)} = \frac{S}{C},\ i = 1, 2, \ldots, C$$

where S is a predefined normalization factor to control the variance of the component. **Here we set it as $S = 10000$ for all datasets.** By connecting $\alpha_i^{(b)}$ with $1/C$, it recovers a uniform distribution. Since the mean of the Dirichlet distribution is exactly $\alpha^{(u)}$, this component represents local variations concentrated around the uniform distribution.

The corresponding p.d.f are expressed as: $\mathsf{Dir}^{(f)}, \mathsf{Dir}^{(u)}, \mathsf{Dir}^{(b)}$. Moreover, the three components are mixed with a uniform distribution. Above all, the p.d.f for the mixture distribution becomes:

$$\mathcal{E} = \frac{1}{3} \cdot \mathsf{Dir}^{(f)} + \frac{1}{3} \cdot \mathsf{Dir}^{(u)} + \frac{1}{3} \cdot \mathsf{Dir}^{(b)}$$

E.6. Sampling Procedure of $\mathbb{P}_{te}$ for the Monte Carlo Approximation

Initially, we employ a Monte Carlo method to sample a set of $\mathbb{P}_{te}$ from the training meta-distribution $\mathcal{E}$ and obtain the generated data $\mathcal{P} = \{\mathbb{P}_j, \xi_j\}_{j=1}^{M}$ for Monte Carlo approximation.

We implement the sampling process by repeating the following two-step procedure for M times.

1. Randomly sampling a Dirichlet distribution component α from $\{\alpha^{(f)}, \alpha^{(u)}, \alpha^{(b)}\}$ with equal probability.
2. Sampling a distribution $\mathbb{P}_{te}$ from the Dirichlet distribution component α.

E.7. Mini-Batch Construction

Denote the number of mini-batches as B. We sample $60 \cdot B$ label distributions to construct the dataset $\mathcal{P}$. We randomly sample 60 label distributions for each mini-batch without replacement in each epoch to get an unbiased estimation.

E.8. Construction of Testing Datasets

We employ two settings, SADE's Setting and our setting, to construct our test data. The details are discussed as follows:

E.8.1. SADE's setting

The test data in this setting directly follows SADE (Zhang et al., 2022). The only difference here is that we include more imbalance ratios, denoted as ρ, in our experiments. For CIFAR 100-LT and CIFAR 10-LT, $\rho \in \{2, 5, 10, 25, 50, 100\}$. For ImageNet-LT, $\rho \in \{2, 5, 10, 25, 50\}$.

E.8.2. Ours setting

According to our meta-distribution, we employ three kinds of Dirichlet components to generate test distributions: the forward Dirichlet distribution $\alpha^{(f_{test})}$, the uniform Dirichlet distribution $\alpha^{(u_{test})}$ and the backward Dirichlet distribution $\alpha^{(b_{test})}$.

From each component of the meta-distribution, we sample three specific distributions as the test class distribution, resulting in a total of **9 test class distributions**.

The Dirichlet distributions are chosen based on a predefined imbalance ratio ρ. The greater the ρ, the more challenging the test distribution. **For CIFAR 100-L and CIFAR 10-LT, we set ρ to be 100, while for ImageNet-LT, we use $\rho = 50$.**

For the forward Dirichlet distribution $\alpha^{(f_{test})}$, we set:

$$\alpha_i^{(f_{test})} = \rho^{-(i-1)/(C-1)},\ i = 1, 2, \ldots, C$$

For the backward Dirichlet distribution $\alpha^{(b_{test})}$, we set:

$$\alpha_i^{(b_{test})} = \rho^{-(C-i)/(C-1)},\ i = 1, 2, \ldots, C$$

For the uniform Dirichlet distribution $\alpha^{(u_{test})}$, we set:

$$\alpha_i^{(u_{test})} = 1/C,\ i = 1, 2, \dots, C$$

Subsequent steps remain consistent across the three components. For simplicity, we denote each Dirichlet distribution as α.

To enhance the simulation of the randomness in the test distribution, we introduce a perturbation to α, allowing up to 5% variations. This involves adjusting the i-th element α_i:

$$\alpha_i = \alpha_i + \alpha_i * \epsilon,\ \epsilon \sim U(-0.05, 0.05)$$

Following this, we normalize α to ensure the sum of its components equals S. The role of S is identical to its function in constructing the Training meta-distribution. We set $S = 100$ for CIFAR 100-LT, $S = 1000$ for CIFAR 10-LT and $S = 10000$ for ImageNet-LT.

F. Additional Experiments

F.1. Overall Performance on the SADE's Settings

The Overall Performance on ImageNet-LT and CIFAR-10-LT are shown in Tab.6-Tab.7.

