Does Mixed Label Imbalance Matter to Minority Collapse in Imbalanced Learning?

Hoyong Kim[1,2]
[1] DGIST InnoCORE Research Group for Bio-Embodied Physical AI
[2] Seoul National University
hoyong.kim@snu.ac.kr

Kangil Kim*
Department of Artificial Intelligence,
Gwangju Institute of Science and Technology
kangil.kim.01@gmail.com

Abstract

Minority collapse, where the class vectors of minority classes converge and become indistinguishable, is a significant challenge in imbalanced learning. It is commonly addressed by Mixup combined with class-balanced sampling. Although minority collapse has been mathematically analyzed using the unconstrained features model (UFM) alongside Neural Collapse (NC), no prior work has analyzed minority collapse under Mixup, particularly from the perspective of *mixed labels*, the class pairs that Mixup interpolates. We investigate this overlooked factor and raise the question: *Is mixed label balance important for alleviating minority collapse?* Our analysis reveals that (i) mixed labels should be balanced, because class-balanced sampling balances them only in expectation and minority collapse persists otherwise, and (ii) when mixed labels are balanced, interpreting them as singleton labels recovers the balanced-case simplex equiangular tight frame geometry for the singleton classifier. Motivated by this analysis, we propose a *Balanced Mixed Label Sampler* (BMLS), which equalizes the frequencies of mixed labels in every epoch, and a *Mixed-Singleton* (MS) classifier, which treats each mixed label as its own target without adding parameters. Through theoretical analysis, direct measurements of the collapse geometry, and experiments on long-tailed benchmarks, we highlight the importance of balancing mixed labels in imbalanced learning. Code is publicly available at https://github.com/khyong/BMLS.

1 Introduction

In imbalanced learning, severe class imbalance often causes a significant degradation of model accuracy, particularly on the minority classes [1–6]. One known cause of this performance drop is the phenomenon termed *minority collapse* [7], wherein the class vectors of minority classes converge and become nearly identical. To mitigate this issue, a wide range of strategies has been explored, including data augmentation [8–10], calibration technique [11], mixture-of-experts models [12–14], and class-balanced loss functions [15, 16] or sampling schemes [15, 17–19]. Among these approaches, Mixup [8], especially when combined with class-balanced sampling, has been shown to effectively improve the model performance under class-imbalanced conditions by linearly interpolating randomly selected pairs of samples in both input and label space. Throughout, a *singleton label* denotes an ordinary one-hot label of a single class, as opposed to a *mixed label* (a, b) produced by Mixup, i.e., the pair of classes whose samples are interpolated.

Meanwhile, Neural Collapse (NC) [20] has emerged as a key framework for analyzing the geometry of last-layer features and classifier at the terminal phase of training, where class means and class vectors converge to a simplex equiangular tight frame (ETF). Although NC has been studied in both

*corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

imbalanced learning [21, 22] and Mixup [23] separately, Mixup in imbalanced settings has not been investigated in conjunction with NC. In particular, the balance of mixed labels has received little attention. This motivates our central research question: *Could the balance of mixed labels be a critical factor in minority collapse?*

We answer this question with a negative result. Following the proof approach of Fang et al. [7], which we deliberately reuse so that the difference from the singleton-label case is attributable to Mixup rather than to a change of technique, we prove that *minority collapse still occurs under Mixup when the frequencies of mixed labels are imbalanced* (Theorem 4.2). This explains why existing class-balanced samplers only partially alleviate it: under random permutation, balancing singleton labels balances mixed labels *in expectation* but not within each epoch, and we observe this *epoch-wise label imbalance* directly (Figure 2). Controlling the fraction of mixed labels covered per epoch shows that this imbalance weakens the mitigation of minority collapse, both in accuracy and in the collapse geometry (Figure 3a and Table 1). To address this issue, we propose the **B**alanced **M**ixed **L**abel **S**ampler (BMLS), which equalizes the frequencies of mixed labels in every epoch (Section 5.1). As a secondary and conditional component, we introduce the **M**ixed-**S**ingleton (MS) classifier, which treats each mixed label as its own target and scores it against the interpolation of the two singleton class vectors (Section 5.2).

Our contributions are summarized as follows.

- **A negative result.** Minority collapse persists under Mixup unless mixed labels are balanced; the operative quantity is the epoch-wise dispersion of mixed-label frequencies, not their mean (Theorem 4.2).
- **Results beyond [7].** We relax the dimension requirement of the semi-definite program (SDP) from $p \geq K^2 + K$ to $p \geq 2K$, show that the UFM with MS attains the global minimum of the balanced K-class UFM (Theorem G.5), and show that restricting mixed labels to $a < b$ preserves the simplex ETF (Proposition G.7).
- **Methods and evidence.** BMLS is applicable on its own, and MS is a parameter-free classifier whose benefit is conditional on balanced mixed labels; collapse metrics respond monotonically to mixed-label coverage, and per-pair coverage explains when MS helps at scale (Section 6).

2 Related Work

Mixup-based Method. Many works address imbalanced learning with Mixup [8], which increases the diversity of sampled data and alleviates overfitting on tail classes, through data augmentation, architecture improvements, and calibration (Section C.1); none has studied the frequency balance of mixed labels in minority collapse.

Class-balanced Methods. Various class-balanced samplers have been proposed (see more references in Section C.2), yet no work has mainly focused on the frequency balance of mixed labels. Additionally, while Logit Adjustment [24] and UniMix [25] have concentrated on the effect of the class vectors of singleton labels, they did not interpret mixed labels as singletons.

Neural Collapse under Mixup in Imbalanced Learning. NC in imbalanced learning has been studied through an unconstrained features model (UFM) [26–32] (or a layer-peeled model [7, 21, 23, 33]). To alleviate the minority collapse, Yang et al. [21] assumed that the classifier is fixed to the K-simplex ETF and proved that UFM with the classifier satisfies NC properties. Fisher et al. [23] proved that Mixup also satisfies NC properties for both same-class and different-class pairs. However, [21] and [23] did not consider the minority collapse from the frequency of mixed labels in the UFM with learnable classifiers. (See more references for NC and discussion on UFM in Section C.3.)

3 Problem Setting

Notations. Let $\mathcal{X}$ be the dataset with N samples where the number of singleton label classes is K and $\mathbb{S}$ be the set of their feature vectors $\boldsymbol{h}$. Then, we formulate them as $\mathcal{X} := [(\boldsymbol{x}_i, c_i)]_{i=1}^N$ where c_i is the class label of the i-th sample x_i and $\mathbb{S} := \{\boldsymbol{h}_i\}_{i=1}^N$. As a result, we define $\boldsymbol{y}_i = \boldsymbol{e}^{(c_i)}$ as the one-hot vector of x_i. Then, we denote the subset of $\mathbb{S}$ which has only k-th class feature vectors $\boldsymbol{h}_{k,i}$ as $\mathbb{S}_k := \{\boldsymbol{h}_{k,i}\}_{i=1}^{n_k}$ where n_k is the number of k-th class samples and $k \in [K]$. Thus, $N = \sum_{k=1}^K n_k$. All symbols are summarized in the notation table (Section A).

Mixup. Mixup constructs sample pairs by permuting sample indices, then linearly interpolates each sample with its paired counterpart in both input and label space. Let $\mathcal{I} := [i]_{i=1}^N$ be the indices of $\mathcal{X}$ and $\pi(\mathcal{I}) := [\pi(i)]_{i=1}^N$ be the permuted one where $\pi(i)$ represents the index number corresponding to i-th element of $\mathcal{I}$. Therefore, the index pairs of mixed samples $\mathcal{I}^\lambda$ is denoted as $\mathcal{I}^\lambda := [(i, \pi(i))]_{i\in\mathcal{I}}$. In this case, we denote $\mathcal{I}^\lambda_{(a,b)}$ as the index pairs of $(c_i, c_{\pi(i)}) = (a, b)$, and $\mathbb{S}^\lambda_{(a,b)}$ as the mixed feature set of (a, b)-label samples. Therefore, $\mathbb{S}^\lambda_{(a,b)} := \{\lambda \boldsymbol{h}_{a,i} + (1-\lambda)\boldsymbol{h}_{b,j} \mid (i,j) \in \mathcal{I}^\lambda_{(a,b)}\} = \{\boldsymbol{h}^\lambda_{(a,b),i}\}_{i=1}^{n_{(a,b)}}$ where $(a, b) \in \mathbb{K}^2$, $n_{(a,b)} = |\mathcal{I}^\lambda_{(a,b)}|$, and $\mathbb{K}^2 = \{(a, b) | 1 \le a \le K,\ 1 \le b \le K\}$. Thus, $N = \sum_{(a,b)\in\mathbb{K}^2} n_{(a,b)}$.

Based on the notations, we perform mixup on each pair defined by $\mathcal{I}^\lambda$ to create mixed-label samples by linearly interpolating them:

$$\begin{aligned} \boldsymbol{x}^\lambda_i &= \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_{\pi(i)}, \\ \boldsymbol{y}^\lambda_i &= \lambda \boldsymbol{y}_{c_i} + (1-\lambda)\boldsymbol{y}_{c_{\pi(i)}}, \forall (i, \pi(i)) \in \mathcal{I}^\lambda, \end{aligned} \tag{1}$$

where mixup ratio $\lambda \in (0, 1)$ is sampled from the beta distribution D_λ, i.e., $\lambda \sim D_\lambda(\alpha, \alpha)$ and α is a hyperparameter.

Unconstrained Features Model with Mixup (UFM$_\lambda$). The unconstrained features model (UFM) [7] is the optimization program of simplified neural network, modeled by only last-layer features and classifier. Following the definition of UFM, we obtain the unconstrained features model with Mixup (UFM$_\lambda$):

$$\begin{aligned} &\min_{\boldsymbol{W}, \boldsymbol{H}^\lambda} \mathbb{E}_\lambda \frac{1}{N} \sum_{k\in\mathbb{K}^2} \sum_{i=1}^{n_k} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}^\lambda_{k,i}, \boldsymbol{y}^\lambda_k) \\ &\text{s.t. } \frac{1}{K}\sum_{k=1}^K \|\boldsymbol{w}_k\|^2 \le E_W, \frac{1}{K^2}\sum_{k\in\mathbb{K}^2} \frac{1}{n_k}\sum_{i=1}^{n_k} \left\|\boldsymbol{h}^\lambda_{k,i}\right\|^2 \le E_H, \end{aligned} \tag{2}$$

where $\boldsymbol{y}^\lambda_{(a,b)} = \lambda \boldsymbol{e}^{(a)} + (1-\lambda)\boldsymbol{e}^{(b)}$. For simplicity, we hereafter denote $\boldsymbol{W} = [\boldsymbol{w}_k]_{k=1}^K \in \mathbb{R}^{K\times p}$ for the weights of the classifier, whose row $\boldsymbol{w}_k$ is the class vector of class k, and $E_W \propto 1/K$ and $E_H \propto 1/K$ for the positive norm budgets of class vectors and features, respectively. $\mathcal{L}$ is the cross-entropy loss with softmax function,

$$\mathcal{L}(\boldsymbol{z}^\lambda(k), \boldsymbol{y}^\lambda_k) = -\lambda \log\left(\frac{\exp(\boldsymbol{z}^\lambda(a))}{\sum_{k'=1}^K \exp(\boldsymbol{z}^\lambda(k'))}\right) - (1-\lambda)\log\left(\frac{\exp(\boldsymbol{z}^\lambda(b))}{\sum_{k'=1}^K \exp(\boldsymbol{z}^\lambda(k'))}\right), \tag{3}$$

where $\boldsymbol{z}^\lambda(k')$ denotes the k'-th entry of the logit $\boldsymbol{z}^\lambda_i = \boldsymbol{W}\boldsymbol{h}^\lambda_{k,i}$, and $k = (a, b)$.

Convex Optimization Program. We present a convex optimization program that serves as a relaxation of the non-convex UFM$_\lambda$ (Equation (2)), leveraging the established result that a quadratically constrained quadratic program can be transformed into the following semi-definite program [34]:

$$\begin{aligned} &\min_{\boldsymbol{X}} \sum_{k\in\mathbb{K}^2} \frac{n_k}{N} \mathcal{L}(\boldsymbol{z}(k)^\lambda, \boldsymbol{y}^\lambda_k) \\ &\text{s.t. } \frac{1}{K^2}\sum_{k=1}^{K^2} \boldsymbol{X}(k,k) \le E_H, \frac{1}{K}\sum_{k=K^2+1}^{K^2+K} \boldsymbol{X}(k,k) \le E_W, \end{aligned} \tag{4}$$

where $\boldsymbol{X} \succeq 0 \in \mathbb{R}^{(K^2+K)\times(K^2+K)}$ is the Gram matrix of $[\boldsymbol{h}^\lambda_{(1,1)}, \ldots, \boldsymbol{h}^\lambda_{(K,K)}, \boldsymbol{W}^\top]$ with $\boldsymbol{h}^\lambda_k$ the mean feature of mixed label k, $\boldsymbol{X}(i,j)$ is its (i,j)-th entry, and $\boldsymbol{z}_k = \left[\boldsymbol{X}(k, K^2+1), \boldsymbol{X}(k, K^2+2), \ldots, \boldsymbol{X}(k, K^2+K)\right]^\top$ is the logit of mixed label k for all $1 \le k \le K^2$. Thus, the first K^2 diagonal entries of $\boldsymbol{X}$ are squared feature norms and the last K are squared class-vector norms. The derivation of Equation (4) and the justification of the SDP relaxation are in Section D and Section E.

4 Main Result

In this section, we present two key theoretical results grounded in the UFM under Mixup (Equation (2)) and its convex optimization program via the semi-definite relaxation (Equation (4)). We first establish

that Equation (2) and Equation (4) share the same global minimizer (Lemma 4.1), and subsequently prove that (i) the imbalanced mixed label distribution gives rise to minority collapse (Theorem 4.2). Finally, we demonstrate that (ii) balancing mixed labels alleviates minority collapse in UFM_λ (Proposition 4.3).

Proof sketch. As in Fang et al. [7], we follow the four-step argument so that the difference from the singleton-label case is attributable to Mixup rather than to a change of technique: (1) define the UFM_λ (Equation (2)); (2) NC at the balanced global optimum, which holds under Mixup by [23]; (3) show that the SDP relaxation is tight (Lemma 4.1) and that, as $R \to \infty$, the loss of the minor group saturates at $\log K$ regardless of λ, which forces the minority class vectors to coincide (Theorem 4.2); and (4) show that reweighting minor mixed labels by w_r acts as oversampling (Proposition 4.3). What is new is the object of the analysis, the frequencies of mixed labels, and the results without counterpart in [7]: the negative result of Theorem 4.2, the SDP dimension relaxation below, and the analysis of MS (Section 5.2); full proofs are in Section G.

4.1 Mixed label imbalance causes minority collapse under Mixup

Lemma 4.1. *Assume $p \geq 2K$ and the loss function $\mathcal{L}$ is convex in its first argument. Let $\boldsymbol{X}^\star$ be a minimizer of the convex program (Equation (4)). Define $\left(\boldsymbol{W}^\star, \boldsymbol{H}^{\lambda\star}\right)$ as*

$$\left[\boldsymbol{h}^{\lambda\star}_{(1,1)}, \boldsymbol{h}^{\lambda\star}_{(1,2)}, \ldots, \boldsymbol{h}^{\lambda\star}_{(K,K)}, \ (\boldsymbol{W}^\star)^\top\right] = \boldsymbol{P}(\boldsymbol{X}^\star)^{1/2} \tag{5}$$
$$\boldsymbol{h}^{\lambda\star}_{k,i} = \boldsymbol{h}^{\lambda\star}_k, \ \text{ for all } i \in \mathcal{I}^\lambda_k, k \in \mathbb{K}^2,$$

where $(\boldsymbol{X}^\star)^{1/2}$ denotes the positive square root of $\boldsymbol{X}^\star$ and $\boldsymbol{P} \in \mathbb{R}^{p\times(K^2+K)}$ is any partial orthogonal matrix such that $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_{K^2+K}$. Then, $(\boldsymbol{W}^\star, \boldsymbol{H}^{\lambda\star})$ is a minimizer of Equation (2). Moreover, if all $\boldsymbol{X}^\star$'s satisfy $\frac{1}{K^2}\sum_{k=1}^{K^2} \boldsymbol{X}^\star(k,k) = E_H$, then all the solutions of Equation (2) are in the form of Equation (5). (Proof in Section G.2)

In general, SDP requires a strong constraint $p \geq K^2 + K$ to guarantee the existence of the left-hand side term in Equation (5). However, $p \geq 2K$ alone is sufficient to guarantee it, since mixed features $\boldsymbol{h}^{\lambda\star}$ are linearly dependent on singleton features $\boldsymbol{h}^\star$. The rationale of this SDP constraint relaxation is provided in Theorem G.5 and more discussion of it is deferred to Section E. Note that $p \geq 2K$ is a condition for representing the solution, not for the collapse mechanism: Theorem 4.2 below requires only $p \geq K$, which holds in all our experiments (e.g., $p = 128$ and $K = 100$ on CIFAR100-LT), and the same asymmetry appears in [7].

Building upon Lemma 4.1, we analyze the imbalanced case in which singleton classes are partitioned into two distinct groups (a major group $\mathbb{K}_A := \{k \mid 1 \leq k \leq K_A\}$ and a minor group $\mathbb{K}_B := \{k \mid K_A + 1 \leq k \leq K\}$) with differing label frequencies ($|\mathbb{S}_{k\in\mathbb{K}_A}| = n_A > |\mathbb{S}_{k\in\mathbb{K}_B}| = n_B$).

Theorem 4.2. *Assume $p \geq K$ and $n_A/n_B \to \infty$, and fix K_A and K_B. Let $\left(\boldsymbol{W}^\star, \boldsymbol{H}^{\lambda\star}\right)$ be any global minimizer of the UFM_λ (Equation (2)). As the imbalance factor $R \equiv n_A/n_B \to \infty$, we have*

$$\lim \boldsymbol{w}^\star_k - \boldsymbol{w}^\star_{k'} = \boldsymbol{0}_p, \ \text{ for all } K_A < k < k' \leq K.$$

(Proof in Section G.3)

Thus, when the frequencies of minority mixed labels are extremely small, the minority class vectors collapse toward one another, so UFM_λ is inherently susceptible to minority collapse under imbalanced mixed label distributions.

4.2 Balancing mixed labels mitigates minority collapse in UFM with Mixup

To formalize the behavior of a neural network trained by minimizing a new program with balanced samples including mixed-label ones, we propose that it may perform as if it were trained on a larger dataset containing n_A examples in the majority class and $w_r n_B$ examples in the minority class. We

begin by analyzing the UFM$_\lambda$ in the context of balancing mixed labels:

$$\min_{\boldsymbol{W},\boldsymbol{H}^\lambda} \frac{1}{N'}\sum_{k\in\mathbb{K}_A^2}\sum_{i=1}^{n_A}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}^\lambda,\boldsymbol{y}_k^\lambda)+\frac{w_r}{N'}\sum_{k\in\mathbb{K}_B^2}\sum_{i=1}^{n_B}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}^\lambda,\boldsymbol{y}_k^\lambda)$$
$$\text{s.t. } \frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2\le E_W,\ \frac{1}{|\mathbb{K}_A^2|}\sum_{k\in\mathbb{K}_A^2}\frac{1}{n_A}\sum_{i=1}^{n_A}\left\|\boldsymbol{h}_{k,i}^\lambda\right\|^2+\frac{1}{|\mathbb{K}_B^2|}\sum_{k\in\mathbb{K}_B^2}\frac{1}{n_B}\sum_{i=1}^{n_B}\left\|\boldsymbol{h}_{k,i}^\lambda\right\|^2\le E_H, \tag{6}$$

where $N'=n_A|\mathbb{K}_A^2|+w_rn_B|\mathbb{K}_B^2|$, $\mathbb{K}_A^2=\{(a,b)|1\le a\le K_A, 1\le b\le K_A\}$ and $\mathbb{K}_B^2=\{(a,b)|K_A+1\le a\le K, K_A+1\le b\le K\}$.

The following result supports the intuition that balancing mixed labels mitigates minority collapse in the UFM$_\lambda$ by enhancing the size of the minority classes.