Table 5. Performance Comparison on CIFAR-100-LT **(SADE's Setting)**

Method	Forward-LT						Uni.	Backward-LT						Mean
	100	50	25	10	5	2	1	2	5	10	25	50	100	
LDAM	66.03	63.03	59.98	55.37	51.01	45.02	39.99	35.04	28.75	24.33	19.49	16.49	13.66	39.86(±17.66)
LA	60.68	58.89	56.90	54.08	51.74	48.45	45.40	42.30	39.44	37.23	33.69	31.98	30.22	45.46(±10.12)
VS	58.80	57.16	55.38	52.55	49.66	45.74	42.84	39.46	35.04	32.33	28.52	26.43	24.78	42.21(±11.58)
LADE	59.22	57.84	55.55	52.40	49.70	46.82	44.47	41.53	38.46	36.46	33.49	31.98	30.41	44.49(± 9.60)
DDC	59.36	58.33	56.60	53.95	51.48	49.01	46.98	44.01	41.06	39.22	36.20	34.55	33.13	46.45(± 8.82)
RIDE	64.57	62.91	61.06	58.02	55.33	51.67	48.40	44.66	40.43	37.54	34.10	32.46	30.41	47.81(±11.62)
SADE	67.81	65.45	62.75	58.69	56.04	51.91	**49.53**	45.90	44.04	43.34	42.32	42.48	42.75	51.77(± 9.02)
BalPoE	**69.22**	**67.02**	**64.45**	**60.19**	**57.26**	52.05	48.66	44.54	40.57	37.13	33.25	31.58	29.24	48.86(±13.44)
DirMixE	68.32	66.21	63.09	59.49	56.35	**52.62**	48.38	**46.40**	**45.05**	**44.79**	**43.71**	**44.41**	**44.25**	**52.54**(± 8.74)

Table 6. ImageNet-LT (**SADE's Setting**)

Method	Forward-LT					Uni.	Backward-LT					Mean
	50	25	10	5	2	1	2	5	10	25	50	
LDAM	63.18	61.31	58.10	55.29	51.28	47.92	44.63	40.19	37.11	33.80	31.17	47.63(±10.65)
LA	60.57	59.83	57.63	55.64	52.92	50.66	48.39	45.43	43.35	41.08	39.07	50.42(± 7.21)
VS	60.27	59.33	57.77	56.05	53.76	52.20	50.42	48.36	46.64	44.97	43.55	52.12(± 5.56)
LADE	63.92	62.41	60.17	57.91	54.98	52.71	50.29	47.10	45.19	42.89	40.55	52.56(± 7.66)
DDC	59.49	58.37	57.02	55.16	53.15	51.32	49.45	47.39	45.53	44.10	42.46	51.22(± 5.64)
RIDE	64.98	63.73	62.20	60.15	57.09	54.98	52.53	49.59	47.75	44.86	42.88	54.61(± 7.35)
SADE	69.93	68.19	**66.02**	63.53	**60.94**	**59.00**	**57.53**	**55.91**	54.60	53.58	53.15	60.22(± 5.69)
BalPoE	69.66	68.28	65.87	**63.77**	60.91	58.95	57.00	55.22	53.85	52.59	51.88	59.82(± 6.05)
DirMixE	**70.09**	**68.46**	65.93	63.22	60.50	58.61	57.27	55.27	**55.04**	**55.38**	**55.33**	**60.46**(± 5.36)

F.2. Results on iNaturalist

In this subsection, we show the results on the `iNaturalist` dataset. First, we clarify the setting we adopt to carry the experiments.

Training Meta-Distribution $\mathcal{E}$ Construction: For `iNaturalist2018`, we construct the training meta-distribution $\mathcal{E}$ similarly to CIFAR-100-LT, CIFAR-10-LT, and ImageNet-LT. The difference is in the normalization factor S, the sum of the Dirichlet distribution hyperparameter α, which we employ to control component variance. Given `iNaturalist2018`'s larger class size, we set S to $100,000$.

Table 7. CIFAR-10-LT (**SADE's Setting**)

Method	Forward-LT						Uni.	Backward-LT						Mean
	100	50	25	10	5	2	1	2	5	10	25	50	100	
LDAM	90.56	89.12	87.18	84.28	81.58	77.60	74.49	71.41	68.06	65.35	62.40	60.86	60.09	74.84$_{(\pm 10.60)}$
LA	89.79	88.55	87.18	85.50	83.70	81.30	79.17	77.59	76.13	74.93	73.52	73.38	72.76	80.27$_{(\pm 5.89)}$
VS	84.26	83.97	83.49	82.52	81.66	80.78	80.59	80.11	80.73	80.66	80.30	80.75	81.23	81.62$_{(\pm 1.39)}$
LADE	87.45	86.73	85.41	83.91	82.82	81.14	79.77	79.02	78.21	77.64	76.74	76.74	76.55	80.93$_{(\pm 3.78)}$
DDC	86.72	86.23	85.35	84.30	83.63	82.49	81.64	81.31	80.61	80.12	79.31	79.32	79.38	82.34$_{(\pm 2.56)}$
RIDE	87.01	86.26	85.32	84.60	83.94	82.57	81.80	81.74	81.64	81.44	80.92	80.89	81.28	83.03$_{(\pm 2.05)}$
SADE	90.15	89.19	87.95	86.24	84.83	83.51	83.10	83.55	84.37	85.01	**86.62**	**87.91**	**89.10**	86.27$_{(\pm 2.31)}$
BalPoE	**92.13**	**91.31**	**89.97**	**87.93**	**86.01**	**83.92**	81.70	80.57	79.94	80.14	78.10	77.89	77.80	83.65$_{(\pm 5.05)}$
DirMixE	90.92	90.16	89.04	87.10	85.83	83.66	**83.26**	**84.16**	**85.16**	**86.17**	**86.62**	87.55	88.30	**86.76**$_{(\pm 2.31)}$

Construction of Testing Datasets: The test dataset is constructed from the following procedure.

1. **SADE's Setting:** We directly apply setting in SADE's paper [a] for 'iNaturalist2018' test data, maintaining imbalance ratios $\rho \in$
   $2, 3$.

2. **Ours' Setting**: As with CIFAR datasets and ImageNet-LT, we use three Dirichlet components for test distributions: $\alpha^{(f_{test})}$, $\alpha^{(u_{test})}$, and $\alpha^{(b_{test})}$. We sample three distributions from each meta-distribution component, creating 9 test class distributions in total, with a predefined imbalance ratio $\rho = 3$. The approach for `iNaturalist2018` mirrors that for the CIFAR and ImageNet datasets. The main adjustment is the normalization factor S to $100,000$ to accommodate `iNaturalist2018`'s larger class dimension.

The results are shown in Tab.8, Tab.9.

Table 8. iNaturalist (Ours Setting)

Method	Forward-LT			Uniform			Backward			Mean
	1	2	3	1	2	3	1	2	3	
LDAM	61.33	61.36	60.31	61.58	63.08	63.59	64.10	62.14	63.42	62.32$_{(\pm 1.14)}$
LA	61.88	61.46	60.19	61.61	62.04	62.09	62.33	60.89	61.84	61.59$_{(\pm 0.60)}$
VS	58.58	58.00	57.80	60.72	61.46	62.20	62.79	62.25	62.56	60.71$_{(\pm 1.82)}$
LADE	64.81	63.56	63.31	64.11	64.13	65.07	65.50	64.06	64.93	64.39$_{(\pm 0.65)}$
DDC	58.58	58.46	56.80	61.31	61.40	62.56	64.81	64.00	64.33	61.36$_{(\pm 2.57)}$
RIDE	67.33	68.28	66.96	68.76	68.39	69.27	69.05	68.72	69.04	68.42$_{(\pm 0.71)}$
SADE	68.94	70.13	69.52	68.64	68.92	69.82	69.95	69.44	70.73	69.57$_{(\pm 0.60)}$
BalPoE	69.46	69.03	66.99	68.70	69.73	70.22	**71.30**	71.75	71.84	69.89$_{(\pm 1.42)}$
DirMixE	**69.75**	**70.49**	**69.88**	**69.13**	**70.00**	**70.34**	71.24	**71.91**	**72.02**	**70.53**$_{(\pm 0.89)}$

Table 9. iNaturalist (SADE's Setting)

Method	Foward-LT		Uni.	Backward-LT		Mean
	3	2	1	2	3	
LDAM	65.95	66.21	66.66	67.45	67.23	66.70$_{(\pm 0.52)}$
LA	66.17	66.31	66.28	66.27	66.26	66.26$_{(\pm 0.04)}$
VS	64.06	64.71	65.34	66.10	66.05	65.25$_{(\pm 0.72)}$
LADE	69.53	69.80	70.06	70.41	70.36	70.03$_{(\pm 0.30)}$
DDC	64.31	64.92	65.93	66.86	66.90	65.78$_{(\pm 0.94)}$
RIDE	71.10	71.52	71.59	72.17	71.69	71.61$_{(\pm 0.31)}$
SADE	71.95	72.58	72.70	73.16	73.27	72.73$_{(\pm 0.43)}$
BalPoE	72.23	**73.00**	**73.31**	**73.99**	73.55	73.22$_{(\pm 0.54)}$
DirMixE	**72.53**	72.88	73.21	73.66	**73.94**	**73.24**$_{(\pm 0.47)}$

F.3. Ablation Study on the Effect of Semi-Variance

We show the performance with and without employing the semi-variance regularization term in Tab.10-Tab.13.The results consistently show that the proposed regularization induces a better average performance on all the datasets.

Table 10. Ablation of Semi-Variance on CIFAR-100 **(Ours Setting)**

Semi-Var	**Uniform**			**Forward-LT**			**Backward-LT**			**Mean**
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	
w/o	47.31	48.76	44.67	66.41	65.80	69.33	43.20	45.47	42.89	52.65
w	47.99	49.41	44.21	66.85	66.40	69.44	44.41	47.01	44.35	53.34

Table 11. Ablation of Semi-Variance on CIFAR-100 **(SADE Setting)**

Semi-Var	**Forward-LT**						**Uni.**	**Backward-LT**						**Mean**
	100	**50**	**25**	**10**	**5**	**2**	**1**	**2**	**5**	**10**	**25**	**50**	**100**	
w/o	68.18	65.73	62.69	58.85	55.60	52.11	48.85	46.89	45.13	43.60	42.96	43.00	43.13	52.06
w	68.32	66.21	63.09	59.49	56.35	52.62	48.38	46.40	45.05	44.79	43.71	44.41	44.25	52.54

Table 12. Ablation of Semi-Variance on CIFAR-10 **(Ours Setting)**

Semi-Var	**Uniform**			**Forward-LT**			**Backward-LT**			**Mean**
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	
w/o	82.87	82.99	83.36	89.86	89.58	90.80	90.04	89.15	88.96	87.51
w	83.24	82.98	83.71	90.46	89.90	91.30	89.39	88.78	88.40	87.57

F.4. Experts Assignment

In this part, we validate the ability of the test-time self-supervised aggregation by visualizing the weight assignments for different label distributions in Fig.6-8. The forward and backward experts always tend to have a significant weight for their corresponding distributions. Uniform distributions tend to utilize all three experts. This is because tail and head classes are equally crucial for uniform distribution.