Proposition 4.3. *Assume $p\ge 2K$ and the loss function $\mathcal{L}$ is convex in the first argument. Let $\boldsymbol{X}^\star$ be any minimizer of the convex program (Equation (4)) with $n_{(1,1)}=n_{(1,2)}=\cdots=n_{(K_A,K_A)}=n_A$ and $n_{(K_A+1,K_A+1)}=n_{(K_A+1,K_A+2)}=\cdots=n_{(K,K)}=w_rn_B$. Define $\left(\boldsymbol{W}^\star,\boldsymbol{H}^{\lambda\star}\right)$ as*

$$\left[\boldsymbol{h}_{(1,1)}^{\lambda\star},\boldsymbol{h}_{(1,2)}^{\lambda\star},\ldots,\boldsymbol{h}_{(K,K)}^{\lambda\star},\ (\boldsymbol{W}^\star)^\top\right]=\boldsymbol{P}(\boldsymbol{X}^\star)^{1/2} \tag{7}$$
$$\boldsymbol{h}_{k_A,i}^{\lambda\star}=\boldsymbol{h}_{k_A}^{\lambda\star},\ \textit{for all } i\in\mathcal{I}_{k_A}^\lambda, k_A\in\mathbb{K}_A^2,$$
$$\boldsymbol{h}_{k_B,i}^{\lambda\star}=\boldsymbol{h}_{k_B}^{\lambda\star},\ \textit{for all } i\in\mathcal{I}_{k_B}^\lambda, k_B\in\mathbb{K}_B^2,$$

where $\boldsymbol{P}\in\mathbb{R}^{p\times(K^2+K)}$ is any partial orthogonal matrix such that $\boldsymbol{P}^\top\boldsymbol{P}=\boldsymbol{I}_{K^2+K}$ and $\boldsymbol{W}^\star\in\mathbb{R}^{K\times p}$, $\boldsymbol{H}^{\lambda\star}\in\mathbb{R}^{p\times N}$. Then, $(\boldsymbol{W}^\star,\boldsymbol{H}^{\lambda\star})$ is a global minimizer of the mixed-label balanced UFM$_\lambda$ (Equation (6)). Moreover, if all $\boldsymbol{X}^\star$'s satisfy $\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\boldsymbol{X}^\star(k,k)=E_H$, then all the solutions of Equation (6) are in the form of Equation (7). (Proof in Section G.4.)

Proposition 4.3 adapts the oversampling result of [7] to mixed labels, and its role is to enable the comparison in Remark 4.4. In conjunction with Lemma 4.1, Proposition 4.3 demonstrates that the number of training examples in each minority mixed label is effectively w_rn_B instead of n_B in the UFM$_\lambda$. In the special case where $w_r=n_A/n_B\equiv R$, the results indicate that the angles between any pair of class vectors are equal, regardless of whether they belong to the majority or minority classes.

Remark 4.4. According to Theorem 4.2, Mixup also experiences the minority collapse. Additionally, as proven in Proposition 4.3, even when using class-balanced samplers to alleviate label suppression and learn an unbiased classifier, minority collapse is partially mitigated but not fully resolved, as the frequency of mixed labels remains imbalanced. Under random permutation, class-balanced samplers balance mixed labels only in expectation; the epoch-wise dispersion of their frequencies remains (Figure 2). For this reason, when using Mixup in imbalanced learning, the frequency of not only singleton labels but also mixed ones should be balanced.

4.3 Assumptions of the theoretical model

Our analysis rests on three simplifications. **(A1) UFM and SDP relaxation.** Real backbones cannot realize arbitrary last-layer features; the UFM is the standard abstraction for NC [7, 23], and SDP relaxations of quadratically constrained quadratic programs are known to give approximate solutions with bounded gaps for deep networks [35, 36] (Section D). **(A2) Global optimality at the terminal phase of training.** SGD reaches an approximate rather than exact optimum; however, NC metrics tracked over 200 epochs show the predicted geometry emerging in practice (Figure 3b and Section I.1). **(A3)** $R=n_A/n_B\to\infty$**.** Practical imbalance is finite, so Theorem 4.2 is a limiting statement, but it makes a testable prediction, that collapse should weaken as mixed labels become more balanced, which we verify directly by controlling mixed-label coverage (Section 6.1). Accordingly, we treat the theory as identifying the operative quantity, the frequencies of mixed labels, rather than as a quantitative model of deep network training.

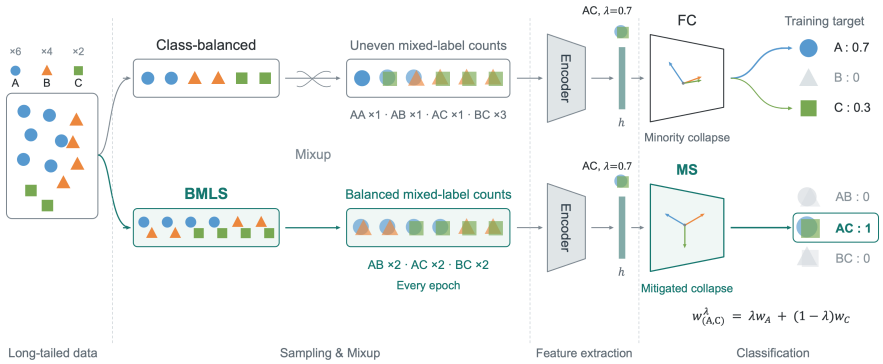


Figure 1: Overview of the Balanced Mixed Label Sampler (BMLS) and the Mixed-Singleton (MS) classifier on a toy long-tailed dataset with three classes (6, 4, and 2 samples). *Sampling & Mixup:* after Mixup, a random sampler and a class-balanced sampler still produce mixed labels with unequal frequencies (e.g., 1:1:1:3), whereas BMLS produces every mixed label equally often (2:2:2). *Classification:* a standard fully connected (FC) classifier scores a mixed sample against the singleton classes, whereas MS scores it against its own mixed label, whose class vector is the interpolation of the two singleton class vectors.

5 Method

As summarized in Figure 1 and Algorithm 1, BMLS, the primary method, changes *which* pairs are formed, and MS, a secondary component whose benefit is conditional on balanced mixed labels, changes *how* the mixed target is scored.

5.1 Balanced Mixed Label Sampler

We propose the Balanced Mixed Label Sampler (BMLS), where the frequency of all mixed-label samples is equal in each epoch. When using BMLS, the probability of sampling of a i-th mixed sample is

$$P\left((i,\pi(i))\right) = \frac{1}{N}, \quad \forall (i,\pi(i)) \in \tilde{\mathcal{I}}^{\lambda}, \tag{8}$$

where $\tilde{\mathcal{I}}^{\lambda}$ is the index pairs of samples where the frequency of mixed labels is balanced, i.e., $n_{(a,b)} = n$ for all $(a,b) \in \mathbb{K}^2$. As done in the class-aware sampler (CAS) [19], we remove the randomness by pre-defining $\tilde{\mathcal{I}}^{\lambda}$ for every epoch. After generating $\tilde{\mathcal{I}}^{\lambda}$, we simply replace $\mathcal{I}^{\lambda}$ to $\tilde{\mathcal{I}}^{\lambda}$ in Equation (1).

As proven in Theorem 4.2 and Proposition 4.3, *the minority collapse observed in Mixup arises from the imbalanced frequency of mixed-label samples*. Consequently, the proposed sampler mitigates the minority collapse of Mixup by performing sampling after pre-balancing the frequency of all label samples, including mixed labels, as formulated in Equation (8). BMLS relies only on these two results, not on MS, and works with a standard fully connected (FC) classifier.

We generate mixed labels (a,b) only for $a < b$, which reduces the mixed-label space to $|\mathbb{K}^{<}| = K(K-1)/2$. This restriction is theoretically admissible: the simplex ETF over singleton classes survives it (Proposition G.7), so Theorem 4.2 and Proposition 4.3 still hold.

5.2 Mixed-Singleton Classifier

Intuition. A standard Mixup classifier scores a mixed sample by combining the scores of the two source classes, weighted by λ. The Mixed-Singleton (MS) classifier instead treats the pair (a,b) as its own target and scores it against the interpolated class vector $\lambda \boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_b$. No parameters are added, because the mixed class vectors remain functions of the K singleton class vectors.

Motivation. In the proof of Theorem 4.2, the loss of the minor group saturates at $\mathcal{L}_B^{\ell}\left(\boldsymbol{W}^{\ell,A}, \boldsymbol{H}^{\ell,A}\right) = \lambda \cdot \log(K) + (1-\lambda) \cdot \log(K) = \log(K)$ (Equation (19)): inside the collapse regime, λ does not affect which configuration is optimal. This λ-*independence* is the premise of MS, not its justification: if λ does not determine the optimum, the soft-label cross-entropy over

Algorithm 1 Mixup Training with Balanced Mixed Label Sampler and Mixed-Singleton Classifier

Require: Training dataset $\mathcal{X} = [(\boldsymbol{x}_i, c_i)]_{i=1}^N$, indices of training dataset $\mathcal{I} = [i]_{i=1}^N$, number of classes K, mixed label set $\mathbb{K}^2 = \{(a,b) | 1 \le a \le K,\ 1 \le b \le K\}$, mixing coefficient $\alpha > 0$, number of epochs T

for $t = 1, 2, \ldots, T$ **do**
  Sample $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$
  Set $\lambda \leftarrow \max(\lambda, 1-\lambda)$
  Set $\tilde{\mathcal{I}}^\lambda$ where $P\left((i, \pi(i))\right) = \frac{1}{N},\ \forall (i, \pi(i)) \in \tilde{\mathcal{I}}^\lambda$ ▷ Construct mixed index pairs (Eq. 8)
  Set $\boldsymbol{W}^\lambda \leftarrow [\lambda \boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_b]_{(a,b)\in\mathbb{K}^2}$ ▷ Construct mixed class vectors (Eq. 9)
  for each mini-batch $\{(i, \pi(i))\}_{i=1}^B \sim \tilde{\mathcal{I}}^\lambda$ **do**
    for $i = 1, 2, \ldots, B$ **do**
      $\tilde{x}_i \leftarrow \lambda\, x_i + (1-\lambda)\, x_{\pi(i)}$
      $\tilde{y}_i \leftarrow \boldsymbol{e}^{\mathcal{I}^2(c_i, c_{\pi(i)})}$ ▷ Construct mixed samples (Eq. 10)
    end for
    $\mathcal{L} \leftarrow \frac{1}{B}\sum_{i=1}^B \ell\big(\boldsymbol{W}^\lambda f_\theta(\tilde{x}_i),\ \tilde{y}_i\big)$ ▷ Update Model
    $\theta \leftarrow \theta - \eta\, \nabla_\theta \mathcal{L}$
  end for
end for

two singleton logits can be replaced by a hard label over the mixed label with a λ-interpolated class vector (see Section F.2 for evidence in real networks [23]).

Mixed-Singleton Classifier. Let $\boldsymbol{W} \in \mathbb{R}^{K\times p}$ be a classifier of singleton labels, which is a fully-connected layer. We define the Mixed-Singleton classifier (MS) as

$$\boldsymbol{W}^\lambda = [\lambda \boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_b]_{(a,b)\in\mathbb{K}^2}, \tag{9}$$

where p is the last-layer feature dimension and $\boldsymbol{W}^\lambda \in \mathbb{R}^{K^2\times p}$. We replace the singleton classifier with MS and perform Mixup with BMLS, where mixed-label samples $\tilde{\boldsymbol{x}}_i^\lambda$ and their one-hot vectors $\tilde{\boldsymbol{y}}_i^\lambda$ are defined as:

$$\begin{aligned} \tilde{\boldsymbol{x}}_i^\lambda &= \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_{\pi(i)}, \\ \tilde{\boldsymbol{y}}_i^\lambda &= \boldsymbol{e}^{\mathcal{I}^2(c_i, c_{\pi(i)})}, \forall (i, \pi(i)) \in \tilde{\mathcal{I}}^\lambda. \end{aligned} \tag{10}$$

$\mathcal{I}^2$ denotes the index pairs of $\mathbb{K}^2$, and $\mathcal{I}^2(a,b)$ gives the index number of $(a,b) \in \mathbb{K}^2$.

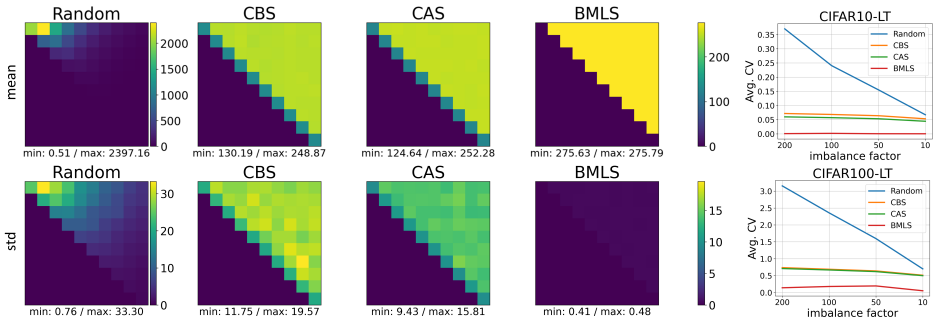
Theoretical Analysis on Mixed-Singleton Classifier. By replacing the classifier of Equation (2) with the Mixed-Singleton classifier, we obtain the UFM$_\lambda$ with Mixed-Singleton classifier (UFM$_\lambda$-MS; Equation (13) in Section F.1). The actual benefit of MS is the following result (Theorem G.5): in the balanced case, the stationarity condition of UFM$_\lambda$-MS reduces to

$$\bar{\boldsymbol{h}}_i - \boldsymbol{h}_{k,i} = -C_e \boldsymbol{w}_k, \quad k \in [K], \tag{11}$$

for a positive constant C_e, which is exactly the optimality condition of the balanced K-class UFM. That is, MS makes the learning target of the singleton classifier formally identical to the balanced case, so its global minimizer recovers the K-simplex ETF at the singleton level. Theorem 5.1 then establishes that balancing mixed labels remains effective in this setting.

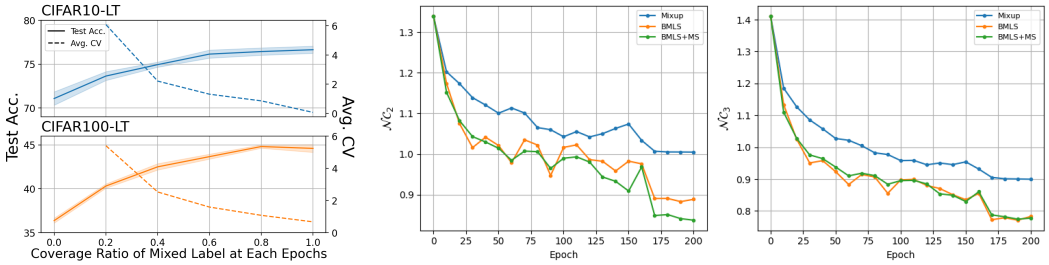
Theorem 5.1. *Assume $p \ge 2K$ and the loss function $\mathcal{L}$ is convex in the first argument. Let $\boldsymbol{X}^{\lambda\star}$ be any minimizer of the convex program of UFM$_\lambda$-MS (Equation (58)) with the mixed-label frequencies of Proposition 4.3, and define $\left(\boldsymbol{W}^{\lambda\star}, \boldsymbol{H}^{\lambda\star}\right)$ from $\boldsymbol{X}^{\lambda\star}$ as in Equation (7), where $\boldsymbol{P} \in \mathbb{R}^{p\times 2K^2}$ satisfies $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}_{2K^2}$, $\boldsymbol{W}^{\lambda\star} \in \mathbb{R}^{K^2\times p}$, and $\boldsymbol{H}^{\lambda\star} \in \mathbb{R}^{p\times N}$. Then $(\boldsymbol{W}^{\lambda\star}, \boldsymbol{H}^{\lambda\star})$ is a global minimizer of the mixed-label balanced UFM$_\lambda$-MS. (Proof in Section G.7; as in Proposition 4.3, $p \ge 2K$ suffices because $\boldsymbol{w}^{\lambda\star}$ is linearly dependent on $\boldsymbol{w}^\star$.)*

Remark 5.2 (Scope of MS). BMLS follows from Theorem 4.2 and Proposition 4.3 and does not rely on the λ-independence observation. MS is a secondary component whose theoretical warrant is partial: λ-independence is established within the UFM, and its transfer to deep networks is supported by prior empirical evidence [23] but not proved. We therefore present MS as an empirically validated design that the analysis motivates. Used standalone, MS helps when the per-epoch sample budget exceeds $|\mathbb{K}^<|$, and BMLS should be used alone otherwise; integrated into another Mixup-based method, both remain usable at large scale (Section 6.3).



(a) Epoch-wise label frequencies on CIFAR10-LT (imbalance factor 100) (b) $\overline{\text{CV}}$ (lower is better)

Figure 2: Epoch-wise label imbalance under the random sampler, the class-balanced sampler (CBS), the class-aware sampler (CAS), and BMLS, all followed by Mixup. (a) Each cell (a, b) is a label (diagonal: singleton labels; upper triangle: mixed labels), colored by the mean (top) and the standard deviation (bottom) of its frequency across epochs. CBS and CAS make the mean nearly uniform, but its epoch-wise standard deviation remains large; only BMLS makes both uniform. (b) Average coefficient of variation ($\overline{\text{CV}}$) of label frequencies across epochs for each imbalance factor; values close to 0 mean that every label appears equally often in every epoch.



(a) Mixed-label frequency control (b) NC metric measurement

Figure 3: **(a)** Mixed-label frequency control on CIFAR10/100-LT (imbalance factor 100). The coverage ratio is the fraction of mixed labels used for training in each epoch, rotated across epochs (e.g., at 0.6 on CIFAR100-LT, mixed labels are formed among 60 classes that change every epoch). Solid: test accuracy (higher is better); dashed: $\overline{\text{CV}}$ of label frequencies (lower is better). Accuracy rises as more mixed labels are covered each epoch. **(b)** NC metrics over all classes of CIFAR100-LT (imbalance factor 100) during training, for Mixup with a random sampler, BMLS, and BMLS_{MS}: $\mathcal{NC}_2$ (simplex ETF error) and $\mathcal{NC}_3$ (self-duality error), lower is better. Balancing mixed labels keeps the class means and class vectors closer to the simplex ETF (per-category results in Section I.1).

6 Experimental Results

In all tables, *imb* refers to the imbalance factor, *C10/100* represents the CIFAR10/100-LT datasets, *Clf.* refers to the classifier, and BMLS_{MS} denotes the method using both BMLS and MS; all experiments include Mixup unless otherwise specified, and best results are in bold. Implementation details and evaluation criteria are in Section H and Section I.6. Since all test sets are class-balanced, top-1 accuracy equals balanced accuracy (BAcc); we also report macro F1, which accounts for precision, and many/medium/few accuracy.

6.1 Empirical Validation

Epoch-wise Label Imbalance. To test Remark 4.4, we examine the mean and standard deviation of label frequencies across epochs under the random sampler, the class-balanced sampler (CBS) [18], the class-aware sampler (CAS) [19], and BMLS (Figure 2), measuring dispersion by the average coefficient of variation $\overline{\text{CV}} = \frac{1}{K}\sum_{k=1}^{K} \sigma_k/\mu_k$ [37]. The mean of label frequencies is almost balanced under all samplers, but epoch-wise balance is not: $\overline{\text{CV}}$ remains high for CBS and CAS and is near zero only for BMLS. When we control the fraction of mixed labels covered in each epoch, the more imbalanced the mixed labels are from epoch to epoch, the lower the performance (Figure 3a).

Table 1: Collapse geometry under mixed-label coverage control on CIFAR100-LT (imbalance factor 100; mean_{std} over five seeds). $\mathcal{NC}_2$/$\mathcal{NC}_3$: simplex ETF / self-duality error on the *Few* category (↓); Cos: mean cosine between *Few*-class vectors (ideal $-1/(K-1) \approx -0.0101$). Last row: Pearson r with coverage ($n = 25$, all $p < 10^{-8}$).

Coverage	$\mathcal{NC}_2$	$\mathcal{NC}_3$	Cos	F1	Acc.
0.2	$0.6026_{.0115}$	$0.5608_{.0124}$	$0.0102_{.0017}$	$36.25_{0.31}$	$40.36_{0.30}$
0.4	$0.5486_{.0064}$	$0.5042_{.0099}$	$0.0063_{.0022}$	$38.80_{0.65}$	$42.94_{0.57}$
0.6	$0.5327_{.0065}$	$0.4748_{.0054}$	$-0.0007_{.0018}$	$39.97_{0.35}$	$44.31_{0.36}$
0.8	$0.5091_{.0115}$	$0.4511_{.0050}$	$-0.0040_{.0013}$	$40.51_{0.46}$	$44.83_{0.54}$
1.0	$\mathbf{0.5021}_{.0071}$	$\mathbf{0.4375}_{.0046}$	$\mathbf{-0.0049}_{.0006}$	$\mathbf{40.77}_{0.51}$	$\mathbf{45.15}_{0.50}$
r	-0.920	-0.950	-0.935	$+0.886$	$+0.895$

Table 2: Per-pair coverage and the effect of MS. *Cov.*: samples seen per epoch divided by the number of mixed labels $|\mathbb{K}^<| = K(K-1)/2$. Δ: top-1 accuracy of BMLS_{MS} minus BMLS (CIFAR: imbalance factor 100).