F.5. Fine-grained Performance

In addition to the overall accuracy of test datasets, we also examine the effectiveness of DirMixE across many-shot, medium-shot, and few-shot classes. For CIFAR 100-LT and ImageNet-LT, classes are divided into many-shot (> 100), medium-shot ($20 \sim 100$), and few-shot (< 20) categories. In the case of CIFAR 10-LT, classes are split into many-shot (> 1000), medium-shot ($200 \sim 1000$), and few-shot (< 200) categories. All these statistics are based on the training label distribution. The results are illustrated in heat maps in Fig.10, Fig.9 for CIFAR-10 and 100, respectively.

F.6. Comparison with Stronger Baselines

In this subsection, we further compare our method with two stronger basline: GPACO (Cui et al., 2023), and BCL (Zhu et al.). The results are shown in Tab.14-17.

F.7. The Effect of the Normalization Factor S

As shown in Tab.18-Tab.21, our investigation into S, the sum of the Dirichlet distribution hyperparameter α, shows performance improves with higher S values before a slight decline. A larger S reduces the randomness in the sampled test-label distributions. When S is too small, local variation overshadows global trends, leading to misassigned experts and

Table 13. Ablation of Semi-Variance on CIFAR-10 **(SADE Setting)**

Semi-Var	**Forward-LT**						**Uni.**	**Backward-LT**						**Mean**
	100	**50**	**25**	**10**	**5**	**2**	**1**	**2**	**5**	**10**	**25**	**50**	**100**	
w/o	90.72	90.02	88.91	87.07	85.65	84.01	83.36	83.80	84.53	84.82	85.48	86.37	86.84	86.28
w	90.92	90.16	89.04	87.10	85.83	83.66	83.26	84.16	85.16	86.17	86.62	87.55	88.30	86.76

Table 14. Performance Comparison on CIFAR-100-LT (SADE's Setting)

Method	Foward-LT						Uni.	Backward-LT						Mean
	100	50	25	10	5	2	1	2	5	10	25	50	100	
GPaCo-ResNet32	64.01	62.59	61.06	58.49	56.55	53.71	51.53	48.8	45.36	43.19	40.02	38.86	37.07	50.86$_{(\pm\ 9.02)}$
BCL-ResNet32	65.37	63.56	61.43	58.26	56.00	52.43	49.82	46.43	42.19	39.37	35.35	33.43	31.30	48.84$_{(\pm 11.33)}$
Ours-ResNet32	68.32	66.21	63.09	59.49	56.35	53.63	48.38	46.40	45.05	44.79	43.71	44.41	44.25	52.62$_{(\pm\ 8.75)}$

Table 15. Performance Comparison on CIFAR-100-LT (Ours Setting)

Method	Foward-LT			Uniform			Backward-LT			Mean
	1	2	3	1	2	3	1	2	3	
GPaCo-ResNet32	62.89	60.31	64.50	50.18	48.86	49.58	37.15	37.09	34.94	49.50$_{(\pm 10.75)}$
BCL-ResNet32	64.56	62.85	66.49	47.37	47.51	48.27	32.59	30.17	32.74	48.06$_{(\pm 13.44)}$
Ours-ResNet32	66.85	66.40	69.44	47.99	49.41	44.21	44.41	47.01	44.35	53.34$_{(\pm 10.22)}$

Table 16. Performance Comparison on ImageNet-LT (SADE's Setting)

Method	Foward-TL					Uni.	Backward-LT					Mean
	50	25	10	5	2	1	2	5	10	25	50	
GPaCo-ResNeXt50	65.68	64.67	63.91	62.31	60.27	58.69	56.86	54.3	52.56	50.49	48.37	58.01$_{(\pm 5.69)}$
BCL-ResNeXt50	67.44	66.41	64.33	62.29	59.49	57.25	54.83	51.63	49.39	47.2	44.6	56.81$_{(\pm 7.55)}$
Ours-ResNeXt50	70.09	68.48	65.93	63.22	60.5	58.61	57.27	55.27	55.04	55.38	55.33	60.47$_{(\pm 5.37)}$

Table 17. Performance Comparison on ImageNet-LT (Ours Setting)

Method	Foward-LT			Uniform			Backward-LT			Mean
	1	2	3	1	2	3	1	2	3	
GPaCo-ResNeXt50	65.74	64.69	64.92	59.52	58.83	59.06	49.43	50.04	49.08	57.92$_{(\pm 6.44)}$
BCL-ResNeXt50	66.75	66.69	66.31	57.41	57.52	58.43	45.40	46.71	44.64	56.65$_{(\pm 8.63)}$
Ours-ResNext50	70.13	70.88	70.29	58.38	58.85	58.02	55.59	55.09	56.25	61.50$_{(\pm 6.43)}$

performance loss. Conversely, excessively large S values minimize local variation, making the meta-distribution collapse to fix distributions. Thus, we find a moderate value is optimal.

Table 18. Ablation of S on CIFAR-100-LT (SADE's Setting)

S	Foward-LT						Uni.	Backward-LT						Mean
	100	50	25	10	5	2	1	2	5	10	25	50	100	
100	55.33	52.29	49.22	44.07	39.31	33.31	28.27	23.53	17.60	14.50	9.64	7.32	5.40	29.21(±17.03)
500	56.92	54.51	51.89	47.99	44.34	39.17	34.18	29.31	25.93	20.51	16.85	15.12	19.43	35.09(±14.44)
1000	69.73	66.81	64.17	59.26	55.64	49.87	45.47	41.68	40.39	37.90	36.33	33.27	31.72	48.63(±12.70)
5000	68.51	65.93	63.53	59.52	56.06	51.90	50.37	46.99	45.52	45.05	43.84	43.72	43.64	52.66(± 8.73)
10000	68.32	66.21	63.09	59.49	56.35	52.62	48.38	46.40	45.05	44.79	43.71	44.41	44.25	52.54(± 8.74)
50000	68.51	66.21	63.46	59.57	55.86	52.14	49.35	46.71	44.44	43.42	42.25	42.60	42.14	52.05(± 9.30)
100000	67.57	65.12	62.65	59.65	56.53	52.14	49.42	46.27	44.77	44.32	42.83	42.92	42.61	52.06(± 8.82)
500000	67.81	66.05	63.46	59.91	56.53	52.14	49.12	45.48	44.28	43.83	42.93	43.28	43.83	52.20(± 9.04)
1000000	67.43	65.57	62.52	58.75	55.66	51.76	48.03	45.13	42.98	42.31	41.14	41.27	41.34	51.07(± 9.47)

Table 19. Ablation of S on CIFAR-100-LT (Ours Setting)

S	Foward-LT			Uniform			Backward			Mean
	1	2	3	1	2	3	1	2	3	
100	54.41	53.35	58.11	29.24	26.00	28.70	9.12	5.81	9.41	30.46(±19.42)
500	55.47	54.42	59.24	31.50	34.43	35.69	28.77	16.75	23.64	37.77(±14.25)
1000	67.47	66.47	69.23	45.18	42.86	48.20	31.01	26.50	32.53	47.72(±15.62)
5000	66.80	66.40	68.96	48.78	47.89	49.35	42.64	41.88	42.47	52.80(±10.66)
10000	65.96	65.80	68.10	50.55	48.00	47.12	42.83	44.53	41.84	52.75(±10.14)
50000	66.85	65.33	68.74	47.25	48.00	47.89	40.22	41.28	40.48	51.78(±11.15)
100000	65.23	66.33	67.62	47.37	48.86	48.04	43.85	43.68	42.68	52.63(± 9.95)
500000	66.63	66.20	68.10	48.35	48.16	47.81	42.27	43.25	43.10	52.65(±10.37)
1000000	65.85	63.72	67.29	46.52	47.78	45.82	41.99	41.03	38.70	50.97(±10.73)

Table 20. Ablation of S on ImageNet-LT (SADE's Setting)

S	Foward-LT					Uni.	Backward-LT					Mean
	50	25	10	5	2	1	2	5	10	25	50	
100	70.24	68.02	64.64	61.60	57.17	54.06	49.91	47.10	42.88	43.14	41.13	54.54(±10.04)
500	70.42	67.87	64.12	60.31	56.09	51.27	46.23	43.07	37.40	30.67	27.25	50.43(±14.11)
1000	70.53	68.05	64.00	60.18	56.33	52.03	47.13	41.79	50.14	48.37	46.71	55.02(± 9.06)
5000	69.94	68.07	65.35	62.35	59.42	57.40	56.15	54.30	54.31	54.25	53.66	59.56(± 5.69)
10000	70.09	68.46	65.93	63.22	60.50	58.61	57.27	55.27	55.04	55.38	55.33	60.46(± 5.36)
50000	69.42	67.99	65.54	63.15	60.27	58.55	57.36	55.96	55.26	55.00	55.23	60.34(± 5.11)
100000	69.31	67.65	65.50	62.81	60.41	58.56	57.36	56.15	55.47	55.11	55.25	60.33(± 4.98)
500000	69.31	67.75	65.61	63.28	60.53	58.59	57.29	56.21	55.66	55.40	55.23	60.44(± 4.99)
1000000	68.81	67.47	65.18	62.73	60.45	58.63	57.44	56.49	55.98	55.63	55.88	60.43(± 4.66)

F.8. The Effect of The Number of Experts

In the DirMixE method, three experts are initially considered. However, the DirMixE framework can be extended to accommodate varying numbers of experts. We explore the impact of the number of experts on CIFAR 100-LT, specifically examining cases where the number is set to 1, 2, 3, 5, 7, and 9, where each expert corresponds to a distinct Dirichlet distribution component.

The forward and backward components are characterized by an imbalance ratio denoted as ρ.

Denote the training class distribution as $P_1, P_2, \ldots, P_C$, where C denotes the number of classes.

Specifically, we will need the following notations for experts handling different levels of imbalance ratio.