Dataset	K	Cov.	Δ
CIFAR10-LT	10	275.7	+0.82
CIFAR100-LT	100	2.19	+2.42
Places-LT	365	2.78	+0.62
ImageNet-LT	1,000	0.23	-1.29
iNaturalist18	8,142	0.013	-10.38

Table 3: Experiments on CIFAR10/100-LT datasets with various imbalance factors. (*diff.*: the difference between Cyan bg. and Gray bg. . More references in Table 16)

		CIFAR10-LT				CIFAR100-LT			
Method	imb	200	100	50	10	200	100	50	10
Mixup		66.77	72.94	78.64	88.05	39.06	42.88	48.31	63.03
+LOM		70.17	76.63	81.15	89.24	39.61	44.24	49.99	63.90
+CAS		69.90	76.43	81.42	89.24	40.28	44.65	50.07	63.57
+BMLS_{MS}		74.70	79.67	83.46	88.51	41.71	47.62	52.74	64.47
	diff.	+7.93	+6.73	+4.82	+0.46	+2.65	+4.74	+4.43	+1.44
ETF+DR		71.58	76.82	81.25	87.59	41.20	45.07	50.71	63.08
BMLS+WETF_{MS}+CE		77.73	80.31	84.22	88.26	42.73	47.10	52.44	64.10
	diff.	+6.15	+3.49	+2.97	+0.67	+1.53	+2.03	+1.73	+1.02
Remix		69.58	75.15	80.41	88.61	41.03	44.95	50.19	63.45
+BMLS		73.95	80.10	83.92	88.62	39.95	46.34	51.53	64.42
+BMLS_{MS}		73.18	78.00	83.70	88.20	40.25	46.82	49.78	63.54
	diff.	+3.60	+2.85	+3.29	-0.41	-0.78	+1.87	-0.41	+0.09
DBN-mix		77.40	82.40	86.05	**91.01**	40.71	45.52	50.47	62.68
+BMLS_{MS}		**79.73**	**84.30**	**87.28**	90.93	**44.42**	**49.08**	**55.41**	**65.42**
	diff.	+2.33	+1.90	+1.23	-0.08	+3.71	+3.56	+4.94	+2.74

Mixed-label Coverage and Collapse Geometry. We also measure the collapse itself at each coverage level. On CIFAR100-LT, every quantity is monotone in coverage (Table 1): $\mathcal{NC}_2$ and $\mathcal{NC}_3$ on the *Few* category fall, the cosine between *Few*-class vectors moves from positive values, the signature of minority collapse, toward the simplex ETF value, and F1 and accuracy rise. All ten coverage–metric correlations on CIFAR10/100-LT are significant and in the predicted direction (Section I.2); CIFAR10-LT plateaus after coverage 0.6, since its 45 mixed labels are covered over training anyway. Across samplers, the random sampler leaves the *Few*-class vectors of CIFAR10-LT positively correlated (cosine $+0.0633$), whereas BMLS_{MS} reaches -0.1075 against the ideal -0.1111. This cosine loses resolution as K grows, so we use $\mathcal{NC}_2$ and $\mathcal{NC}_3$, which characterize the full frame, as the primary collapse measurements (Figure 3b).

The Effect of Balanced Mixed Label Sampler. Enforcing balance among mixed labels through BMLS keeps the classifier closer to the simplex ETF (Figure 3b; class-vector visualization and uniformity in Section I.1) and improves performance even with a standard FC classifier: BMLS alone improves over the random sampler from 72.94 to 78.85 on CIFAR10-LT and from 42.88 to 45.20 on CIFAR100-LT (imbalance factor 100; Table 5).

The Effect of Mixed-Singleton Classifier. Under BMLS, MS outperforms FC at every imbalance factor except the mild 10 (Table 20) and improves the *few* category by 4.10 on CIFAR10-LT and 7.43 on CIFAR100-LT (Table 5). With a random sampler MS is harmful (64.08 vs. 72.94 on CIFAR10-LT), the conditional profile the analysis predicts.

6.2 Standard Imbalanced Learning Benchmarks

Comparison of class-balanced samplers. We compare with Mixup, LOM [18], and CAS, the most similar to our approach, under identical settings. As shown in Table 3, our proposed method achieves the highest performance on CIFAR10/100-LT in all settings except the mildest imbalance (imbalance factor 10 on CIFAR10-LT). The ordering of methods is preserved under F1 (Table 14), so the gains do not come from trading precision for recall. The largest improvement is on the *few* category (Table 17).

Integration with ETF classifier, Remix, and DBN-mix. We apply our methods to (i) ETF+DR [21], which fixes the classifier to a simplex ETF; (ii) Remix [38], which re-balances the Mixup lambda according to class sample counts; and (iii) DBN-mix [39], which uses bilateral Mixup and a double-branch architecture (Section H). BMLS_{MS} improves ETF+DR and DBN-mix in almost all settings, by up to $+4.94$ (Table 3); gains are smaller and mixed for Remix, whose rebalancing of λ partially overlaps with our mechanism. Details are in Tables 23 and 26, and an integration with a superclass-aware method [40] is in Section I.4.

6.3 Large-scale Datasets and Per-pair Coverage

On large-scale datasets, BMLS alone remains competitive, whereas MS helps on Places-LT ($K = 365$) but degrades on ImageNet-LT and iNaturalist2018 (Table 7), so K itself is not the cause. What tracks the outcome is *per-pair coverage*, the number of samples seen per epoch per mixed label (Table 2): MS improves exactly where it exceeds one and degrades where it falls below one, more so the further below. This is the uncontrolled counterpart of the coverage sweep in Figure 3a, and restoring coverage restores MS: enlarging the ImageNet-LT training set by a factor of N raises the accuracy of BMLS_{MS} from 47.54 to 48.20 to 49.76 and its *few* accuracy from 24.86 to 25.27 to 28.81 (Table 8). When integrated with DBN-mix, BMLS and MS remain usable even at $K = 8{,}142$ (Table 4). They narrow the spread across categories, and BMLS_{MS} attains the best overall accuracy. However, it remains 6.85 points below DBN-mix on the *few* category, so at this scale the benefit lies in overall and cross-category balance rather than in tail recognition (Section I.5).

Table 4: Integration with DBN-mix on iNaturalist2018 (two-branch ResNet-50, 15 epochs for all methods; top-1 accuracy).

Method	many	med	few	all
DBN-mix	21.42	37.63	**42.78**	38.08
+BMLS	38.58	40.76	35.69	38.58
+BMLS_{MS}	**40.39**	**41.55**	35.93	**39.22**

6.4 Ablation Study

Applying both BMLS and MS yields the largest improvement (Table 5), including under extreme imbalance (Table 29), and BMLS_{MS} is robust to the Mixup α (Table 30) and mini-batch size (Figure 7) at 1.1–1.3$\times$ the training time (Table 32).

K^2 Classifier. MS adds no parameters (Table 32), whereas a genuine K^2 classifier has far more capacity, yet the latter performs dramatically worse, as each of its class vectors sees far fewer samples; the gain of MS thus comes from interpolating class vectors, not from capacity.

Table 5: Ablation on CIFAR10/100-LT (imbalance factor 100; *few* and *all* top-1 accuracy). K^2: a K^2-way classifier with independent class vectors. Full results in Table 21.

Sampler	Clf.	C10		C100	
		few	all	few	all
random	FC	50.99	72.94	8.77	42.88
BMLS	FC	70.53	78.85	14.37	45.20
random	MS	41.15	64.08	2.78	36.87
BMLS	K^2	25.93	39.01	1.21	8.72
BMLS	MS	**74.63**	**79.67**	**21.80**	**47.62**

7 Conclusion

We proved that minority collapse persists under Mixup unless mixed labels are balanced, which class-balanced samplers achieve only in expectation; BMLS, which balances them in every epoch, follows from this analysis, and MS recovers the balanced-case geometry once they are balanced. The collapse geometry responds monotonically to mixed-label coverage, and BMLS and MS improve minority classes and several Mixup-based methods. **Limitations.** Our strongest evidence is on small-to-medium class regimes such as CIFAR-LT; the theory relies on the UFM, an SDP relaxation, and an asymptotic imbalance (Section 4.3), and the transfer of λ-independence to deep networks is not proved. Used standalone, MS degrades when per-pair coverage falls below one, and BMLS treats all class pairs uniformly, which structure-aware pair selection could improve (Section B).

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2022R1A2C2012054, Development of AI for Canonicalized Expression of Trained Hypotheses by Resolving Ambiguity in Various Relation Levels of Representation Learning) (80%), Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)) (10%), and the InnoCORE program of the Ministry of Science and ICT(26-InnoCORE-01) (10%).

References

[1] Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 915–922, 2014.

[2] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 864–873, 2016.

[3] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

[4] Yin Cui, Yang Song, Chen Sun, Andrew G. Howard, and Serge J. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4109–4118, 2018.

[5] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2537–2546, 2019.

[6] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6908–6917, 2022.

[7] Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

[8] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

[9] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447. PMLR, 2019.

[10] Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How re-sampling helps for long-tail learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[11] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16489–16498, 2021.

[12] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. ACE: Ally complementary experts for solving long-tailed recognition in one-shot. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 112–121, 2021.

[13] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[14] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision (ECCV)*, volume 12350 of *Lecture Notes in Computer Science*, pages 247–263. Springer, 2020.

[15] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019.

[16] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9268–9277, 2019.

[17] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020.

[18] Shaoyu Zhang, Chen Chen, Xiujuan Zhang, and Silong Peng. Label-occurrence-balanced mixup for long-tailed recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3224–3228, 2022.

[19] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, volume 9911 of *Lecture Notes in Computer Science*, pages 467–482. Springer, 2016.

[20] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[21] Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[22] Xuantong Liu, Jianfeng Zhang, Tianyang Hu, He Cao, Yuan Yao, and Lujia Pan. Inducing neural collapse in deep long-tailed learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 11534–11544. PMLR, 2023.

[23] Quinn LeBlanc Fisher, Haoming Meng, and Vardan Papyan. Pushing boundaries: Mixup's influence on neural collapse. In *International Conference on Learning Representations (ICLR)*, 2024.

[24] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021.

[25] Zhengzhuo Xu, Zenghao Chai, and Chun Yuan. Towards calibrated model for long-tailed visual recognition from prior perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

[26] Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2):11, 2022.

[27] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 29820–29834, 2021.

[28] Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27225–27238, 2022.

[29] Hien Dang, Tho Tran, Stanley Osher, Hung Tran-The, Nhat Ho, and Tan Nguyen. Neural collapse in deep linear networks: From balanced to imbalanced data. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 6873–6947, 2023.

[30] Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 21478–21505, 2022.

[31] Hien Dang, Tho Tran Huu, Tan Minh Nguyen, and Nhat Ho. Neural collapse for cross-entropy class-imbalanced learning with unconstrained ReLU features model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10017–10040. PMLR, 2024.

[32] Wanli Hong and Shuyang Ling. Neural collapse for unconstrained feature model under cross-entropy loss with imbalanced data. *Journal of Machine Learning Research*, 25(192):1–48, 2024.

[33] Hoyong Kim and Kangil Kim. Fixed non-negative orthogonal classifier: Inducing zero-mean neural collapse with feature dimension separation. In *International Conference on Learning Representations (ICLR)*, 2024.

[34] Jos F. Sturm and Shuzhong Zhang. On cones of nonnegative quadratic functions. *Mathematics of Operations Research*, 28(2):246–267, 2003.

[35] Mahyar Fazlyab, Manfred Morari, and George J. Pappas. Probabilistic verification and reachability analysis of neural networks via semidefinite programming. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2726–2731, 2019.

[36] Godai Azuma, Sunyoung Kim, and Makoto Yamashita. Tight semidefinite relaxations for verifying robustness of neural networks. *arXiv preprint arXiv:2504.09934*, 2025.

[37] Yadolah Dodge. *Coefficient of Variation*, pages 95–96. Springer New York, New York, NY, 2008.

[38] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced mixup. In *Computer Vision – ECCV 2020 Workshops*, pages 95–110. Springer International Publishing, 2020.

[39] Jae Soon Baik, In Young Yoon, and Jun Won Choi. DBN-Mix: Training dual branch network using bilateral mixup augmentation for long-tailed visual recognition. *Pattern Recognition*, 147:110107, 2024.

[40] Sunguk Jang, Jinwoo Jeon, and Byung-Jun Lee. SCAD: Super-class-aware debiasing for long-tailed semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2026.

[41] Youngseok Yoon, Sangwoo Hong, Hyungjun Joo, Yao Qin, Haewon Jeong, and Jungwoo Lee. Mix from failure: Confusion-pairing mixup for long-tailed recognition. *arXiv preprint arXiv:2411.07621*, 2024.

[42] Kisoo Kwon, Kuhwan Jeong, Sanghyun Park, Sangha Park, Hoshik Lee, Seung-Yeon Kwak, Sungmin Kim, and Kyunghyun Cho. ExtraMix: Extrapolatable data augmentation for regression using generative models, 2023. OpenReview preprint.

[43] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019.

[44] Caidan Zhao and Yang Lei. Intra-class cutmix for unbalanced data augmentation. In *Proceedings of the 2021 13th International Conference on Machine Learning and Computing*, pages 246–251. Association for Computing Machinery, 2021.

[45] Haolin Pan, Yong Guo, Mianjie Yu, and Jian Chen. Enhanced long-tailed recognition with contrastive CutMix augmentation. *IEEE Transactions on Image Processing*, 33:4215–4230, 2024.

[46] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6887–6896, 2022.

[47] Jintong Gao, He Zhao, Zhuo Li, and Dandan Guo. Enhancing minority classes by mixing: An adaptative optimal transport approach for long-tailed classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[48] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9719–9728, 2020.

[49] Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 973–978, 2001.

[50] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 872–881. PMLR, 2019.

[51] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 4175–4186. Curran Associates, Inc., 2020.

[52] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13896–13905, 2020.

[53] Anubha Kabra, Ayush Chopra, Nikaash Puri, Pinkesh Badjatiya, Sukriti Verma, Piyush Kumar Gupta, and K Balaji. MixBoost: Synthetic oversampling with boosted mixup for handling extreme imbalance. *arXiv preprint arXiv:2009.01571*, 2020.

[54] Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J. Su. An unconstrained layer-peeled perspective on neural collapse. In *International Conference on Learning Representations (ICLR)*, 2022.

[55] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[56] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[57] Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022.

[58] Weinan E and Stephan Wojtowytsch. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference (MSML)*, volume 145 of *Proceedings of Machine Learning Research*, pages 270–290, 2022.

[59] X. Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations (ICLR)*, 2022.

[60] Tomaso A. Poggio and Qianli Liao. Explicit regularization and implicit bias in deep network classifiers trained with the square loss. *arXiv preprint arXiv:2101.00072*, 2021.

[61] Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under MSE loss: Global optimality with unconstrained features. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 27179–27202, 2022.

[62] Peter Súkeník, Marco Mondelli, and Christoph H. Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[63] Akshay Rangamani and Andrzej Banburski-Fahey. Neural collapse in deep homogeneous classifiers and the role of weight decay. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4243–4247, 2022.

[64] Can Yaras, Peng Wang, Zhihui Zhu, Laura Balzano, and Qing Qu. Neural collapse with normalized features: A geometric analysis over the Riemannian manifold. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 11547–11560, 2022.

[65] Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[66] Tomer Galanti, András György, and Marcus Hutter. Improved generalization bounds for transfer learning via neural collapse. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML*, 2022.

[67] Jarrod Haas, William Yolland, and Bernhard T. Rabus. Linking neural collapse and L2 normalization with improved out-of-distribution detection in deep neural networks. *Transactions on Machine Learning Research (TMLR)*, 2023.

[68] Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. NECO: NEural collapse based out-of-distribution detection. In *International Conference on Learning Representations (ICLR)*, 2024.

[69] Tom Tirer, Haoxiang Huang, and Jonathan Niles-Weed. Perturbation analysis of neural collapse. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 34301–34329, 2023.

[70] Ido Ben-Shaul and Shai Dekel. Nearest class-center simplification through intermediate layers. In *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, volume 196 of *Proceedings of Machine Learning Research*, pages 37–47, 2022.

[71] Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

[72] Yongyi Yang, Jacob Steinhardt, and Wei Hu. Are neurons actually collapsed? On the fine-grained structure in neural representations. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 39453–39487, 2023.

[73] Federico Pernici, Matteo Bruni, Claudio Baecchi, and Alberto Del Bimbo. Maximally compact and separated features with regular polytope networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 46–53, 2019.

[74] Federico Pernici, Matteo Bruni, Claudio Baecchi, and Alberto Del Bimbo. Regular polytope networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4373–4387, 2022.

[75] Elad Hoffer, Itay Hubara, and Daniel Soudry. Fix your classifier: the marginal value of training the last weight layer. In *International Conference on Learning Representations (ICLR)*, 2018.

[76] Tong Liang and Jim Davis. Inducing neural collapse to a fixed hierarchy-aware frame for reducing mistake severity. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1443–1452, 2023.

[77] Liang Xie, Yibo Yang, Deng Cai, and Xiaofei He. Neural collapse inspired attraction–repulsion-balanced loss for imbalanced learning. *Neurocomputing*, 527:60–70, 2023.

[78] Hongren Yan, Yuhua Qian, Furong Peng, Jiachen Luo, Zheqing Zhu, and Feijiang Li. Neural collapse to multiple centers for imbalanced data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

[79] Donghao Li, Yang Cao, and Yuan Yao. NeuroMixGDP: A neural collapse-inspired random mixup for private data release. In *Conference on Parsimony and Learning (CPAL)*, volume 234 of *Proceedings of Machine Learning Research*, pages 480–514. PMLR, 2024.

[80] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, 2016.

[81] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[82] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[83] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[84] Sunil Thulasidasan, Gopinath Chennupati, Jeff Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 13888–13899, 2019.

[85] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, and James Zou. When and how mixup improves calibration. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 26135–26160, 2022.

[86] Hongyu Guo, Yongyi Mao, and Richong Zhang. MixUp as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019.

[87] Arslan Chaudhry, Aditya Krishna Menon, Andreas Veit, Sadeep Jayasumana, Srikumar Ramalingam, and Sanjiv Kumar. When does mixup promote local linearity in learned representations? In *NeurIPS 2022 Workshop on Interpolation Regularizers and Beyond (INTERPOLATE)*, 2022. arXiv:2210.16413.

[88] Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. On mixup regularization. *Journal of Machine Learning Research*, 23(325):1–31, 2022.

[89] Chanwoo Park, Sangdoo Yun, and Sanghyuk Chun. A unified analysis of mixed sample data augmentation: A loss function perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[90] Muthu Chidambaram, Xiang Wang, Yuzheng Hu, Chenwei Wu, and Rong Ge. Towards understanding the data dependency of mixup-style training. In *International Conference on Learning Representations (ICLR)*, 2022.

[91] Muthu Chidambaram, Xiang Wang, Chenwei Wu, and Rong Ge. Provably learning diverse features in multi-view data with midpoint mixup. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5563–5599, 2023.

[92] Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. The benefits of mixup for feature learning. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 43423–43479, 2023.

[93] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle Mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 5275–5285, 2020.

[94] Pengyu Li, Xiao Li, Yutong Wang, and Qing Qu. Neural collapse in multi-label learning with pick-all-label loss. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 28060–28094. PMLR, 2024.

[95] Wanli Hong and Shuyang Ling. Beyond unconstrained features: Neural collapse for shallow neural networks with general data. *Journal of Machine Learning Research*, 27(82):1–46, 2026.

[96] Tolga Ergen and Mert Pilanci. Revealing the structure of deep neural networks via convex duality. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3004–3014. PMLR, 2021.

[97] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[98] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

[99] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778, 2018.

[100] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[101] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 2020.

[102] Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, 2023.

[103] Md Yousuf Harun, Jhair Gallardo, and Christopher Kanan. Controlling neural collapse enhances out-of-distribution detection and transfer learning. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 22149–22176, 2025.