Table 21. Ablation of S on ImageNet-LT (Ours Setting)

S	Forward-LT			Uniform			Backward-LT			Mean
	1	2	3	1	2	3	1	2	3	
100	70.21	70.26	70.74	53.82	53.36	53.22	42.44	40.73	42.38	55.24$_{(\pm 11.73)}$
500	70.30	70.33	70.24	50.90	51.89	51.10	27.32	24.87	24.96	49.10$_{(\pm 18.28)}$
1000	70.65	70.42	70.45	51.01	52.14	51.04	46.58	46.42	46.42	56.13$_{(\pm 10.37)}$
5000	70.03	69.94	70.47	56.75	57.53	56.55	54.11	53.72	55.03	60.46$_{(\pm\ 6.95)}$
10000	70.13	70.88	70.29	58.38	58.85	58.02	55.59	55.09	56.25	61.50$_{(\pm\ 6.43)}$
50000	69.50	69.55	69.70	58.58	59.29	58.37	54.96	54.52	55.31	61.09$_{(\pm\ 6.21)}$
100000	69.44	69.20	69.37	58.29	58.90	58.36	55.15	55.53	56.14	61.15$_{(\pm\ 5.91)}$
500000	69.68	70.09	69.90	58.43	59.10	58.26	54.95	54.61	55.87	61.21$_{(\pm\ 6.31)}$
1000000	69.02	69.24	69.12	58.26	58.47	58.26	56.02	55.83	56.18	61.16$_{(\pm\ 5.72)}$

The forward component $\alpha^{(f_\rho)}$ with imbalance ratio ρ is set element-wisely as:

$$\alpha_i^{(f_\rho)} = \frac{S * \rho^{(i-1)/(C-1)}}{\sum_{j=1}^{C} \rho^{(j-1)/(C-1)}},\ i = 1, 2, \dots, C,$$

where S is the normalization factor.

The backward component $\alpha^{(b_\rho)}$ with imbalance ratio ρ is set element-wisely as:

$$\alpha_i^{(b_\rho)} = \frac{S * \rho^{(C-i)/(C-1)}}{\sum_{j=1}^{C} \rho^{(j-1)/(C-1)}},\ i = 1, 2, \dots, C$$

where S is the normalization factor.

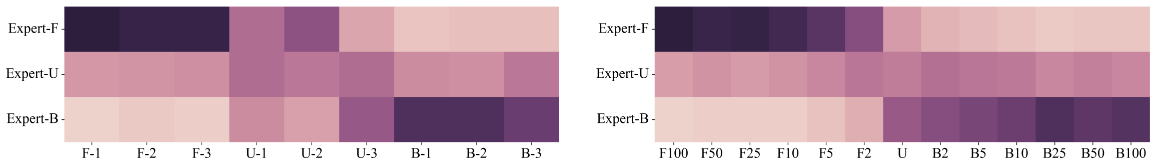
The uniform component $\alpha^{(u)}$ is set element-wisely as:

$$\alpha_i^{(u)} = \frac{S}{C}, i = 1, 2, \dots, C$$

where S is the normalization factor.

In the case of a single expert, we utilize one Dirichlet component $\alpha^{(u)}$. For two experts, we employ two components, $\alpha^{(f_{100})}$ and $\alpha^{(b_{100})}$. For three experts, we select the uniform component $\alpha^{(u)}$ and the forward and backward component with $\rho = 100$. For scenarios with more than three experts, we will symmetrically add the forward components and backward components based on the 3-experts configuration. For five experts, we select the uniform component $\alpha^{(u)}$ and the forward/backward components with $\rho = 100, 50$. For seven experts, we select the uniform component $\alpha^{(u)}$ and the forward/backward component with $\rho = 100, 50, 25$. For nine experts, we select the uniform component $\alpha^{(u)}$ and the forward/backward components with $\rho = 100, 50, 25, 10$.

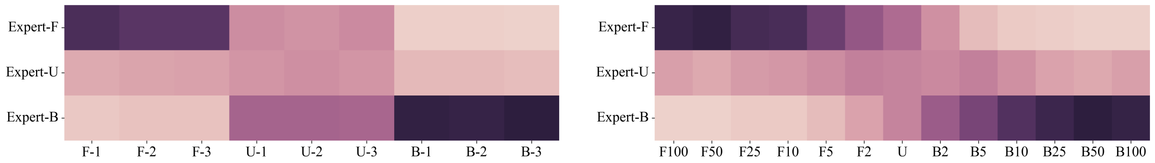
The results are reported in Fig.11. This indicates that the average ACC is monotonically increasing as more and more experts are employed.



(a) Our's Setting

(b) SADE's Setting

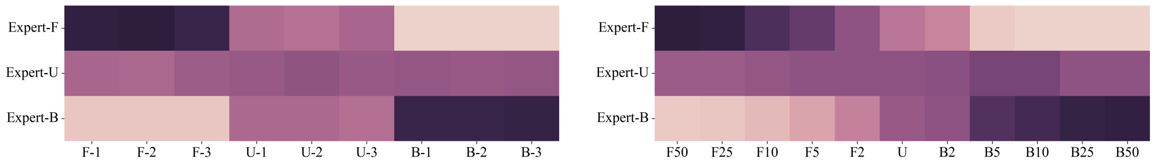
Figure 6. **Weight Assignment in the Self-supervised Aggregations on CIFAR-100. F,U,B** represent the forward, uniform and backward distributions. For the x-axis, **F-1,F-2,F-3** in **(a)** denote the three observed label distributions in the test data respectively. **F-2,F-5,$\cdots$,F-100** represents the corresponding imbalance ratio of the forward distribution under SADE's setting. The case for **the suffices of U and B** are similar. **Expert-F,U,B** represents the experts assigned to the forward, backward, uniform Dirichlet distribution, respectively.