Table of Contents

A Notation **18**

B Limitations and Future Work **18**

C Related Work **19**

C.1 Mixup-based Methods 19

C.2 Class-balanced Methods 20

C.3 Neural Collapse Inspired Methods 20

D Convex Optimization Program **22**

E SDP Constraint Relaxation **23**

F Theoretical Model and Discussion about Mixed-Singleton classifier **23**

F.1 UFM with Mixed-Singleton classifier 23

F.2 Discussion about Mixup lambda-independence 24

G Proofs **24**

G.1 Proof Sketch 24

G.2 Proof of Lemma 4.1 24

G.3 Proof of Theorem 4.2 26

G.4 Proof of Proposition 4.3 34

G.5 Proof of Theorem G.5 36

G.6 Proof of Proposition G.7 39

G.7 Proof of Theorem 5.1 40

H Experimental Setup **44**

I Additional Experimental Results **46**

I.1 Validation of BMLS 46

I.2 Collapse Geometry under Mixed-label Coverage Control 48

I.3 F1 Score and Balanced Accuracy 49

I.4 Integration with a Superclass-aware Method 49

I.5 Integration with DBN-mix on iNaturalist2018 50

I.6 Category-wise Accuracy 50

I.7 Comparison Experiments for ETF+DR 54

I.8 Comparison Experiments for Remix 55

I.9 Ablation Study including K2 Classifier 56

I.10 An Empirical Study on Mixup Alpha 57

I.11 Ablation study on mini-batch size 58

I.12 Computational cost 58

A Notation

Table 6: Notation used throughout the paper.

Symbol	Description
Data and labels	
N, K	Number of training samples and number of singleton classes
$\mathcal{X} = [(\boldsymbol{x}_i, c_i)]_{i=1}^N$	Training set; $c_i \in [K]$ is the class of sample $\boldsymbol{x}_i$
$\boldsymbol{y}_i = \boldsymbol{e}^{(c_i)}$	One-hot (singleton) label of $\boldsymbol{x}_i$
singleton label	An ordinary one-hot label of a single class
mixed label (a, b)	The pair of classes whose samples are interpolated by Mixup
$\mathbb{K}^2$	Set of all mixed labels, $\{(a, b) \mid 1 \le a, b \le K\}$, of size K^2
$\mathbb{K}^<$	Mixed labels restricted to $a < b$, of size $K(K-1)/2$ (Section 5.1)
n_k, $n_{(a,b)}$	Number of samples of singleton class k and of mixed label (a, b)
Mixup	
$\mathcal{I}$, π	Sample indices and a random permutation of them
$\mathcal{I}^\lambda$, $\mathcal{I}^\lambda_{(a,b)}$	Index pairs $(i, \pi(i))$ of mixed samples, and those with $(c_i, c_{\pi(i)}) = (a, b)$
$\tilde{\mathcal{I}}^\lambda$	Index pairs pre-defined by BMLS such that $n_{(a,b)}$ is equal for all mixed labels
λ, α	Mixup ratio $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$ and its hyperparameter
$\boldsymbol{x}_i^\lambda$, $\boldsymbol{y}_i^\lambda$	Mixed input and mixed (soft) label (Equation (1))
$\mathcal{I}^2(a, b)$	Index of mixed label (a, b) in $\mathbb{K}^2$, used as a hard label by MS
Unconstrained features model	
p	Dimension of last-layer features
$\boldsymbol{h}_{k,i}$, $\mathbb{S}_k$	Last-layer feature of the i-th sample of class k, and the set of such features
$\boldsymbol{h}^\lambda_{(a,b),i}$, $\mathbb{S}^\lambda_{(a,b)}$	Mixed feature of the i-th sample of mixed label (a, b), and the set of such features
$\boldsymbol{h}^\lambda_k$, $\bar{\boldsymbol{h}}$	Mean feature of mixed label k, and the global mean feature
$\boldsymbol{W} = [\boldsymbol{w}_k]_{k=1}^K \in \mathbb{R}^{K \times p}$	Singleton classifier; the row $\boldsymbol{w}_k$ is the class vector of class k
$\boldsymbol{W}^\lambda = [\lambda \boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_b]_{(a,b)}$	Mixed-Singleton classifier (Equation (9)); no additional parameters
E_W, E_H	Norm budgets of class vectors and features
$\mathcal{L}$, $\boldsymbol{z}$	Cross-entropy loss with softmax, and the logit $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{h}$
$\boldsymbol{X}$, $\boldsymbol{X}(i, j)$	Gram matrix of $[\boldsymbol{h}^\lambda_{(1,1)}, \ldots, \boldsymbol{h}^\lambda_{(K,K)}, \boldsymbol{W}^\top]$ in the SDP (Equation (4)), and its (i, j)-th entry
$\mathbb{K}_A$, $\mathbb{K}_B$	Major and minor groups of singleton classes, of sizes K_A and K_B
n_A, n_B, R	Samples per major and minor class, and the imbalance factor $R = n_A/n_B$
$\mathbb{K}^2_A$, $\mathbb{K}^2_B$	Mixed labels formed within the major group and within the minor group
w_r	Oversampling weight of minor mixed labels (Equation (6))
Abbreviations	
BMLS, MS	Balanced Mixed Label Sampler, Mixed-Singleton classifier
NC, ETF	Neural Collapse, (simplex) equiangular tight frame
UFM, UFM$_\lambda$, UFM$_\lambda$-MS	Unconstrained features model, its Mixup variant, and its variant with MS
SDP, TPT	Semi-definite program, terminal phase of training
CAS	Class-aware sampler [19]
CBS, LOM	Class-balanced sampler, Label-Occurrence-Balanced Mixup [18]
$\overline{\mathrm{CV}}$	Average coefficient of variation of epoch-wise label frequencies
$\mathcal{NC}_1, \mathcal{NC}_2, \mathcal{NC}_3$	Within-class variability, simplex ETF error, self-duality error

B Limitations and Future Work

Scope of the evidence. The strongest evidence for our claims is on small-to-medium class regimes such as CIFAR10/100-LT, where the pair space can be covered within each epoch. On large-class datasets, BMLS alone remains competitive with other class-balanced samplers (Table 7), whereas BMLS$_{\mathrm{MS}}$ is less reliable.

Scalability and the MS classifier. The number of mixed labels grows quadratically with K, while we match the number of samples learned per epoch to that of a random sampler for a fair comparison (e.g., on iNaturalist2018, 437,513 images per epoch against $K(K-1)/2 = 33{,}142{,}011$ mixed labels

Table 7: Experiments on large datasets. (*:use pre-trained model) (More detailed results in Table 18 and Table 19)

Method	Places-LT	Places-LT*	ImageNet-LT	iNaturalist18
random	22.06	25.90	45.19	64.62
CBS	24.79	37.32	47.49	67.06
CAS	24.26	37.44	47.31	**67.55**
BMLS	27.33	37.39	**48.83**	66.98
BMLS_{MS}	**27.95**	**37.81**	47.54	56.60

Table 8: Scale-up experiments of BMLS_{MS} (BMLS with the MS classifier) on ImageNet-LT. The results are the mean of three repeated experiments with random seeds. (N: the scaling factor with respect to the size of the dataset used for training.) The $\times 1$ row coincides with BMLS_{MS} in Table 19.

Method	N	many	med	few	all
BMLS_{MS}	$\times 1$	$59.03_{0.89}$	$45.87_{1.01}$	$24.86_{0.92}$	$47.54_{0.94}$
	$\times 2$	$59.65_{0.89}$	$46.67_{0.09}$	$25.27_{0.20}$	$48.20_{0.08}$
	$\times 4$	$\mathbf{60.54}_{0.89}$	$\mathbf{48.32}_{0.49}$	$\mathbf{28.81}_{0.20}$	$\mathbf{49.76}_{0.48}$

with $K = 8{,}142$). As shown in Section 6.3, the operative variable is per-pair coverage rather than K itself. When it falls below one, the MS classifier degrades, by 1.29 points on ImageNet-LT and by 10.38 points on iNaturalist2018 (Table 7). We regard this as a limitation of the MS component, not only a general scalability issue: in the standalone configuration MS should be enabled only when the per-epoch sample budget exceeds the number of mixed labels, and BMLS should be used alone otherwise. Increasing the number of training samples restores MS (Table 8), but excessive oversampling may introduce overfitting [7], and finding an appropriate trade-off is left for future work.

Uniform balancing of class pairs. BMLS treats all class pairs as equally worth balancing. Semantically unrelated pairs are not a defect of Mixup but part of its mechanism, since interpolations between distant classes extend the support of the training distribution; nevertheless, in large-scale long-tailed recognition many pairs may be uninformative for the classifier. A more scalable strategy could group or prioritize pairs by superclass structure, semantic similarity, or confusion patterns. Our analysis constrains such schemes: Proposition G.7 shows that the simplex ETF over singleton classes survives restricting the pair space, provided that for every class i there exist j and k such that $\{i, j\}$, $\{j, k\}$, and $\{i, k\}$ are all retained (Equation (56)). Moreover, Proposition 4.3 already admits non-uniform weighting of mixed labels through w_r. We leave the design of such pruning schemes to future work.

Simplifications in the theory. The assumptions of our theoretical model and their gaps with respect to practice are discussed in Section 4.3. In particular, λ-independence, the premise of MS, is established within the UFM, and its transfer to deep networks is supported by prior empirical evidence [23] but not proved.

Integration with other methods. We integrated BMLS and MS with ETF+DR, Remix, DBN-mix, and SCAD. The benefit is not uniform: it is smaller and mixed for Remix, whose rebalancing of λ partially overlaps with our mechanism, and on iNaturalist2018 the integration with DBN-mix improves overall and cross-category balance while lowering *few*-category accuracy (Table 4). A detailed investigation of how BMLS and MS interact with other Mixup-based methods is left for future work.

C Related Work

C.1 Mixup-based Methods

Data augmentation. Mixup [8] generates mixed-label samples by interpolating between input samples, extending the training distribution support. Manifold Mixup [9] applies this interpolation to intermediate layers, regularizing the network by encouraging less confident predictions. CP-Mix, or Confusion-Pairing Mixup [41], augments samples based on confusion pairs, addressing data deficiency by enhancing the model's ability to distinguish frequently misclassified class pairs. ExtraMix [42] introduces a mixup technique capable of extrapolation, broadening both feature and

label distributions, which minimizes label imbalance more effectively than traditional interpolation methods. CutMix [43–45] focuses on mixed-label sample generation by cutting and pasting image patches, creating a regional dropout effect. CMO [46] extends this idea by pasting minority-class images onto majority-class backgrounds, enriching minority-class samples with context from majority-class images. OTMix [47] improves upon this by using Optimal Transport to adaptively combine majority-class backgrounds with minority-class foregrounds, ensuring semantically reasonable mixed images.

Architecture. BBN [48] and SBN-Mix/DBN-Mix [39] utilize different architectures to enhance both representation and classifier learning. These methods incorporate bilateral mixup or decoupling strategies to optimize performance for imbalanced datasets. OLTR [5] uses dynamic meta-embedding and modulated attention to map images into a feature space that respects both closed-world classification and the novelty of the open world, improving the generalization of imbalanced datasets.

Calibration or two-stage. UniMix [25] balances class distributions by introducing a novel mixing factor and sampler that favors the minority class. MiSLAS [11] decouples representation and classifier learning, improving both calibration and performance in imbalanced-data scenarios.

While many attempts have been made to address the challenges of imbalanced learning using Mixup, including data augmentation, architecture improvements, and calibration methods, no research has specifically focused on the balance of mixed labels in such contexts.

C.2 Class-balanced Methods

Re-balance. Remix [38] applies a higher mixup ratio to minority classes, rebalancing the data without re-sampling. Re-weighting [49, 50, 16] adjusts the loss function by tuning class weights, with methods like Balanced SoftMax [51] explicitly considering label distribution shifts during optimization. Logit Adj [24] adjusts logits based on label frequencies, promoting a larger margin between rare positive and dominant negative labels. τ-Norm [17] normalizes classifier weight norms according to class size, rebalancing decision boundaries. LDAM loss [15] improves generalization by replacing standard cross-entropy with a margin-based approach tailored to handle imbalanced datasets. cRT [17] re-trains the classifier using class-balanced sampling, improving the model's generalization ability. LWS [17] focuses on re-scaling classifier weights to ensure a balanced learning process for imbalanced datasets.

Re-/Over-Sampling. M2M [52] augments minority classes by translating samples from majority classes, enhancing generalization for minority-class features. MixBoost [53] iteratively selects and combines majority- and minority-class instances to create hybrid samples, improving model performance. The Meta Sampler [51], built on Balanced SoftMax, adapts the sampling rate through meta-learning to alleviate over-balancing issues. CB Sampling [17] ensures that each class has an equal probability of being selected, balancing the dataset during training. Class-Aware Sampler (CAS) [19] is a more specific form of CB Sampling that explicitly ensures class-frequency balance in each mini-batch. Label-Occurrence Mixup (LOM) [18] uses two CB samplers to draw input pairs independently. CSA [10] generates diverse training images for tail classes by maintaining a context bank assembled from head-class images.

Various class-balanced samplers have been proposed, yet no research has specifically focused on the balance of mixed labels. Additionally, while methods such as Logit Adjustment and UniMix have concentrated on singleton labels, they did not interpret mixed labels as singletons.

C.3 Neural Collapse Inspired Methods

Balanced Settings. Neural Collapse (NC) [20] is a geometric phenomenon observed at the terminal phase of training (TPT) of deep neural networks trained on balanced datasets. At the TPT, four simultaneous properties emerge: ($\mathcal{NC}_1$) within-class variability collapse, where the features of each class converge to their class mean; ($\mathcal{NC}_2$) convergence of the class means to the vertices of a simplex equiangular tight frame (ETF), a structure that is simultaneously equiangular and equinorm; ($\mathcal{NC}_3$) convergence of the linear classifier weights to the same ETF; and ($\mathcal{NC}_4$) convergence of the decision rule to the nearest-class-mean classifier. To understand how this geometry emerges, the unconstrained feature model (UFM) [26, 27] and the layer-peeled model (LPM) [7, 54] have been proposed as tractable surrogates, treating the last-layer features as free optimization variables alongside the classifier weights. The UFM assumption is theoretically grounded in the universal

approximation theorem for over-parameterized networks [55, 56]: because a sufficiently expressive feature map can realize essentially arbitrary feature configurations, replacing the feature map with unconstrained vectors is a principled simplification. Under both CE loss [57, 58, 54] and MSE loss [59, 60, 27, 61], the global minimizers of the regularized empirical risk in the UFM provably satisfy NC1–NC3. The optimization landscape is benign [27, 61] in the sense that all critical points that are not global minimizers are strict saddle points. Further theoretical extensions study NC under deep linear networks [29, 30, 62], the dynamics of gradient descent or gradient flow converging to the NC configuration [54, 63], and NC under normalized features on a Riemannian manifold [64]. Empirically, NC has been connected to generalization and transfer learning [65, 66], to out-of-distribution detection [67, 68], to perturbation analysis [69], and to intermediate layer geometry [70]. Several works have questioned whether NC is a reliable proxy for generalization [71, 72]. In balanced settings, a line of work exploits the ETF geometry to design principled fixed-classifier training strategies, fixing the last-layer weights to a simplex ETF or other symmetric frames [21, 73–76] and training only the feature extractor.

Imbalanced Settings. When training data are class-imbalanced, the NC structure is disrupted. Fang et al. [7] first documented the *minority collapse* phenomenon via the LPM: the class vectors of minority classes collapse toward each other as the imbalance ratio grows, severely degrading minority-class accuracy. Thrampoulidis et al. [28] extended the UFM analysis to imbalanced data using an unconstrained-feature SVM, deriving the Simplex-Encoded-Labels Interpolation (SELI) geometry that characterizes the structure of global minimizers and generalizes the balanced ETF. Dang et al. [29] studied NC in deep linear networks from balanced to imbalanced data under MSE and CE losses, while Dang et al. [31] generalized NC to imbalanced regimes under cross-entropy loss with an unconstrained ReLU feature model, proving that class means converge to orthogonal vectors of different lengths that scale with the square root of the class-sample sizes. Hong and Ling [32] provided a complete characterization of NC under cross-entropy loss in the UFM for imbalanced data, precisely identifying the sharp threshold for minority collapse and showing that pairwise angles between class means depend on class sample sizes. On the practical side, NC-inspired methods have been developed to mitigate minority collapse: Yang et al. [21] fix the classifier to the simplex ETF; Xie et al. [77] propose the Attraction-Repulsion-Balanced (ARB) loss; and Liu et al. [22] introduce explicit feature regularization terms that plug into existing long-tail methods. Yan et al. [78] extend NC to Neural Collapse to Multiple Centers (NCMC), allowing each class to collapse to multiple per-class prototypes to better accommodate minority classes under a cosine regression loss with a fixed multi-center frame.

Neural Collapse under Mixup. NeuroMixGDP [79] draws on the ETF geometry of NC to design a differentially private Mixup scheme that mixes last-layer features rather than raw inputs, exploiting the simplex structure to release noisy mixed features with improved privacy (utility trade-offs). Fisher et al. [23] provide the first systematic empirical and theoretical investigation of what Mixup does to last-layer geometry. Their study, spanning multiple architectures (WideResNet-40-10 [80], ViT-B [81]) and datasets (FashionMNIST [82], CIFAR-10/100 [83]), reveals that: (i) same-class mixed activations form a simplex ETF aligned with the respective classifier, regardless of the mixing coefficient λ; (ii) different-class activations delineate channels along the decision boundary, with activation density increasing as $\lambda \to 0.5$; and (iii) regular Mixup promotes manifold-mixup-like behaviour [9] in earlier layers as evidenced by the layer-wise CLS-token trajectory in ViT. A theoretical characterization under the UFM—assuming the classifier is a simplex ETF at optimality, which is empirically verified—shows that same-class features are λ-independent and ETF-aligned, while different-class features are dynamic linear combinations of the relevant classifier rows [23]. The paper also demonstrates that when the classifier is fixed as a simplex ETF, as in [21], the empirical features approach the theoretical optimum more closely without degrading test accuracy. This distinctive geometric configuration explains Mixup's calibration benefit [84, 85]: as λ approaches 0.5, the model becomes less confident and more prone to misclassification, aligning softmax probabilities with the true mixed label. Earlier work has sought to understand Mixup's success from complementary perspectives: Guo et al. [86] identified manifold intrusion as a limitation; Chaudhry et al. [87] analysed the linearity of learned representations; Carratino et al. [88] interpret Mixup as an empirical risk minimizer with transformed data; Park et al. [89] offer a unified analysis of mixed-sample augmentation from a loss perspective; Chidambaram et al. [90, 91] study the data dependency and feature-learning properties of Mixup training; and Zou et al. [92] analyse the benefits of Mixup for feature learning. Additional Mixup variants have been proposed, including PuzzleMix [93] and other methods [86]. Li et al. [94] generalize NC to the multi-label setting via the pick-all-label loss, establishing multi-label NC (M-lab

NC) under the UFM: while single-label instance features still form a simplex ETF, multi-label instance features emerge as scaled tag-wise averages of the associated single-label prototypes.