(a) Our's Setting

(b) SADE's Setting

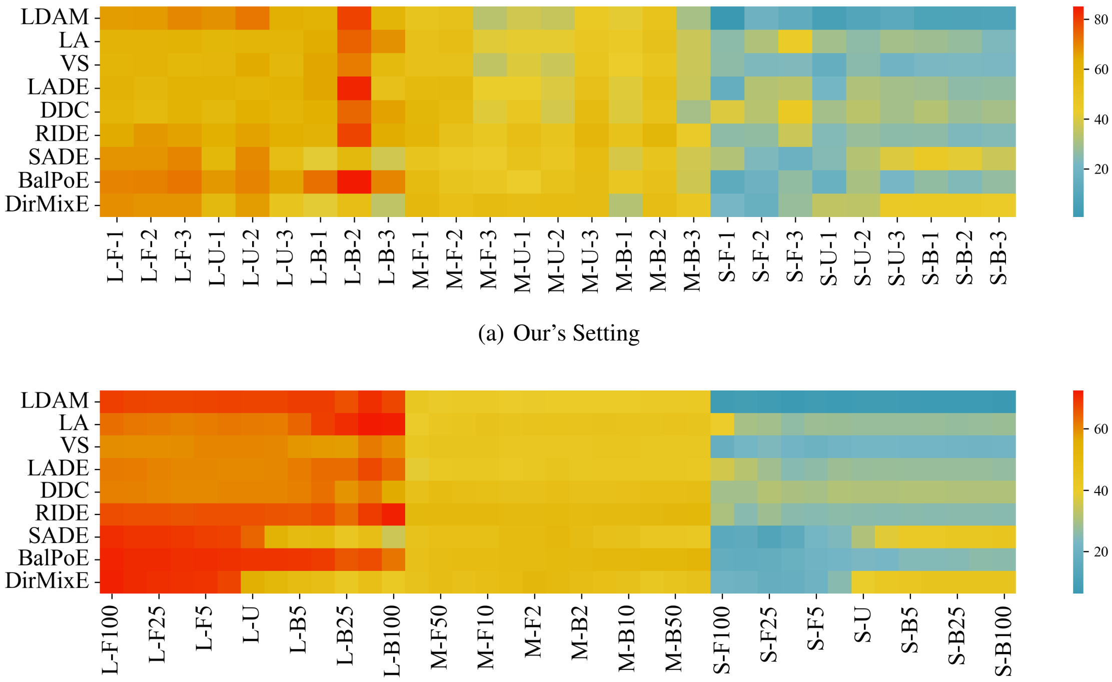
Figure 7. **Weight Assignment in the Self-supervised Aggregations on CIFAR-10. F,U,B** represent the forward, uniform and backward distributions. For the x-axis, **F-1,F-2,F-3** in **(a)** denote the three observed label distributions in the test data respectively. **F-2,F-5,$\cdots$,F-100** represents the corresponding imbalance ratio of the forward distribution under SADE's setting. The case for **the suffices of U and B** are similar. **Expert-F,U,B** represents the experts assigned to the forward, backward, uniform Dirichlet distribution, respectively.