Neural Collapse under Mixup in Imbalanced Learning. To alleviate minority collapse, Yang et al. [21] assumed that the classifier is fixed to the K-simplex ETF and proved that the UFM with this fixed classifier satisfies the NC properties. Furthermore, combining the fixed ETF classifier with Mixup training substantially improves model performance in imbalanced settings. Building on these theorems, Fisher et al. [23] proved that Mixup also satisfies NC properties for both same-class and different-class activations when the classifier is a simplex ETF, providing the first UFM-grounded explanation of Mixup's geometric effect. The theoretical justification for applying the UFM in this analysis is further supported by Hong and Ling [95]: their Theorem 2.1 shows that the global minimizer of the regularized empirical risk under the unconstrained positive feature model already exhibits the same ETF geometry that the fixed-classifier approach seeks to impose, confirming that the UFM analysis in [21, 23] is grounded in the expressive power of practical architectures. However, Yang et al. [21] and Fisher et al. [23] did not consider the minority collapse arising from the *frequency of mixed labels* in the UFM with a learnable classifier. The only related finding comes from M-lab NC [94], which observes that even when multi-label samples are imbalanced, NC occurs at the singleton-class level as long as singleton-label samples are balanced, with multi-label classes emerging as combinations of singletons. However, whether the balance of input samples still holds for mixed labels under Mixup remains unclear. This motivates our central research question: *Could the balance of mixed labels be a critical factor in minority collapse?*

D Convex Optimization Program

To begin with, defining $\boldsymbol{h}_k^\lambda = \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{h}_{k,i}^\lambda$ as the feature mean of the $\mathbb{S}_k^\lambda$ where $k \in \mathbb{K}^2$, we introduce a new decision variable $\boldsymbol{X} = [\boldsymbol{h}_{(1,1)}^\lambda, \boldsymbol{h}_{(1,2)}^\lambda, \dots, \boldsymbol{h}_{(K,K)}^\lambda, \boldsymbol{W}^\top]^\top[\boldsymbol{h}_{(1,1)}^\lambda, \boldsymbol{h}_{(1,2)}^\lambda, \dots, \boldsymbol{h}_{(K,K)}^\lambda, \boldsymbol{W}^\top] \in \mathbb{R}^{(K^2+K)\times(K^2+K)}$. By definition, $\boldsymbol{X}$ is positive semi-definite and satisfies

$$\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}(k,k) = \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\|\boldsymbol{h}_k^\lambda\|^2 \overset{a}{\leq} \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}_{k,i}^\lambda\right\|^2 \leq E_H$$

and

$$\frac{1}{K}\sum_{k=K^2+1}^{K^2+K}\boldsymbol{X}(k,k) = \frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2 \leq E_W,$$

where $\overset{a}{\leq}$ follows from the Cauchy-Schwarz inequality. Thus, we consider the following semi-definite programming problem:

$$\begin{aligned}
\min_{\boldsymbol{X}\in\mathbb{R}^{(K^2+K)\times(K^2+K)}} \quad & \sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\mathcal{L}(\boldsymbol{z}(k)^\lambda, \boldsymbol{y}_k^\lambda) \\
\text{s.t. } & \boldsymbol{X} \succeq 0, \\
& \frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}(k,k) \leq E_H,\ \frac{1}{K}\sum_{k=K^2+1}^{K^2+K}\boldsymbol{X}(k,k) \leq E_W, \\
& \text{for all } 1 \leq k \leq K^2, \\
& \boldsymbol{z}_k = \left[\boldsymbol{X}(k,K^2+1), \boldsymbol{X}(k,K^2+2), \dots, \boldsymbol{X}(k,K^2+K)\right]^\top.
\end{aligned} \tag{12}$$

When $\mathcal{L}$ is the cross-entropy loss with softmax function,

$$\mathcal{L}(\boldsymbol{z}^\lambda(k), \boldsymbol{y}_k^\lambda) = -\lambda\log\left(\frac{\exp(\boldsymbol{z}^\lambda(a))}{\sum_{k'=1}^{K}\exp(\boldsymbol{z}^\lambda(k'))}\right) - (1-\lambda)\log\left(\frac{\exp(\boldsymbol{z}^\lambda(b))}{\sum_{k'=1}^{K}\exp(\boldsymbol{z}^\lambda(k'))}\right),$$

where $\boldsymbol{z}^\lambda(k')$ denotes the k'-th entry of the logit $\boldsymbol{z}_i^\lambda = \boldsymbol{W}\boldsymbol{h}_{k,i}^\lambda$, and $k = (a,b)$.

Sturm and Zhang [34] have shown that minimizing an indefinite quadratic function can be solved via an SDP formulation. Building on [34] and [96], Fang et al. [7] relaxed the unconstrained features

model with cross-entropy loss, originally formulated as a quadratically constrained quadratic program, into an SDP, and proved that the solution of the relaxed SDP coincides with that of the original non-convex program when the inequality constraint becomes an equality. Following the same reasoning, our paper proves that the minimizer of the convex program obtained via SDP relaxation (Equation (4)) recovers the solution of UFM$_\lambda$ (Equation (2)) when the equality constraint holds.

A potential concern may be the applicability of SDP beyond the UFM setting, particularly for deep networks with multiple layers. However, Fazlyab et al. [35] demonstrate that the safety of an abstracted network, which is sufficient for ensuring the safety of the original deep network, can be analyzed using semidefinite programming. As further discussed in [36], SDP relaxations of quadratically constrained quadratic programs generally provide approximate solutions with bounded gaps for deep networks.

E SDP Constraint Relaxation

In Equation (7) (and its counterpart in Theorem 5.1), SDP requires a strong constraint $p \geq K^2 + K$ ($p \geq 2K^2$) to guarantee the existence of $\left[\boldsymbol{h}^{\lambda\star}_{(1,1)}, \boldsymbol{h}^{\lambda\star}_{(1,2)}, \ldots, \boldsymbol{h}^{\lambda\star}_{(K,K)}, (\boldsymbol{W}^\star)^\top\right]$ $\left(\left[\boldsymbol{h}^{\lambda\star}_{(1,1)}, \boldsymbol{h}^{\lambda\star}_{(1,2)}, \ldots, \boldsymbol{h}^{\lambda\star}_{(K,K)}, (\boldsymbol{W}^{\lambda\star})^\top\right]\right)$. However, $p \geq 2K$ ($p \geq 2K$) alone is sufficient to guarantee it, since mixed features $\boldsymbol{h}^\lambda$ (mixed class vectors $\boldsymbol{W}^\lambda$) are linearly dependent on singleton features $\boldsymbol{h}$ (singleton class vectors $\boldsymbol{W}$). We have demonstrated that this SDP constraint relaxation can be justified by showing that $\bar{\boldsymbol{h}}_i - \boldsymbol{h}_{k,i} = -C_e \boldsymbol{w}_k$ is derived from $\bar{\boldsymbol{h}}^\lambda_i - \boldsymbol{h}^\lambda_{k,i} = -C_e \boldsymbol{w}^\lambda_k$ (see Equation (48) in Theorem G.5).

Practical realism should also be taken into account with respect to BMLS and MS while the assumption is just necessary to guarantee the theoretical robustness. While relaxing the strong assumption into $p \geq 2K$, it might be still strong. Furthermore, since the claim in this paper that balancing mixed labels mitigates minority collapse stems from Proposition 4.3 and Theorem 5.1, accounting for practical realism is also warranted.

BMLS and MS empirically demonstrate their effectiveness. Through the empirical validation of BMLS and MS on CIFAR100-LT, where our theoretical assumption is not satisfied ($p = 128 < 2K = 200$), we empirically substantiate the practical efficacy of our theoretical findings in resolving minority collapse in imbalanced learning.

Lemma 4.1 requires $p \geq 2K$, while Theorem 4.2 only requires $p \geq K$. The conditions $p \geq 2K$ in Lemma 4.1 is imposed to ensure the existence of the left-hand side terms in Equation (7) within the SDP. However, since the number of learnable class vectors remains K, the condition $p \geq K$ holds in Theorem 4.2, which characterizes the collapse among minority classes. This is consistent with the same pattern observed in [7] (Lemma 1: $p \geq 2K$ vs. Theorem 5: $p \geq K$).

F Theoretical Model and Discussion about Mixed-Singleton classifier

F.1 UFM with Mixed-Singleton classifier

By replacing the classifier as Mixed-Singleton classifier defined in Section 5, we obtain the UFM$_\lambda$ with Mixed-Singleton classifier (UFM$_\lambda$-MS):

$$\begin{aligned} \min_{\boldsymbol{W}^\lambda, \boldsymbol{H}^\lambda} \; & \mathbb{E}_\lambda \frac{1}{N} \sum_{k \in \mathbb{K}^2} \sum_{i=1}^{n_k} \mathcal{L}(\boldsymbol{W}^\lambda \boldsymbol{h}^\lambda_{k,i}, \boldsymbol{y}^\lambda_k) \\ \text{s.t.} \; & \frac{1}{|\mathbb{K}^2|} \sum_{k \in \mathbb{K}^2} \left\|\boldsymbol{w}^\lambda_k\right\|^2 \leq E_W, \\ & \frac{1}{|\mathbb{K}^2|} \sum_{k \in \mathbb{K}^2} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\|\boldsymbol{h}^\lambda_{k,i}\right\|^2 \leq E_H, \end{aligned} \tag{13}$$

where $\boldsymbol{W}^\lambda = [\lambda \boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_b]_{(a,b) \in \mathbb{K}^2}$.

F.2 Discussion about Mixup lambda-independence

Building on Theorem 4.2 and Proposition 4.3, we raise a conjecture: *If mixed labels are interpreted as singletons, then the mitigation of minority collapse will be enhanced.*

During the proof of Theorem 4.2, we focused on the observation that *the minority collapse of Mixup occurs regardless of the mixup lambda* λ in Equation (19) ($\mathcal{L}_B^{\ell}\left(\boldsymbol{W}^{\ell,A}, \boldsymbol{H}^{\ell,A}\right) = \lambda \cdot \log(K) + (1-\lambda) \cdot \log(K) = \log(K)$). This λ-independence stems from the characteristic that last-layer features are linearly interpolated when Mixup is applied to UFM, where the last-layer features are free optimization variables, and thus, this may be regarded as an artifact of UFM. However, according to [23], although this does not perfectly conform to the theoretical framework in real deep networks (such as WideResNet, ViT) where the role of λ may become more complex, the last-layer features of mixed samples exhibit a tendency to be positioned between the two classes used in Mixup, following a linear interpolation scheme. Therefore, it is reasonable to posit that λ-independence may also manifest to a certain degree in real deep networks beyond UFM. This suggests that the proposed Mixed-Singleton classifier can serve as a method applicable in general settings. At the same time, rather than computing the loss of mixed samples as a simple sum of scalar predictions for the classes engaged in the mixed label, the Mixed-Singleton classifier, which treats the mixed label as a singleton class and derives the loss therefrom, has the effect of potentially reducing the mixup loss's negligible influence on classifier learning, which has been reported that Mixup primarily facilitates representation learning while exerting a minimal or adverse effect on classifier learning [11], by maintaining mixed-label samples but removing the mixup loss.

G Proofs

G.1 Proof Sketch

We first present a proof sketch that outlines the approach we followed to propose and prove our theorems. [7] proved that oversampling mitigates minority collapse when singleton label samples are imbalanced, following the sequence outlined below. (Gray indicates the part as defined in [7].)

(1) Define the unconstrained features model. (Equation (7))
(2) Prove that NC properties are satisfied when the UFM has global optimality in the case where singleton label samples are balanced. (Theorem 1)
(3) Demonstrate that the UFM suffers from minority collapse in the case where singleton label samples are imbalanced. (Lemma 1 and Theorem 5)
(4) Show that oversampling alleviates minority collapse in the imbalanced case. (Proposition 1)

Our theorem and proof leverages strategies similar to those in [7], but we extend these concepts to Mixup focusing on the balance of mixed label samples.

In Section 4, (1) we define the unconstrained features model with Mixup (UFM$_\lambda$) and omit step (2), which holds true according to the theorem of [23]; (3) we prove that in the imbalanced case the UFM$_\lambda$ also suffers from minority collapse; and in closing, (4) we show that the Balanced Mixed Label Sampler (BMLS) alleviates the minority collapse. In Section 5.2, we extend the UFM$_\lambda$ by modifying the classifier: (1) we newly define the unconstrained features model with Mixup and Mixed-Singleton classifier (UFM$_\lambda$-MS); (2) we prove that when this model achieves global optimality, it also satisfies the NC properties; and finally, following the same reasoning as in Theorem 4.2, (3–4) we show that in the imbalanced case the UFM$_\lambda$-MS suffers from minority collapse, and that BMLS is effective to the minority collapse even in this setting in Theorem 5.1.

G.2 Proof of Lemma 4.1

Restated Lemma 4.1. Assume $p \geq 2K$ and the loss function $\mathcal{L}$ is convex in its first argument. Let $\boldsymbol{X}^\star$ be a minimizer of the convex program (Equation (4)). Define $\left(\boldsymbol{W}^\star, \boldsymbol{H}^{\lambda\star}\right)$ as

$$\left[\boldsymbol{h}^{\lambda\star}_{(1,1)}, \boldsymbol{h}^{\lambda\star}_{(1,2)}, \ldots, \boldsymbol{h}^{\lambda\star}_{(K,K)}, \ (\boldsymbol{W}^\star)^\top\right] = \boldsymbol{P}(\boldsymbol{X}^\star)^{1/2},$$

$$\boldsymbol{h}^\star_{k,i} = \boldsymbol{h}^\star_k, \quad \text{for all } i \in \mathcal{I}^\lambda_k, k \in \mathbb{K}^2,$$

where $(\boldsymbol{X}^\star)^{1/2}$ denotes the positive square root of $\boldsymbol{X}^\star$ and $\boldsymbol{P} \in \mathbb{R}^{p\times(K^2+K)}$ is any partial orthogonal matrix such that $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_{K^2+K}$. Then, $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ is a minimizer of Equation (2). Moreover, if all $\boldsymbol{X}^\star$'s satisfy $\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}^\star(k,k) = E_H$, then all the solutions of Equation (2) are in the form of Equation (5).

Proof. For any feasible solution $\left(\boldsymbol{W}, \boldsymbol{H}^\lambda\right)$ for the original program Equation (2), we define

$$\boldsymbol{h}_k^\lambda := \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{h}_{k,i}^\lambda,\ k \in \mathbb{K}^2,$$

and

$$\boldsymbol{X} := \left[\boldsymbol{h}_{(1,1)}^\lambda, \boldsymbol{h}_{(1,2)}^\lambda, \ldots, \boldsymbol{h}_{(K,K)}^\lambda, \boldsymbol{W}^\top\right]^\top \left[\boldsymbol{h}_{(1,1)}^\lambda, \boldsymbol{h}_{(1,2)}^\lambda, \ldots, \boldsymbol{h}_{(K,K)}^\lambda, \boldsymbol{W}^\top\right].$$

Clearly, $\boldsymbol{X} \succeq 0$. For the other two constraints of Equation (4), we have

$$\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}(k,k) = \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\|\boldsymbol{h}_k^\lambda\|^2 \overset{a}{\leq} \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}_{k,i}^\lambda\right\|^2 \overset{b}{\leq} E_H$$

and

$$\frac{1}{K}\sum_{k=K^2+1}^{K^2+K}\boldsymbol{X}(k,k) = \frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2 \overset{c}{\leq} E_W,$$

where $\overset{a}{\leq}$ applies Jensen's inequality and $\overset{b}{\leq}$ and $\overset{c}{\leq}$ use that $\left(\boldsymbol{W}, \boldsymbol{H}^\lambda\right)$ is a feasible solution. So $\boldsymbol{X}$ is a feasible solution for the convex program Equation (4). Letting L_0 be the global minimum of Equation (4), for any feasible solution $\left(\boldsymbol{W}, \boldsymbol{H}^\lambda\right)$, we obtain

$$\begin{aligned}\frac{1}{N}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n_k}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}^\lambda, \boldsymbol{y}_k^\lambda) &= \sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\left[\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}^\lambda, \boldsymbol{y}_k^\lambda)\right] \\ &\overset{a}{\geq} \sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_k^\lambda, \boldsymbol{y}_k^\lambda) = \sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\mathcal{L}(\boldsymbol{z}(k)^\lambda, \boldsymbol{y}_k^\lambda) \geq L_0, \end{aligned} \tag{14}$$

where in $\overset{a}{\geq}$, we use $\mathcal{L}$ is convex on the first argument, and so $\mathcal{L}(\boldsymbol{W}\boldsymbol{h}^\lambda, \boldsymbol{y}_k^\lambda)$ is convex on $\boldsymbol{h}$ given $\boldsymbol{W}$ and $k \in \mathbb{K}^2$.

For the simplicity of our expressions, we hereafter remove the superscript λ of $\boldsymbol{H}^\lambda$, $\boldsymbol{h}^\lambda$ and $\boldsymbol{z}^\lambda$.

On the other hand, considering the solution $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ defined in Equation (5) with $\boldsymbol{X}^\star$ being a minimizer of Equation (4), we have $\left[\boldsymbol{h}_{(1,1)}^\star, \boldsymbol{h}_{(1,2)}^\star, \ldots, \boldsymbol{h}_{(K,K)}^\star, \boldsymbol{W}^\top\right]^\top \left[\boldsymbol{h}_{(1,1)}^\star, \boldsymbol{h}_{(1,2)}^\star, \ldots, \boldsymbol{h}_{(K,K)}^\star, \boldsymbol{W}^\top\right] = \boldsymbol{X}^\star$ ($p \geq 2K$ guarantees the existence of $\left[\boldsymbol{h}_{(1,1)}^\star, \boldsymbol{h}_{(1,2)}^\star, \ldots, \boldsymbol{h}_{(K,K)}^\star, (\boldsymbol{W}^\star)^\top\right]$). We can verify that $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ is a feasible solution for Equation (2) and have

$$\frac{1}{N}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n_k}\mathcal{L}(\boldsymbol{W}^\star\boldsymbol{h}_{k,i}^\star, \boldsymbol{y}_k^\lambda) = \sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\mathcal{L}(\boldsymbol{z}(k)^\star, \boldsymbol{y}_k^\lambda) = L_0, \tag{15}$$

where $\boldsymbol{z}(k)^\star = \left[\boldsymbol{X}^\star(k, K^2+1), \boldsymbol{X}^\star(k, K^2+2), \ldots, \boldsymbol{X}^\star(k, K^2+K)\right]^\top$ for $k \in \mathbb{K}^2$.

Combining Equation (14) and Equation (15), we conclude that L_0 is the global minimum of Equation (2) and $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ is a minimizer.

Suppose there is a minimizer $(\boldsymbol{W}', \boldsymbol{H}')$ that cannot be written as Equation (5). Let

$$\boldsymbol{h}_k' = \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{h}_{k,i}',\ k \in \mathbb{K}^2,$$

and

$$\boldsymbol{X}' = \left[\boldsymbol{h}_{(1,1)}', \boldsymbol{h}_{(1,2)}', \ldots, \boldsymbol{h}_{(K,K)}', (\boldsymbol{W}')^\top\right]^\top \left[\boldsymbol{h}_{(1,1)}', \boldsymbol{h}_{(1,2)}', \ldots, \boldsymbol{h}_{(K,K)}', (\boldsymbol{W}')^\top\right].$$

Equation (14) implies that $\boldsymbol{X}'$ is a minimizer of Equation (4). As $(\boldsymbol{W}', \boldsymbol{H}')$ cannot be written as Equation (5) with $\boldsymbol{X}^\star = \boldsymbol{X}'$, then there is a $k' \in \mathbb{K}^2$, $i, j \in [n_{k'}]$ with $i \neq j$ such that $\boldsymbol{h}_{k',i} \neq \boldsymbol{h}_{k',j}$. We have

$$
\begin{aligned}
&\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}'(k,k) = \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\|\boldsymbol{h}'_k\|^2 \\
=&\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}'_{k,i}\right\|^2 - \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\|\boldsymbol{h}'_{k,i} - \boldsymbol{h}'_k\|^2 \\
\leq&\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}'_{k,i}\right\|^2 - \frac{1}{K^2}\frac{1}{n_{k'}}(\|\boldsymbol{h}'_{k',i} - \boldsymbol{h}'_{k'}\|^2 + \|\boldsymbol{h}'_{k',j} - \boldsymbol{h}'_{k'}\|^2) \\
\leq&\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}'_{k,i}\right\|^2 - \frac{1}{K^2}\frac{1}{2n_{k'}}\|\boldsymbol{h}'_{k',i} - \boldsymbol{h}'_{k',j}\|^2 \\
<&E_H.
\end{aligned}
$$

By contraposition, if all $\boldsymbol{X}^\star$ satisfy that $\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}^\star(k,k) = E_H$, then all the solutions of Equation (2) are in the form of Equation (5). We complete the proof. □

G.3 Proof of Theorem 4.2

Restated Theorem 4.2. Assume $p \geq K$ and $n_A/n_B \to \infty$, and fix K_A and K_B. Let $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ be any global minimizer of the UFM$_\lambda$ (Equation (2)). As the imbalance factor $R \equiv n_A/n_B \to \infty$, we have

$$\lim \boldsymbol{w}_k^\star - \boldsymbol{w}_{k'}^\star = \boldsymbol{0}_p, \text{ for all } K_A < k < k' \leq K.$$

To prove Theorem 4.2, we first study a limit case where we only learn the classification for partial classes. We solve the optimization program:

$$
\begin{aligned}
\min_{\boldsymbol{W},\boldsymbol{H}^\lambda}\ \mathbb{E}_{\lambda\sim D_\lambda}\quad & \frac{1}{|\mathbb{K}_A^2|\cdot n_A}\sum_{k\in\mathbb{K}_A^2}\sum_{i=1}^{n_A}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}^\lambda, \boldsymbol{y}_k^\lambda) \\
\text{s.t.}\quad & \frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2 \leq E_W, \\
& \frac{1}{|\mathbb{K}_\cup^2|}\sum_{k\in\mathbb{K}_\cup^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}_{k,i}^\lambda\right\|^2 \leq E_H,
\end{aligned}
\tag{16}
$$

where $\boldsymbol{y}_{(a,b)}^\lambda = \lambda\boldsymbol{y}_a + (1-\lambda)\boldsymbol{y}_b$, $\mathbb{K}_A^2 = \{(a,b)|1 \leq a \leq K_A \ \wedge\ 1 \leq b \leq K_A\}$, $\mathbb{K}_B^2 = \{(a,b)|K_A+1 \leq a \leq K \ \wedge\ K_A+1 \leq b \leq K\}$, $\mathbb{K}_\cup^2 = \mathbb{K}_A^2 \cup \mathbb{K}_B^2$ and

$$n_k = \begin{cases} n_A & \text{if } k = (a,b) \in \mathbb{K}_A^2 \\ n_B & \text{if } k = (a,b) \in \mathbb{K}_B^2 \\ 0 & \text{otherwise} \end{cases}.$$

For the simplicity of our expressions, we remove the superscript λ of $\boldsymbol{H}^\lambda$ and $\boldsymbol{h}^\lambda$.

Lemma G.1 characterizes useful properties for the minimizer of Equation (16).

Lemma G.1. *Let $(\boldsymbol{W}, \boldsymbol{H})$ be a minimizer of Equation (16). We have $\boldsymbol{h}_{k,i}^\lambda = \boldsymbol{0}_p$ for all $k \in \mathbb{K}_B^2$ and $i \in [n_B]$. Let L_0 be the global minimum of Equation (16). We have*

$$L_0 = \frac{1}{|\mathbb{K}_A^2|\cdot n_A}\sum_{k\in\mathbb{K}_A^2}\sum_{i=1}^{n_A}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}, \boldsymbol{y}_k^\lambda).$$

Then L_0 only depends on K_A, n_A, E_H, and E_W. Moreover, for any feasible solution $(\boldsymbol{W}', (\boldsymbol{H})')$, if there exist $k, k' \in \mathbb{K}_B^2$ such that $\|\boldsymbol{w}_k - \boldsymbol{w}_{k'}\| = \epsilon > 0$, we have

$$\frac{1}{|\mathbb{K}_A^2|\cdot n_A}\sum_{k\in\mathbb{K}_A^2}\sum_{i=1}^{n_A}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}, \boldsymbol{y}_k^\lambda) \geq L_0 + \epsilon',$$

where $\epsilon' > 0$ depends on ϵ, $|\mathbb{K}_A^2|$, n_A, E_H, and E_W.

Now we are ready to prove Theorem 4.2. The proof is based on the contradiction.

Proof of Theorem 4.2. Consider sequences n_A^ℓ and n_B^ℓ with $R^\ell := n_A^\ell / n_B^\ell$ for $\ell = 1, 2, \ldots$. We have $R^\ell \to \infty$. For each optimization program indexed by $\ell \in \mathbb{N}_+$, we introduce $(\boldsymbol{W}^{\ell,\star}, \boldsymbol{H}^{\ell,\star})$ as a minimizer and separate the objective function into two parts. We consider

$$\mathcal{L}^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) = \frac{|\mathbb{K}_A^2| \cdot n_A^\ell}{|\mathbb{K}_A^2| \cdot n_A^\ell + |\mathbb{K}_B^2| \cdot n_B^\ell} \mathcal{L}_A^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) + \frac{|\mathbb{K}_B^2| \cdot n_B^\ell}{|\mathbb{K}_A^2| \cdot n_A^\ell + |\mathbb{K}_B^2| \cdot n_B^\ell} \mathcal{L}_B^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right),$$

with

$$\mathcal{L}_A^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) := \frac{1}{|\mathbb{K}_A^2| \cdot n_A^\ell} \sum_{k \in \mathbb{K}_A^2} \sum_{i=1}^{n_A^\ell} \mathcal{L}\left(\boldsymbol{W}^\ell \boldsymbol{h}_{k,i}^\ell, \boldsymbol{y}_k^\lambda\right)$$

and

$$\mathcal{L}_B^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) := \frac{1}{|\mathbb{K}_B^2| \cdot n_B^\ell} \sum_{k \in \mathbb{K}_B^2} \sum_{i=1}^{n_B^\ell} \mathcal{L}\left(\boldsymbol{W}^\ell \boldsymbol{h}_{k,i}^\ell, \boldsymbol{y}_k^\lambda\right).$$

We define $\left(\boldsymbol{W}^{\ell,A}, \boldsymbol{H}^{\ell,A}\right)$ as a minimizer of the optimization program:

$$\begin{aligned} \min_{\boldsymbol{W}^\ell, \boldsymbol{H}^\ell} \quad & \mathcal{L}_A^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) \\ \text{s.t.} \quad & \frac{1}{K} \sum_{k=1}^K \left\|\boldsymbol{w}_k^\ell\right\|^2 \le E_W, \\ & \frac{1}{|\mathbb{K}_A^2|} \sum_{k \in \mathbb{K}_A^2} \frac{1}{n_A^\ell} \sum_{i=1}^{n_A^\ell} \left\|\boldsymbol{h}_{k,i}^\ell\right\|^2 + \frac{1}{|\mathbb{K}_B^2|} \sum_{k \in \mathbb{K}_B^2} \frac{1}{n_B^\ell} \sum_{i=1}^{n_B^\ell} \left\|\boldsymbol{h}_{k,i}^\ell\right\|^2 \le E_H, \end{aligned} \tag{17}$$

and $\left(\boldsymbol{W}^{\ell,B}, \boldsymbol{H}^{\ell,B}\right)$ as a minimizer of the optimization program:

$$\begin{aligned} \min_{\boldsymbol{W}^\ell, \boldsymbol{H}^\ell} \quad & \mathcal{L}_B^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) \\ \text{s.t.} \quad & \frac{1}{K} \sum_{k=1}^K \left\|\boldsymbol{w}_k^\ell\right\|^2 \le E_W, \\ & \frac{1}{|\mathbb{K}_A^2|} \sum_{k \in \mathbb{K}_A^2} \frac{1}{n_A^\ell} \sum_{i=1}^{n_A^\ell} \left\|\boldsymbol{h}_{k,i}^\ell\right\|^2 + \frac{1}{|\mathbb{K}_B^2|} \sum_{k \in \mathbb{K}_B^2} \frac{1}{n_B^\ell} \sum_{i=1}^{n_B^\ell} \left\|\boldsymbol{h}_{k,i}^\ell\right\|^2 \le E_H. \end{aligned} \tag{18}$$

Note that Programs Equation (17) and Equation (18) and their minimizers have been studied in Lemma G.1. We define:

$$L_A := \mathcal{L}_A^\ell\left(\boldsymbol{W}^{\ell,A}, \boldsymbol{H}^{\ell,A}\right) \quad \text{and} \quad L_B := \mathcal{L}_B^\ell\left(\boldsymbol{W}^{\ell,B}, \boldsymbol{H}^{\ell,B}\right).$$

Then Lemma G.1 implies that L_A and L_B only depend on $|\mathbb{K}_A^2|$, K_B, E_H, and E_W, and are independent of ℓ. Moreover, since $\boldsymbol{h}_{k,i}^{\ell,A} = \mathbf{0}_p$ for all $k \in \mathbb{K}_B^2$ and $i \in [n_B]$, we have

$$\mathcal{L}_B^\ell\left(\boldsymbol{W}^{\ell,A}, \boldsymbol{H}^{\ell,A}\right) = \lambda \cdot \log(K) + (1-\lambda) \cdot \log(K) = \log(K). \tag{19}$$

Now we prove Theorem 4.2 by contradiction. Suppose there exists a pair (k, k') such that $\lim_{\ell \to \infty} \boldsymbol{w}_k^{\ell,\star} - \boldsymbol{w}_{k'}^{\ell,\star} \neq \mathbf{0}_p$. Then there exists $\epsilon > 0$ such that for a subsequence $\{(\boldsymbol{w}^{a_\ell,\star}, \boldsymbol{h}^{a_\ell,\star})\}_{\ell=1}^\infty$ and an index ℓ_0 when $\ell \ge \ell_0$, we have $\left\|\boldsymbol{W}_k^{a_\ell,\star} - \boldsymbol{W}_{k'}^{a_\ell,\star}\right\| \ge \epsilon$. Now we figure out a contradiction by estimating the objective function value on $(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star})$. In fact, because $(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star})$ is a minimizer of $\mathcal{L}^\ell(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell)$, we have

$$\begin{aligned} \mathcal{L}^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) \le \; & \mathcal{L}^{a_\ell}\left(\boldsymbol{W}^{a_\ell,A}, \boldsymbol{H}^{a_\ell,A}\right) \\ \overset{\text{Equation (19)}}{=} \; & \frac{|\mathbb{K}_A^2| \cdot n_A^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} L_A + \frac{|\mathbb{K}_B^2| \cdot n_B^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} \log(K) \\ = \; & L_A + \frac{1}{K_R R^{a_\ell} + 1} \left(\log(K) - L_A\right) \overset{\ell \to \infty}{\to} L_A, \end{aligned} \tag{20}$$

where we define $K_R := |\mathbb{K}_A^2|/|\mathbb{K}_B^2|$ and use $R^\ell = n_A^\ell / n_B^\ell$.

However, when $\ell > \ell_0$, because $\left\|\boldsymbol{w}_k^{a_\ell,\star} - \boldsymbol{w}_{k'}^{a_\ell,\star}\right\| \geq \epsilon > 0$, Lemma G.1 implies that

$$\mathcal{L}_A^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) \geq L_A + \epsilon_2,$$

where $\epsilon_2 > 0$ only depends on ϵ, $|\mathbb{K}_A^2|$, K_B, E_H, and E_W, and is independent of ℓ. We obtain

$$\begin{aligned}
\mathcal{L}^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) &= \frac{|\mathbb{K}_A^2| \cdot n_A^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} \mathcal{L}_A^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) \\
&\quad + \frac{|\mathbb{K}_B^2| \cdot n_B^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} \mathcal{L}_B^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) \\
&\overset{a}{\geq} \frac{|\mathbb{K}_A^2| \cdot n_A^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} \mathcal{L}_A^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) \\
&\quad + \frac{|\mathbb{K}_B^2| \cdot n_B^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} \mathcal{L}_B^{a_\ell}\left(\boldsymbol{W}^{a_\ell,B}, \boldsymbol{H}^{a_\ell,B}\right) \\
&= \frac{|\mathbb{K}_A^2| \cdot n_A^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} (L_A + \epsilon_2) + \frac{|\mathbb{K}_B^2| \cdot n_B^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} L_B \\
&= L_A + \epsilon_2 + \frac{1}{K_R R^{a_\ell} + 1}(L_B - L_A - \epsilon_2) \overset{\ell \to \infty}{\to} L_A + \epsilon_2, \qquad (21)
\end{aligned}$$

where $\overset{a}{\geq}$ uses $\left(\boldsymbol{W}^{a_\ell,B}, \boldsymbol{H}^{a_\ell,B}\right)$ is the minimizer of Equation (18). Thus we meet contradiction by comparing Equation (20) with Equation (21) and achieve Theorem 4.2. □

Proof of Lemma G.1. For any constants $C_a > 0$, $C_b > 0$, and $C_c > 0$, define

$$\begin{aligned}
C_a' &:= \frac{C_a}{C_a + (K_A - 1)C_b + K_B C_c} \in (0,1) \\
C_b' &:= \frac{C_b}{C_a + (K_A - 1)C_b + K_B C_c} \in (0,1) \\
C_c' &:= \frac{C_c}{C_a + (K_A - 1)C_b + K_B C_c} \in (0,1) \\
C_d &:= -C_a' \log(C_a') - C_b'(K_A - 1)\log(C_b') - K_B C_c' \log(C_c') \\
C_e &:= \frac{K_A C_b}{K_A C_b + K_B C_c} \in (0,1) \\
C_f &:= \frac{K_B C_c}{K_A C_b + K_B C_c} \in (0,1) \\
C_g &:= \frac{K_A C_b + K_B C_c}{C_a + (K_A - 1)C_b + K_B C_c} > 0.
\end{aligned}$$

Using a similar argument as Theorem G.5, we show in Lemma G.2 (see the end of the proof), for any feasible solution $(\boldsymbol{W}, \boldsymbol{H})$ of Equation (16), the objective value of Equation (16) can be bounded from below by:

$$\begin{aligned}
&\frac{1}{|\mathbb{K}_A^2| n_A} \sum_{k \in \mathbb{K}_A^2} \sum_{i=1}^{n_A} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{(a,b),i}, \boldsymbol{y}_{(a,b)}) \\
&\overset{a}{\geq} -\frac{C_g}{K_A}\sqrt{KE_H}\sqrt{\sum_{k=1}^{K_A} \|C_e \boldsymbol{w}_A + C_f \boldsymbol{w}_B - \boldsymbol{w}_k\|^2} + C_d \\
&\overset{b}{\geq} -\frac{C_g}{K_A}\sqrt{KE_H}\sqrt{KE_W - K_A\left(1/K_R - C_f^2 - \frac{C_f^4}{C_e(2 - C_e)}\right)\|\boldsymbol{w}_B\|^2 - \sum_{k=K_A+1}^{K} \|\boldsymbol{w}_k - \boldsymbol{w}_B\|^2} + C_d
\end{aligned} \qquad (22)$$

where $\boldsymbol{w}_A := \frac{1}{K_A}\sum_{k=1}^{K_A}\boldsymbol{w}_k$, $\boldsymbol{w}_B := \frac{1}{K_B}\sum_{k=K_A+1}^{K}\boldsymbol{w}_k$, and $K_R := \frac{K_A}{K_B}$. Moreover, the equality in $\overset{a}{\geq}$ holds only if $\boldsymbol{h}_{k,i} = \boldsymbol{0}_p$ for all $k \in [K_A+1:K]$ and $i \in [n_B]$.

Though C_a, C_b, and C_c can be any positive numbers, we need to carefully pick them to exactly reach the global minimum of Equation (16). In the following, we separately consider three cases according to the values of K_A, K_B, and $E_H E_W$.

(Case 1) Consider the case when $K_A = 1$. We pick $C_a := \exp\left(\sqrt{K_B(1+K_B)E_H E_W}\right)$, $C_b := 1$, and $C_c := \exp\left(-\sqrt{(1+K_B)E_H E_W/K_B}\right)$.

Then, from $\overset{a}{\geq}$ in Equation (22), we have

$$\begin{aligned}
&\frac{1}{|\mathbb{K}_A^2|n_A}\sum_{k\in\mathbb{K}_A^2}\sum_{i=1}^{n_A}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{(a,b),i},\boldsymbol{y}_{(a,b)}) \\
\overset{a}{\geq}& -C_g C_f\sqrt{KE_H}\sqrt{\|\boldsymbol{w}_1-\boldsymbol{w}_B\|^2}+C_d \\
=& -C_g C_f\sqrt{KE_H}\sqrt{\|\boldsymbol{w}_1\|^2-2\boldsymbol{w}_1^\intercal\boldsymbol{w}_B+\|\boldsymbol{w}_B\|^2}+C_d \\
\overset{b}{\geq}& -C_g C_f\sqrt{KE_H}\sqrt{(1+1/K_B)(\|\boldsymbol{w}_1\|^2+K_B\|\boldsymbol{w}_B\|^2)}+C_d \\
\overset{c}{\geq}& -C_g C_f\sqrt{KE_H}\sqrt{(1+1/K_B)\left(KE_W-\sum_{k=2}^{K}\|\boldsymbol{w}_k-\boldsymbol{w}_B\|^2\right)}+C_d \\
\overset{c}{\geq}& -C_g C_f\sqrt{KE_H}\sqrt{(1+1/K_B)KE_W}+C_d := L_1
\end{aligned} \tag{23}$$

where $\overset{a}{\geq}$ uses $C_e + C_f = 1$, $\overset{b}{\geq}$ follows from $-2ab \leq a^2+b^2$, *i.e.*, $-2\boldsymbol{w}_1^\intercal\boldsymbol{w}_B \leq (1/K_B)\|\boldsymbol{w}_1\|^2 + K_B\|\boldsymbol{w}_B\|^2$, and $\overset{c}{\geq}$ follows from $\sum_{k=2}^{K}\|\boldsymbol{w}_k\|^2 = K_B\|\boldsymbol{w}_B\|^2 + \sum_{k=2}^{K}\|\boldsymbol{w}_k-\boldsymbol{w}_B\|^2$ and the constraint that $\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2 \leq KE_W$.

On the other hand, when $(\boldsymbol{W},\boldsymbol{H})$ satisfies that

$$\begin{aligned}
\boldsymbol{w}_1 = \sqrt{K_B E_W}\boldsymbol{u}, \quad \boldsymbol{w}_k = -\sqrt{\frac{1}{K_B E_W}}\boldsymbol{u}, \ \ k\in[2:K], \\
\boldsymbol{h}_{1,i} = \sqrt{(1+K_B)E_H}\boldsymbol{u}, \ \ i\in[n_A], \quad \boldsymbol{h}_{k,i} = \boldsymbol{0}_p \ \ k\in[2:K], \ \ i\in[n_B],
\end{aligned}$$

where $\boldsymbol{u}$ is any unit vector, the inequalities in Equation (23) reduces to equalities. So, L_1 is the global minimum of Equation (16). Moreover, L_1 is achieved only if $\overset{a}{\geq}$ in Equation (22) reduces to equality. From Lemma G.2, we have that any minimizer satisfies that $\boldsymbol{h}_{k,i} = \boldsymbol{0}_p$ for all $k \in [K_A+1:K]$ and $i \in [n_B]$.

Finally, for any feasible solution $(\boldsymbol{W}',\boldsymbol{H}')$, if there exist $k,k' \in [K_A+1:K]$ such that $\|\boldsymbol{w}_k - \boldsymbol{w}_{k'}\| = \varepsilon > 0$, we have

$$\sum_{k=K_A+1}^{K}\|\boldsymbol{w}_k-\boldsymbol{w}_B\|^2 \geq \|\boldsymbol{w}_k-\boldsymbol{w}_B\|^2 + \|\boldsymbol{w}_{k'}-\boldsymbol{w}_B\|^2 \geq \frac{\|\boldsymbol{w}_k-\boldsymbol{w}_{k'}\|^2}{2} = \varepsilon^2/2. \tag{24}$$

It follows from $\overset{c}{\geq}$ in Equation (23) that

$$\begin{aligned}
&\frac{1}{|\mathbb{K}_A^2|n_A}\sum_{k\in\mathbb{K}_A^2}\sum_{i=1}^{n_A}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{(a,b),i},\boldsymbol{y}_{(a,b)}) \\
\geq& -C_g C_f\sqrt{KE_H}\sqrt{(1+1/K_B)(KE_W-\varepsilon^2/2)}+C_d := L_1+\varepsilon_1,
\end{aligned} \tag{25}$$

with $\varepsilon_1 > 0$ depending on ε, K_A, K_B, E_H, and E_W.

(Case 2) Consider the case when $K_A > 1$ and $\exp\left((1+1/K_R)\sqrt{E_H E_W}/(K_A-1)\right) < \sqrt{1+K_R}+1$. Let us pick $C_a := \exp\left((1+1/K_R)\sqrt{E_H E_W}\right)$, $C_b := \exp\left(-\frac{1}{K_A-1}(1+1/K_R)\sqrt{E_H E_W}\right)$, and $C_c := 1$.

Following from $\overset{b}{\geq}$ in Equation (22), we know if $1/K_R - C_f^2 - \frac{C_f^4}{C_e(2-C_e)} > 0$, then

$$\frac{1}{|\mathbb{K}_A^2| n_A} \sum_{k\in\mathbb{K}_A^2} \sum_{i=1}^{n_A} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{(a,b),i}, \boldsymbol{y}_{(a,b)}) \geq -C_g(1+1/K_R)\sqrt{E_H E_W} + C_d := L_2 \tag{26}$$

In fact, we do have $1/K_R - C_f^2 - \frac{C_f^4}{C_e(2-C_e)} > 0$ because

$$\begin{aligned}
& 1/K_R > C_f^2 + \frac{C_f^4}{C_e(2-C_e)} \\
\overset{a}{\Longleftrightarrow}\ & C_f < \sqrt{\frac{1}{1+K_R}} \\
\overset{b}{\Longleftrightarrow}\ & \frac{C_b}{C_c} > \frac{1}{\sqrt{1+K_R}+1} \\
\Longleftrightarrow\ & \exp\left((1+1/K_R)\sqrt{E_H E_W}/(K_A-1)\right) < \sqrt{1+K_R}+1.
\end{aligned}$$

where in $\overset{a}{\Longleftrightarrow}$, $C_e + C_f = 1$, and in $\overset{b}{\Longleftrightarrow}$, $C_f = \frac{K_B C_c}{K_A C_b + K_B C_c}$.