(a) Our's Setting

(b) SADE's Setting

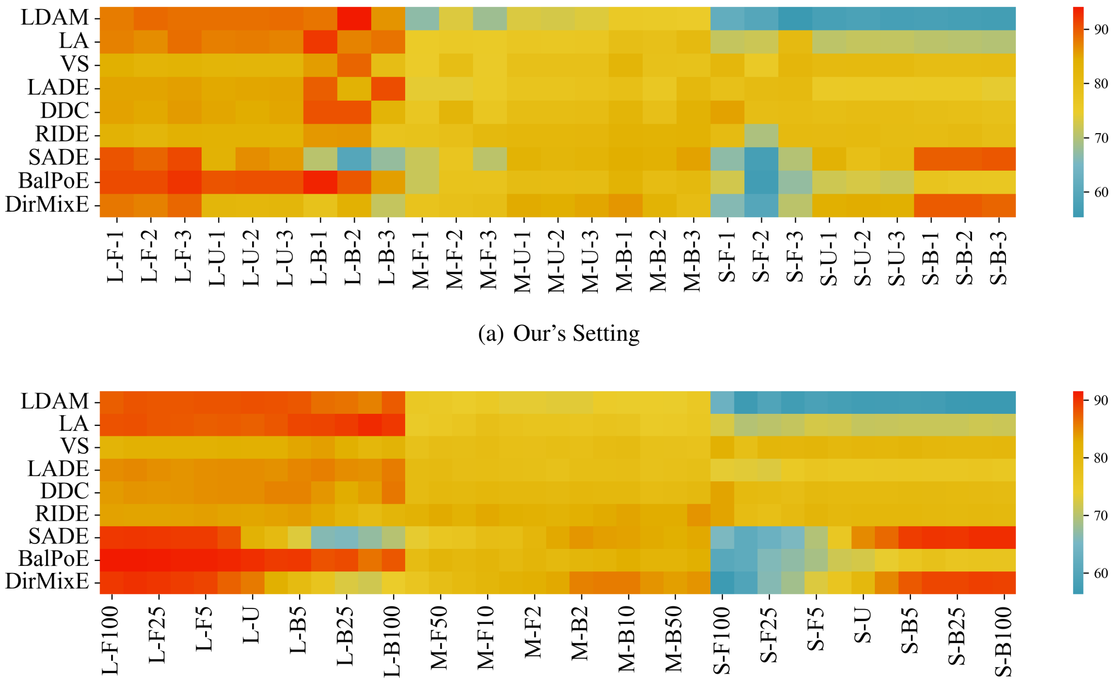
Figure 8. **Weight Assignment in the Self-supervised Aggregations on ImageNet. F,U,B** represent the forward, uniform and backward distributions. For the x-axis, **F-1,F-2,F-3** in **(a)** denote the three observed label distributions in the test data respectively. **F-2,F-5,$\cdots$,F-100** represents the corresponding imbalance ratio of the forward distribution under SADE's setting. The case for **the suffices of U and B** are similar. **Expert-F,U,B** represents the experts assigned to the forward, backward, uniform Dirichlet distribution, respectively.



(a) Our's Setting

(b) SADE's Setting

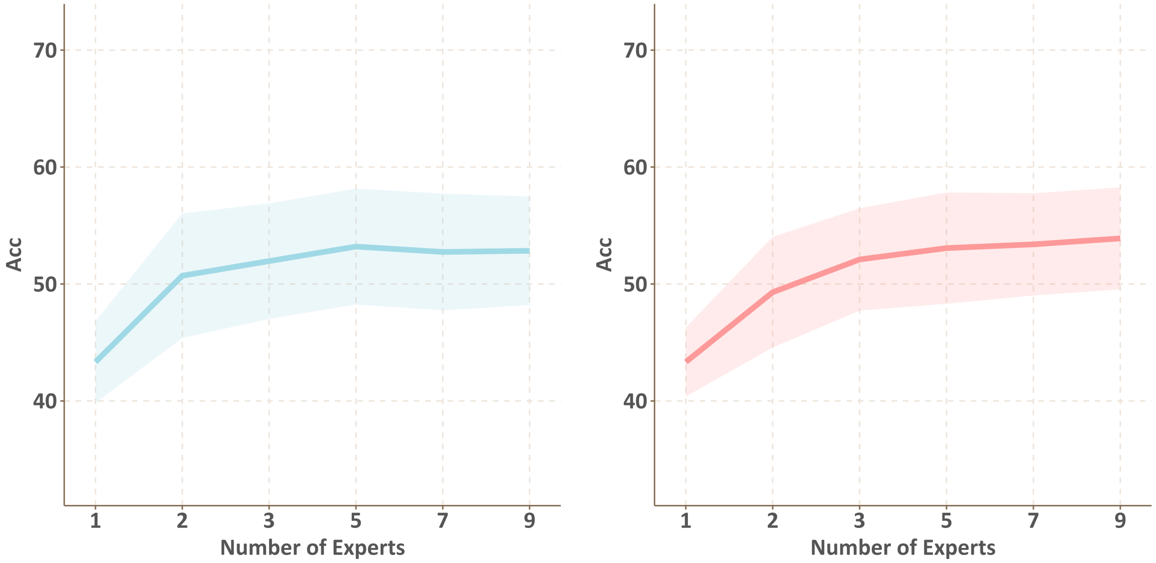
Figure 9. **Fine-grained Performance of DirMixE on CIFAR-100-LT F,U,B** represent the forward, uniform and backward distributions. For the x-axis, **F-1,F-2,F-3** in **(a)** denote the three observed label distributions in the test data respectively. **F2,F5,· · · ,F100** in **(b)** represents the corresponding imbalance ratio of the forward distribution under SADE's setting. The case for **the suffixes of U and B** are similar. **L, M, S** denote the many-shot, medium-shot and few-shot classes respectively. For example, **L-B100** indicates the performance of medium-shot classes on the backward distribution with an imbalance ratio of 100.



(a) Our's Setting

(b) SADE's Setting

Figure 10. **Fine-grained Performance of DirMixE on CIFAR-10-LT F,U,B** represent the forward, uniform and backward distributions. For the x-axis, **F-1,F-2,F-3** in **(a)** denote the three observed label distributions in the test data respectively. **F2,F5,· · · ,F100** in **(b)** represents the corresponding imbalance ratio of the forward distribution under SADE's setting. The case for **the suffixes of U and B** are similar. **L, M, S** denote the many-shot, medium-shot and few-shot classes respectively. For example, **L-B100** indicates the performance of medium-shot classes on the backward distribution with an imbalance ratio of 100.



(a) Our's Setting

(b) SADE's Setting

Figure 11. **The Effect of Using Different Number of Experts.**