On the other hand, when $(\boldsymbol{W}, \boldsymbol{H})$ satisfies that

$$\begin{aligned}
[\boldsymbol{w}_1, ..., \boldsymbol{w}_{K_A}] = \sqrt{\frac{E_W}{E_H}}\,[\boldsymbol{h}_1, ...\boldsymbol{h}_{K_A}]^\intercal = \sqrt{(1+1/K_R)E_W}(\boldsymbol{M}_A^*)^\intercal, \\
\boldsymbol{h}_{k,i} = \boldsymbol{h}_k, \quad k \in [K_A],\ \ i \in [n_A], \\
\boldsymbol{h}_{k,i} = \boldsymbol{w}_k = \boldsymbol{0}_p, \quad k \in [K_A+1:K],\ \ i \in [n_B],
\end{aligned}$$

where $(\boldsymbol{M}_A^*)$ is a K_A-simplex ETF, Equation (26) reduces to equality. So, L_2 is the global minimum of Equation (16). Moreover, L_2 is achieved only if $\overset{a}{\geq}$ in Equation (22) reduces to equality. From Lemma G.2, we have that any minimizer satisfies that $\boldsymbol{h}_{k,i} = \boldsymbol{0}_p$ for all $k \in [K_A+1:K]$ and $i \in [n_B]$.

Finally, for any feasible solution $(\boldsymbol{W}', \boldsymbol{H}')$, if there exist $k, k' \in [K_A+1:K]$ such that $\|\boldsymbol{w}_k - \boldsymbol{w}_{k'}\| = \varepsilon > 0$, plugging Equation (24) into $\overset{b}{\geq}$ in Equation (22), we have

$$\begin{aligned}
& \frac{1}{|\mathbb{K}_A^2| n_A} \sum_{k\in\mathbb{K}_A^2} \sum_{i=1}^{n_A} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{(a,b),i}, \boldsymbol{y}_{(a,b)}) \\
\geq & -\frac{C_g}{K_A}\sqrt{KE_H}\sqrt{KE_W - \varepsilon^2/2)} + C_d := L_2 + \varepsilon_2,
\end{aligned} \tag{27}$$

with $\varepsilon_2 > 0$ depending on ε, K_A, K_B, E_H, and E_W.

(Case 3) Consider the case when $K_A > 1$ and $\exp\left((1+1/K_R)\sqrt{E_H E_W}/(K_A-1)\right) \geq \sqrt{1+K_R}+1$. Let $C'_f := \frac{1}{\sqrt{1+K_R}}$ and $C'_e = 1 - C'_f$. For $x \in [0,1]$, we define:

$$g_N(x) := \sqrt{\frac{(1+K_R)E_W}{K_R x^2 + K_R(1+K_R)(1-x)^2}},$$

$$g_a(x) := \exp\left(\frac{g_N(x)\sqrt{(1+K_R)E_H/K_R}}{x^2 + \left(1+\frac{C'_e}{C'_f}\right)^2(1-x)^2}\left[x^2 + \left(1+\frac{C'_e}{C'_f}\right)(1-x)^2\right]\right),$$

$$g_b(x) := \exp\left(\frac{g_N(x)\sqrt{(1+K_R)E_H/K_R}}{x^2 + \left(1+\frac{C'_e}{C'_f}\right)^2(1-x)^2}\left[-\frac{1}{K_A-1}x^2 + \left(1+\frac{C'_e}{C'_f}\right)(1-x)^2\right]\right),$$

$$g_c(x) := \exp\left(\frac{g_N(x)\sqrt{(1+K_R)E_H/K_R}}{x^2 + \left(1+\frac{C'_e}{C'_f}\right)^2(1-x)^2}\left[-\left(1+\frac{C'_e}{C'_f}\right)K_R(1-x)^2\right]\right),$$

Let $x_0 \in [0,1]$ be a root of the equation

$$g_b(x)/g_c(x) = \frac{1/C'_f - 1}{K_R}.$$

We first show that the solution x_0 exists. First of all, one can directly verify then $x \in [0,1]$, $g_b(x)/g_c(x)$ is continuous. It suffices to prove that (A) $g_b(0)/g_c(0) \geq \frac{1/C'_f-1}{K_R}$ and (B) $g_b(1)/g_c(1) \leq \frac{1/C'_f-1}{K_R}$.

(A) When $x=0$, we have $g_b(x)/g_c(x) \geq \exp(0) = 1$. At the same time, $\frac{1/C'_f-1}{K_R} = \frac{\sqrt{1+K_R}-1}{K_R} = \frac{1}{\sqrt{1+K_R}+1} \leq 1$. Thus, $g_b(0)/g_c(0) \geq \exp(0) = 1 \geq \frac{1/C'_f-1}{K_R}$ is achieved.

(B) When $x=1$, we have $g_N(1) = \sqrt{(1+1/K_R)E_W}$. So,

$$g_b(1)/g_c(1) = \exp\left(-(1+1/K_R)\sqrt{E_H E_W}/(K_A-1)\right) \overset{a}{\leq} \frac{1}{\sqrt{1+K_R}+1} = \frac{1/C'_f-1}{K_R},$$

where $\overset{a}{\leq}$ is obtained by the condition that

$$\exp\left((1+1/K_R)\sqrt{E_H E_W}/(K_A-1)\right) \geq \sqrt{1+K_R}+1.$$

Now, we pick $C_a := g_a(x_0)$, $C_b := g_b(x_0)$, and $C_c := g_c(x_0)$, because $\frac{C_b}{C_c} = \frac{1/C'_f-1}{K_R}$, we have $C_e = C'_e$, $C_f = C'_f$, and $1/K_R = C_f^2 + \frac{C_f^4}{C_e(2-C_e)}$. Then, it follows from $\overset{b}{\geq}$ in Equation (22) that

$$\frac{1}{|\mathbb{K}_A^2| n_A}\sum_{k\in\mathbb{K}_A^2}\sum_{i=1}^{n_A}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{(a,b),i}, \boldsymbol{y}_{(a,b)}) \geq -C_g(1+1/K_R)\sqrt{E_H E_W} + C_d := L_2. \tag{28}$$

On the other hand, consider the solution $(\boldsymbol{W}, \boldsymbol{H})$ that satisfies

$$\boldsymbol{w}_k = g_N(x_0)\boldsymbol{P}_A\left[\frac{x_0}{\sqrt{(K_A-1)K_A}}(K_A\boldsymbol{y}_k - \boldsymbol{1}_{K_A} + \frac{1-x_0}{\sqrt{K_A}}\boldsymbol{1}_{K_A}\right], \ k \in [K_A],$$

$$\boldsymbol{w}_k = -\frac{C_e(2-C_e)}{C_f^2 K_A}\boldsymbol{P}_A\sum_{k=1}^{K_A}\boldsymbol{w}_k, \ k \in [K_A+1:K],$$

$$\boldsymbol{h}_{k,i} = \frac{\sqrt{(1+1/K_R)E_H}}{\|\boldsymbol{w}_i + \frac{C_e}{C_f K_A}\sum_{k=1}^{K_A}\boldsymbol{w}_k\|}\boldsymbol{P}_A\left[\boldsymbol{w}_i\frac{C_e}{C_f K_A}\sum_{k=1}^{K_A}\boldsymbol{w}_k\right], \ k \in [K_A], \ i \in [n_A]$$

$$\boldsymbol{h}_{k,i} = \boldsymbol{0}_p, \ k \in [K_A+1:K], \ i \in [n_B],$$

where $\boldsymbol{y}_k \in \mathbb{R}^K$ is the one-hot vector of the k-th class label and $\boldsymbol{P}_A \in \mathbb{R}^{p \times K_A}$ is a partial orthogonal matrix such that $\boldsymbol{P}_A^\intercal \boldsymbol{P}_A = \boldsymbol{I}_{K_A}$. We have $\exp\left(\boldsymbol{h}_{k,i}^\intercal \boldsymbol{w}_k\right) = g_a(x_0)$ for $i \in [n_A]$ and $k \in [K_A]$, $\exp\left(\boldsymbol{h}_{k,i}^\intercal \boldsymbol{w}_{k'}\right) = g_b(x_0)$ for $i \in [n_A]$ and $k, k' \in [K_A]$ such that $k \neq k'$, and $\exp\left(\boldsymbol{h}_{k,i}^\intercal \boldsymbol{w}_{k'}\right) = g_c(x_0)$ for $i \in [n_A]$, $k \in [K_A]$, and $k' \in [K_A + 1 : K]$. Moreover, $(\boldsymbol{W}, \boldsymbol{H})$ can achieve the equality in Equation (28). Finally, following the same argument as **(Case 2)**, we have that (1) L_2 is the global minimum of Equation (16); (2) any minimizer satisfies that $\boldsymbol{h}_{k,i} = \boldsymbol{0}_p$ for all $k \in [K_A + 1 : K]$ and $i \in [n_B]$; (3) for any feasible solution $(\boldsymbol{W}', \boldsymbol{H}')$, if there exist $k, k' \in [K_A + 1 : K]$ such that $\|\boldsymbol{w}_k - \boldsymbol{w}_{k'}\| = \varepsilon > 0$, then Equation (27) holds.

Combining the three cases, we obtain Lemma G.1, completing the proof. □

Lemma G.2. *For any constants $C_a > 0$, $C_b > 0$, and $C_c > 0$, define*

$$
\begin{aligned}
C_a' :=& \frac{C_a}{C_a + (K_A - 1)C_b + K_B C_c} \in (0, 1) \\
C_b' :=& \frac{C_b}{C_a + (K_A - 1)C_b + K_B C_c} \in (0, 1) \\
C_c' :=& \frac{C_c}{C_a + (K_A - 1)C_b + K_B C_c} \in (0, 1) \\
C_d :=& -C_a' \log(C_a') - C_b'(K_A - 1)\log(C_b') - K_B C_c' \log(C_c') \\
C_e :=& \frac{K_A C_b}{K_A C_b + K_B C_c} \in (0, 1) \\
C_f :=& \frac{K_B C_c}{K_A C_b + K_B C_c} \in (0, 1) \\
C_g :=& \frac{K_A C_b + K_B C_c}{C_a + (K_A - 1)C_b + K_B C_c} > 0.
\end{aligned}
$$

For any feasible solution $(\boldsymbol{W}, \boldsymbol{H})$ of Equation (16), the objective value of Equation (16) can be bounded from below by:

$$
\begin{aligned}
&\frac{1}{|\mathbb{K}_A^2| n_A} \sum_{k \in \mathbb{K}_A^2} \sum_{i=1}^{n_A} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{(a,b),i}, \boldsymbol{y}_{(a,b)}) \\
&\overset{a}{\geq} -\frac{C_g}{K_A}\sqrt{KE_H}\sqrt{\sum_{k=1}^{K_A} \|C_e \boldsymbol{w}_A + C_f \boldsymbol{w}_B - \boldsymbol{w}_k\|^2 + C_d} \\
&\overset{b}{\geq} -\frac{C_g}{K_A}\sqrt{KE_H}\sqrt{KE_W - K_A\left(1/K_R - C_f^2 - \frac{C_f^4}{C_e(2 - C_e)}\right)\|\boldsymbol{w}_B\|^2 - \sum_{k=K_A+1}^{K} \|\boldsymbol{w}_k - \boldsymbol{w}_B\|^2 + C_d}
\end{aligned}
\tag{29}
$$

where $\boldsymbol{w}_A := \frac{1}{K_A}\sum_{k=1}^{K_A} \boldsymbol{w}_k$, $\boldsymbol{w}_B := \frac{1}{K_B}\sum_{k=K_A+1}^{K} \boldsymbol{w}_k$, and $K_R := \frac{K_A}{K_B}$. Moreover, the equality in $\overset{a}{\geq}$ holds only if $\boldsymbol{h}_{k,i} = \boldsymbol{0}_p$ for all $k \in [K_A + 1 : K]$ and $i \in [n_B]$.

Remark G.3. Note that the case $\boldsymbol{h}_{k,i} = \boldsymbol{0}_p$ does not imply that network activations all die for the classes $k \in [K_A + 1 : K]$. This is because our analysis does not include the bias term for simplicity.

Proof of Lemma G.2. For $(a, b) \in \mathbb{K}_A^2$ and $i \in [n_{(a,b)}]$, we introduce $\boldsymbol{z}_{(a,b),i} = \boldsymbol{W}\boldsymbol{h}_{(a,b),i}^{\lambda}$. Because that $C_a' + (K_A - 1)C_b' + K_B C_c' = 1$, $C_a' > 0$, $C_b' > 0$, and $C_c' > 0$, by the concavity of $\log(\cdot)$, we

have

$$
\begin{aligned}
&-\lambda \log\left(\frac{\exp(\boldsymbol{z}_{(a,b),i}(a))}{\sum_{k'=1}^{K}\exp(\boldsymbol{z}_{(a,b),i}(k'))}\right)-(1-\lambda)\log\left(\frac{\exp(\boldsymbol{z}_{(a,b),i}(b))}{\sum_{k'=1}^{K}\exp(\boldsymbol{z}_{(a,b),i}(k'))}\right)\\
=&-\lambda\boldsymbol{z}_{(a,b),i}(a)-(1-\lambda)\boldsymbol{z}_{(a,b),i}(b)\\
&+\lambda\log\left(C'_a\left(\frac{\exp(\boldsymbol{z}_{(a,b),i}(a))}{C'_a}\right)+\sum_{k'=1,k'\neq a}^{K_A}C'_b\left(\frac{\exp(\boldsymbol{z}_{k,i}(k'))}{C'_b}\right)+\sum_{k'=K_A+1}^{K}C'_c\left(\frac{\exp(\boldsymbol{z}_{k,i}(k'))}{C'_c}\right)\right)\\
&+(1-\lambda)\log\left(C'_a\left(\frac{\exp(\boldsymbol{z}_{(a,b),i}(b))}{C'_a}\right)+\sum_{k'=1,k'\neq b}^{K_A}C'_b\left(\frac{\exp(\boldsymbol{z}_{k,i}(k'))}{C'_b}\right)+\sum_{k'=K_A+1}^{K}C'_c\left(\frac{\exp(\boldsymbol{z}_{k,i}(k'))}{C'_c}\right)\right)\\
\geq&-\lambda\boldsymbol{z}_{(a,b),i}(a)-(1-\lambda)\boldsymbol{z}_{(a,b),i}(b)+C'_a\left(\lambda\boldsymbol{z}_{(a,b),i}(a)+(1-\lambda)\boldsymbol{z}_{(a,b),i}(b)\right)\\
&+C'_b\left(\lambda\sum_{k'=1,k'\neq a}^{K_A}\boldsymbol{z}_{(a,b),i}(k')+(1-\lambda)\sum_{k'=1,k'\neq b}^{K_A}\boldsymbol{z}_{(a,b),i}(k')\right)+C'_c\sum_{k'=K_A+1}^{K}\boldsymbol{z}_{(a,b),i}(k')+C_d\\
=&C_gC_e\left(\frac{1}{K_A}\sum_{k'=1}^{K_A}\boldsymbol{z}_{(a,b),i}(k')-\lambda\boldsymbol{z}_{(a,b),i}(a)-(1-\lambda)\boldsymbol{z}_{(a,b),i}(b)\right)\\
&+C_gC_f\left(\frac{1}{K_B}\sum_{k'=K_A+1}^{K}\boldsymbol{z}_{(a,b),i}(k')-\lambda\boldsymbol{z}_{(a,b),i}(a)-(1-\lambda)\boldsymbol{z}_{(a,b),i}(b)\right)+C_d.
\end{aligned}
\tag{30}
$$

Therefore, integrating Equation (30) with $(a,b)\in\mathbb{K}_A^2$ and $i\in[n_A]$, recalling that $\boldsymbol{w}_A=\frac{1}{K_A}\sum_{k=1}^{K_A}\boldsymbol{w}_k$ and $\boldsymbol{w}_B=\frac{1}{K_B}\sum_{k=K_A+1}^{K}\boldsymbol{w}_k$, we have

$$
\begin{aligned}
&\frac{1}{|\mathbb{K}_A^2|n_A}\sum_{(a,b)\in\mathbb{K}_A^2}\sum_{i=1}^{n_A}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}^{\lambda}_{(a,b),i},\boldsymbol{y}_{(a,b)})\\
\geq&\frac{1}{|\mathbb{K}_A^2|n_A}\sum_{(a,b)\in\mathbb{K}_A^2}\sum_{i=1}^{n_A}C_g\begin{pmatrix}C_e\left(\lambda(\boldsymbol{h}_{a,i}\boldsymbol{w}_A-\boldsymbol{h}_{a,i}\boldsymbol{w}_a)+(1-\lambda)(\boldsymbol{h}_{b,i}\boldsymbol{w}_A-\boldsymbol{h}_{b,i}\boldsymbol{w}_b)\right)\\+C_f\left(\lambda(\boldsymbol{h}_{a,i}\boldsymbol{w}_B-\boldsymbol{h}_{a,i}\boldsymbol{w}_a)+(1-\lambda)(\boldsymbol{h}_{b,i}\boldsymbol{w}_B-\boldsymbol{h}_{b,i}\boldsymbol{w}_b)\right)\end{pmatrix}\\
\overset{a}{=}&\frac{1}{K_An_A}\sum_{k=1}^{K_A}\sum_{i=1}^{n_A}Cg\left[C_e(\boldsymbol{h}_{k,i}\boldsymbol{w}_A-\boldsymbol{h}_{k,i}\boldsymbol{w}_k)+C_f(\boldsymbol{h}_{k,i}\boldsymbol{w}_B-\boldsymbol{h}_{k,i}\boldsymbol{w}_k)\right]+C_d\\
\overset{b}{=}&\frac{C_g}{K_A}\sum_{k=1}^{K_A}\boldsymbol{h}_{k,i}^{\intercal}(C_e\boldsymbol{w}_A+C_f\boldsymbol{w}_B-\boldsymbol{w}_k)+C_d,
\end{aligned}
\tag{31}
$$

where in $\overset{a}{=}$, we use $\sum_{(a,b)\in\mathbb{K}_A^2}\boldsymbol{h}^{\lambda}_{(a,b),i}=K_A\sum_{k=1}^{K_A}\boldsymbol{h}_{k,i}$, and in $\overset{b}{=}$, we introduce $\boldsymbol{h}_k:=\frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{h}_{k,i}$ for $k\in[K]$ and use $C_e+C_f=1$. Then, it is sufficient to bound $\sum_{k=1}^{K_A}\boldsymbol{h}_k^{\intercal}(C_e\boldsymbol{w}_A+C_f\boldsymbol{w}_B-\boldsymbol{w}_k)$. By the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\sum_{k=1}^{K_A}\boldsymbol{h}_k^{\intercal}(C_e\boldsymbol{w}_A+C_f\boldsymbol{w}_B-\boldsymbol{w}_k)&\geq-\sqrt{\sum_{k=1}^{K_A}\|\boldsymbol{h}_k\|^2}\sqrt{\sum_{k=1}^{K_A}\|C_e\boldsymbol{w}_A+C_f\boldsymbol{w}_B-\boldsymbol{w}_k\|^2}\\
&\overset{a}{\geq}-\sqrt{\sum_{k=1}^{K_A}\frac{1}{n_k}\sum_{i=1}^{n_k}\|\boldsymbol{h}_{k,i}\|^2}\sqrt{\sum_{k=1}^{K_A}\|C_e\boldsymbol{w}_A+C_f\boldsymbol{w}_B-\boldsymbol{w}_k\|^2}\\
&\overset{b}{\geq}-\sqrt{KE_H}\sqrt{\sum_{k=1}^{K_A}\|C_e\boldsymbol{w}_A+C_f\boldsymbol{w}_B-\boldsymbol{w}_k\|^2},
\end{aligned}
\tag{32}
$$

where $\overset{a}{\geq}$ follows from Jensen's inequality $\|\boldsymbol{h}_k\|^2 \leq \frac{1}{n_k}\sum_{i=1}^{n_k}\|\boldsymbol{h}_{k,i}\|^2$ for $k \in [K_A]$, and $\overset{b}{\geq}$ uses the constraint that $\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\|\boldsymbol{h}_{k,i}\|^2 \leq E_H$. Moreover, we have $\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\|\boldsymbol{h}_{k,i}\|^2 = E_H$ only if $\boldsymbol{h}_{k,i} = \boldsymbol{0}_p$ for all $k \in [K_A+1, K]$. Plugging Equation (32) to Equation (31), we obtain $\overset{a}{\geq}$ in Equation (29).

We then bound $\sum_{k=1}^{K_A}\|C_e\boldsymbol{w}_A + C_f\boldsymbol{w}_B - \boldsymbol{w}_k\|^2$. First, we have

$$\begin{aligned}
&\frac{1}{K_A}\sum_{k=1}^{K_A}\|C_e\boldsymbol{w}_A + C_f\boldsymbol{w}_B - \boldsymbol{w}_k\|^2 \\
=&\frac{1}{K_A}\sum_{k=1}^{K_A}\|\boldsymbol{w}_k\|^2 - 2\frac{1}{K_A}\sum_{k=1}^{K_A}\boldsymbol{w}_k\cdot(C_e\boldsymbol{w}_A + C_f\boldsymbol{w}_B) + \|C_e\boldsymbol{w}_A + C_f\boldsymbol{w}_B\|^2 \\
\overset{a}{=}&\frac{1}{K_A}\sum_{k=1}^{K_A}\|\boldsymbol{w}_k\|^2 - 2C_f^2\boldsymbol{w}_A^\intercal\boldsymbol{w}_B - C_e(2-C_e)\|\boldsymbol{w}_A\|^2 + C_f^2\|\boldsymbol{w}_B\|^2
\end{aligned} \tag{33}$$

where $\overset{a}{=}$ uses $\sum_{k=1}^{K_A}\boldsymbol{w}_A = K_A\boldsymbol{w}_A$. Then, using the constraint that $\sum_{k=1}^{K_A}\|\boldsymbol{w}_A\|^2 \leq KE_W$ yields that

$$\begin{aligned}
&\frac{1}{K_A}\sum_{k=1}^{K_A}\|\boldsymbol{w}_k\|^2 - 2C_f^2\boldsymbol{w}_A^\intercal\boldsymbol{w}_B - C_e(2-C_e)\|\boldsymbol{w}_A\|^2 + C_f^2\|\boldsymbol{w}_B\|^2 \\
\leq&\frac{K}{K_A}E_W - \frac{1}{K_A}\sum_{k=K_A+1}K\|\boldsymbol{w}_k\|^2 - C_e(2-C_e)\|\boldsymbol{w}_A + \frac{C_f^2}{C_e(2-C_e)}\boldsymbol{w}_B\|^2 + \left(C_f^2 + \frac{C_f^4}{C_e(2-C_e)}\right)\|w_B\|^2 \\
\overset{a}{=}&\frac{K}{K_A}E_W - \left(1/K_R - C_f^2 - \frac{C_f^4}{C_e(2-C_e)}\right)\|w_B\|^2 - \frac{1}{K_A}\sum_{k=K_A+1}^{K}\|\boldsymbol{w}_k - \boldsymbol{w}_B\|^2,
\end{aligned} \tag{34}$$

where $\overset{a}{=}$ applies $\sum_{k=K_A+1}K\|\boldsymbol{w}_k\|^2 = K_B\|\boldsymbol{w}_B\|^2 + \sum_{k=K_A+1}K\|\boldsymbol{w}_k - \boldsymbol{w}_B\|^2$. Plugging Equation (33) and Equation (34) into $\overset{a}{\geq}$ in Equation (29), we obtain $\overset{b}{\geq}$ in Equation (29), completing the proof. □

G.4 Proof of Proposition 4.3

Restated Proposition 4.3. Assume $p \geq 2K$ and the loss function $\mathcal{L}$ is convex in the first argument. Let $\boldsymbol{X}^\star$ be any minimizer of the convex program (Equation (4)) with $n_{(1,1)} = n_{(1,2)} = \cdots = n_{(K_A,K_A)} = n_A$ and $n_{(K_A+1,K_A+1)} = n_{(K_A+1,K_A+2)} = \cdots = n_{(K,K)} = w_r n_B$. Define $\left(\boldsymbol{W}^\star, \boldsymbol{H}^{\lambda\star}\right)$ as

$$\left[\boldsymbol{h}_{(1,1)}^{\lambda\star}, \boldsymbol{h}_{(1,2)}^{\lambda\star}, \ldots, \boldsymbol{h}_{(K,K)}^{\lambda\star},\ (\boldsymbol{W}^\star)^\top\right] = \boldsymbol{P}(\boldsymbol{X}^\star)^{1/2},$$

$$\boldsymbol{h}_{k_A,i}^{\lambda\star} = \boldsymbol{h}_{k_A}^{\lambda\star},\ \text{ for all } i \in \mathcal{I}_{k_A}^{\lambda}, k_A \in \mathbb{K}_A^2,\ \ \boldsymbol{h}_{k_B,i}^{\lambda\star} = \boldsymbol{h}_{k_B}^{\lambda\star},\ \text{ for all } i \in \mathcal{I}_{k_B}^{\lambda}, k_B \in \mathbb{K}_B^2,$$

where $\boldsymbol{P} \in \mathbb{R}^{p\times(K^2+K)}$ is any partial orthogonal matrix such that $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_{K^2+K}$. Then, $(\boldsymbol{W}^\star, \boldsymbol{H}^{\lambda\star})$ is a global minimizer of the mixed-label balanced UFM$_\lambda$ (Equation (6)). Moreover, if all $\boldsymbol{X}^\star$'s satisfy $\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\boldsymbol{X}^\star(k,k) = E_H$, then all the solutions of Equation (6) are in the form of Equation (7).

Proof. For any feasible solution $\left(\boldsymbol{W}, \boldsymbol{H}^{\lambda}\right)$ for the original program Equation (2), we define

$$\boldsymbol{h}_{k_A}^{\lambda} := \frac{1}{n_A}\sum_{i=1}^{n_A}\boldsymbol{h}_{k_A,i},\ k_A \in \mathbb{K}_A^2, \text{ and } \boldsymbol{h}_{k_B}^{\lambda} := \frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\boldsymbol{h}_{k_B,i},\ k_B \in \mathbb{K}_B^2,$$

and

$$\boldsymbol{X} := \left[\boldsymbol{h}^{\lambda}_{(1,1)}, \boldsymbol{h}^{\lambda}_{(1,2)}, \ldots, \boldsymbol{h}^{\lambda}_{(K,K)}, \boldsymbol{W}^{\top}\right]^{\top} \left[\boldsymbol{h}^{\lambda}_{(1,1)}, \boldsymbol{h}^{\lambda}_{(1,2)}, \ldots, \boldsymbol{h}^{\lambda}_{(K,K)}, \boldsymbol{W}^{\top}\right].$$

Clearly, $\boldsymbol{X} \succeq 0$. For the other two constraints of Equation (4), we have

$$\begin{aligned}
\frac{1}{K^2}\sum_{k=1}^{K^2} \boldsymbol{X}(k,k) &= \frac{1}{K^2}\sum_{k\in\mathbb{K}^2} \|\boldsymbol{h}^{\lambda}_{k}\|^2 \\
&\overset{a}{\leq} \frac{1}{K^2}\left(\sum_{k_A\in\mathbb{K}^2_A} \frac{1}{n_A}\sum_{i=1}^{n_A} \left\|\boldsymbol{h}^{\lambda}_{k_A,i}\right\|^2 + \sum_{k_B\in\mathbb{K}^2_B} \frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B} \left\|\boldsymbol{h}^{\lambda}_{k_B,i}\right\|^2\right) \\
&\overset{b}{\leq} E_H
\end{aligned}$$

and

$$\frac{1}{K}\sum_{k=K^2+1}^{K^2+K} \boldsymbol{X}(k,k) = \frac{1}{K}\sum_{k=1}^{K} \|\boldsymbol{w}_k\|^2 \overset{c}{\leq} E_W,$$

where $\overset{a}{\leq}$ applies Jensen's inequality and $\overset{b}{\leq}$ and $\overset{c}{\leq}$ use that $\left(\boldsymbol{W}, \boldsymbol{H}^{\lambda}\right)$ is a feasible solution. So $\boldsymbol{X}$ is a feasible solution for the convex program Equation (4). Letting L_0 be the global minimum of Equation (4), for any feasible solution $\left(\boldsymbol{W}, \boldsymbol{H}^{\lambda}\right)$, we obtain

$$\begin{aligned}
&\frac{1}{N}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n_k} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}^{\lambda}_{k,i}, \boldsymbol{y}^{\lambda}_{k}) \\
&= \sum_{k_A\in\mathbb{K}^2_A} \frac{n_A}{N}\left[\frac{1}{n_A}\sum_{i=1}^{n_A} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}^{\lambda}_{k_A,i}, \boldsymbol{y}^{\lambda}_{k_A})\right] + \sum_{k_B\in\mathbb{K}^2_B} \frac{w_r n_B}{N}\left[\frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}^{\lambda}_{k_B,i}, \boldsymbol{y}^{\lambda}_{k_B})\right] \\
&\overset{a}{\geq} \sum_{k_A\in\mathbb{K}^2_A} \frac{n_A}{N}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}^{\lambda}_{k_A}, \boldsymbol{y}^{\lambda}_{k_A}) + \sum_{k_B\in\mathbb{K}^2_B} \frac{w_r n_B}{N}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}^{\lambda}_{k_B}, \boldsymbol{y}^{\lambda}_{k_B}) \\
&= \sum_{k_A\in\mathbb{K}^2_A} \frac{n_A}{N}\mathcal{L}(\boldsymbol{z}(k_A)^{\lambda}, \boldsymbol{y}^{\lambda}_{k_A}) + \sum_{k_B\in\mathbb{K}^2_B} \frac{w_r n_B}{N}\mathcal{L}(\boldsymbol{z}(k_B)^{\lambda}, \boldsymbol{y}^{\lambda}_{k_B}) \geq L_0,
\end{aligned} \tag{35}$$

where in $\overset{a}{\geq}$, we use $\mathcal{L}$ is convex on the first argument, and so $\mathcal{L}(\boldsymbol{W}\boldsymbol{h}^{\lambda}, \boldsymbol{y}^{\lambda}_{k})$ is convex on $\boldsymbol{h}$ given $\boldsymbol{W}$ and $k \in \mathbb{K}^2$.

For the simplicity of our expressions, we hereafter remove the superscript λ of $\boldsymbol{H}^{\lambda}$, $\boldsymbol{h}^{\lambda}$ and $\boldsymbol{z}^{\lambda}$.

On the other hand, considering the solution $(\boldsymbol{W}^{\star}, \boldsymbol{H}^{\star})$ defined in Equation (5) with $\boldsymbol{X}^{\star}$ being a minimizer of Equation (4), we have $\left[\boldsymbol{h}^{\star}_{(1,1)}, \boldsymbol{h}^{\star}_{(1,2)}, \ldots, \boldsymbol{h}^{\star}_{(K,K)}, \boldsymbol{W}^{\top}\right]^{\top} \left[\boldsymbol{h}^{\star}_{(1,1)}, \boldsymbol{h}^{\star}_{(1,2)}, \ldots, \boldsymbol{h}^{\star}_{(K,K)}, \boldsymbol{W}^{\top}\right] = \boldsymbol{X}^{\star}$ ($p \geq 2K$ guarantees the existence of $\left[\boldsymbol{h}^{\star}_{(1,1)}, \boldsymbol{h}^{\star}_{(1,2)}, \ldots, \boldsymbol{h}^{\star}_{(K,K)}, (\boldsymbol{W}^{\star})^{\top}\right]$). We can verify that $(\boldsymbol{W}^{\star}, \boldsymbol{H}^{\star})$ is a feasible solution for Equation (2) and have

$$\frac{1}{N}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n_k} \mathcal{L}(\boldsymbol{W}^{\star}\boldsymbol{h}^{\star}_{k,i}, \boldsymbol{y}^{\lambda}_{k}) = \sum_{k_A\in\mathbb{K}^2_A} \frac{n_A}{N}\mathcal{L}(\boldsymbol{z}(k_A)^{\star}, \boldsymbol{y}^{\lambda}_{k_A}) + \sum_{k_B\in\mathbb{K}^2_B} \frac{w_r n_B}{N}\mathcal{L}(\boldsymbol{z}(k_B)^{\star}, \boldsymbol{y}^{\lambda}_{k_B}) = L_0, \tag{36}$$

where $\boldsymbol{z}(k_A)^{\star} = \left[\boldsymbol{X}^{\star}(k_A, K^2+1), \boldsymbol{X}^{\star}(k_A, K^2+2), \ldots, \boldsymbol{X}^{\star}(k_A, K^2+K_A)\right]^{\top}$ for $k_A \in \mathbb{K}^2_A$ and $\boldsymbol{z}(k_B)^{\star} = \left[\boldsymbol{X}^{\star}(k_B, K^2+K_A+1), \boldsymbol{X}^{\star}(k_B, K^2+K_A+2), \ldots, \boldsymbol{X}^{\star}(k_B, K^2+K)\right]^{\top}$ for $k_B \in \mathbb{K}^2_B$.

Combining Equation (35) and Equation (36), we conclude that L_0 is the global minimum of Equation (2) and $(\boldsymbol{W}^{\star}, \boldsymbol{H}^{\star})$ is a minimizer.

Suppose there is a minimizer $(\boldsymbol{W}', \boldsymbol{H}')$ that cannot be written as Equation (5). Let

$$\boldsymbol{h}'_{k_A} = \frac{1}{n_A}\sum_{i=1}^{n_A} \boldsymbol{h}'_{k_A,i},\ k_A \in \mathbb{K}^2_A, \text{ and } \boldsymbol{h}'_{k_B} = \frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B} \boldsymbol{h}'_{k_B,i}, k_B \in \mathbb{K}^2_B$$

and

$$\boldsymbol{X}' = \left[\boldsymbol{h}'_{(1,1)}, \boldsymbol{h}'_{(1,2)}, \ldots, \boldsymbol{h}'_{(K,K)}, (\boldsymbol{W}')^\top\right]^\top \left[\boldsymbol{h}'_{(1,1)}, \boldsymbol{h}'_{(1,2)}, \ldots, \boldsymbol{h}'_{(K,K)}, (\boldsymbol{W}')^\top\right].$$

Equation (35) implies that $\boldsymbol{X}'$ is a minimizer of Equation (4). As $(\boldsymbol{W}', \boldsymbol{H}')$ cannot be written as Equation (5) with $\boldsymbol{X}^\star = \boldsymbol{X}'$, then there is a $k' \in \mathbb{K}^2$, $i, j \in [n'_k]$ with $i \neq j$ such that $\boldsymbol{h}_{k',i} \neq \boldsymbol{h}_{k',j}$. We have

$$\begin{aligned}
&\frac{1}{K^2}\sum_{k=1}^{K^2} \boldsymbol{X}'(k,k) = \frac{1}{K^2}\sum_{k\in\mathbb{K}^2} \|\boldsymbol{h}'_k\|^2 \\
=&\frac{1}{K^2}\sum_{k_A\in\mathbb{K}_A^2}\frac{1}{n_A}\sum_{i=1}^{n_A}\left\|\boldsymbol{h}'_{k_A,i}\right\|^2 - \frac{1}{K^2}\sum_{k_A\in\mathbb{K}_A^2}\frac{1}{n_A}\sum_{i=1}^{n_A}\|\boldsymbol{h}'_{k_A,i} - \boldsymbol{h}'_{k_A}\|^2 \\
&+\frac{1}{K^2}\sum_{k_B\in\mathbb{K}_B^2}\frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\left\|\boldsymbol{h}'_{k_B,i}\right\|^2 - \frac{1}{K^2}\sum_{k_B\in\mathbb{K}_B^2}\frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\|\boldsymbol{h}'_{k_B,i} - \boldsymbol{h}'_{k_B}\|^2 \\
\leq&\frac{1}{K^2}\sum_{k_A\in\mathbb{K}^2}\frac{1}{n_A}\sum_{i=1}^{n_A}\left\|\boldsymbol{h}'_{k_A,i}\right\|^2 - \frac{1}{K^2}\frac{1}{n_{k'_A}}(\|\boldsymbol{h}'_{k'_A,i} - \boldsymbol{h}'_{k'_A}\|^2 + \|\boldsymbol{h}'_{k'_A,j} - \boldsymbol{h}'_{k'_A}\|^2) \\
&+\frac{1}{K^2}\sum_{k_B\in\mathbb{K}_B^2}\frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\left\|\boldsymbol{h}'_{k_B,i}\right\|^2 - \frac{1}{K^2}\frac{1}{n_{k'_B}}(\|\boldsymbol{h}'_{k'_B,i} - \boldsymbol{h}'_{k'_B}\|^2 + \|\boldsymbol{h}'_{k'_B,j} - \boldsymbol{h}'_{k'_B}\|^2) \\
\leq&\frac{1}{K^2}\sum_{k_A\in\mathbb{K}_A^2}\frac{1}{n_A}\sum_{i=1}^{n_A}\left\|\boldsymbol{h}'_{k_A,i}\right\|^2 - \frac{1}{K^2}\frac{1}{2n_{k'_A}}\|\boldsymbol{h}'_{k'_A,i} - \boldsymbol{h}'_{k'_A,j}\|^2 \\
&+\frac{1}{K^2}\sum_{k_B\in\mathbb{K}_B^2}\frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\left\|\boldsymbol{h}'_{k_B,i}\right\|^2 - \frac{1}{K^2}\frac{1}{2n_{k'_B}}\|\boldsymbol{h}'_{k'_B,i} - \boldsymbol{h}'_{k'_B,j}\|^2 \\
<&E_H.
\end{aligned}$$

By contraposition, if all $\boldsymbol{X}^\star$ satisfy that $\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}^\star(k,k) = E_H$, then all the solutions of Equation (2) are in the form of Equation (5). We complete the proof. □

G.5 Proof of Theorem G.5

Definition G.4. A K-simplex ETF is a collection of points in $\mathbb{R}^p$ specified by the columns of the matrix

$$\boldsymbol{M}^\star = \sqrt{\frac{K}{K-1}}\boldsymbol{P}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\intercal\right)$$

where $\boldsymbol{I}_K \in \mathbb{R}^{K\times K}$ is the identity matrix, $\mathbf{1}_K$ is the ones vector, and $\boldsymbol{P} \in \mathbb{R}^{p\times K}$ $(p \geq K)$ is a partial orthogonal matrix such that $\boldsymbol{P}^\intercal\boldsymbol{P} = \boldsymbol{I}_K$.

Theorem G.5. *In the balanced case, although $(\boldsymbol{W}^\lambda, \boldsymbol{H}^\lambda)$ are linearly dependent on $(\boldsymbol{W}, \boldsymbol{H})$ in Equation (13), any global minimizer $\boldsymbol{W}^\star \equiv [\boldsymbol{w}_1^\star, ..., \boldsymbol{w}_K^\star]$, $\boldsymbol{H}^\star \equiv \left[\boldsymbol{h}_{k,i}^\star : 1 \leq k \leq K, 1 \leq i \leq n\right]$ of Equation (13) with the cross-entropy loss obeys*

$$\boldsymbol{h}_{k,i}^\star = C\boldsymbol{w}_k^\star = C'\boldsymbol{m}_k^\star \tag{37}$$

for all $1 \leq i \leq n$, $1 \leq k \leq K$, where the constants $C = \sqrt{E_H/E_W}$, $C' = \sqrt{E_H}$, and the matrix $[\boldsymbol{m}_1^\star, ..., \boldsymbol{m}_K^\star]$ forms a K-simplex ETF specified in Definition G.4

Because there are multiplication of variables in the objective functions, Equation (13) is non-convex. Thus, the KKT condition is not sufficient for optimality. To prove Theorem G.5, we directly determine the global minimum of Equation (13). During this procedure, one key step is to show that minimizing Equation (13) is equivalent to minimize a symmetric quadratic function:

$$\sum_{i=1}^{n}\left[\left(\sum_{k\in\mathbb{K}^2}\boldsymbol{h}_{k,i}^\lambda\right)^\intercal\left(\sum_{k\in\mathbb{K}^2}\boldsymbol{w}_k^\lambda\right) - K^2\sum_{k\in\mathbb{K}^2}\boldsymbol{h}_{k,i}^{\lambda\intercal}\boldsymbol{w}_k^\lambda\right]$$

under suitable conditions. The detail is shown below.

Proof. By the concavity of $\log(\cdot)$, for any $\boldsymbol{z} \in \mathbb{R}^{K^2}$, $k \in [K^2]$, constants $C_a,\ C_b > 0$, letting $C_c = \frac{C_b}{(C_a+C_b)(K^2-1)}$, we have

$$\begin{aligned}
-\log\left(\frac{\boldsymbol{z}(k)}{\sum_{k'=1}^{K^2}\boldsymbol{z}(k')}\right) &= -\log(\boldsymbol{z}(k)) + \log\left(\sum_{k'=1}^{K^2}\boldsymbol{z}(k')\right) \\
&= -\log(\boldsymbol{z}(k)) + \log\left(\frac{C_a}{C_a+C_b}\left(\frac{(C_a+C_b)\boldsymbol{z}(k)}{C_a}\right) + C_c\sum_{k'=1,k'\neq k}^{K^2}\frac{\boldsymbol{z}(k')}{C_c}\right).
\end{aligned} \tag{38}$$

Recognizing the equality

$$\frac{C_a}{C_a+C_b} + \underbrace{C_c+\cdots+C_c}_{K^2-1} = \frac{C_a}{C_a+C_b} + (K^2-1)\frac{C_b}{(C_a+C_b)(K^2-1)} = 1$$

and the concavity of $\log(\cdot)$, we see that the Jensen inequality gives

$$\begin{aligned}
&\log\left(\frac{C_a}{C_a+C_b}\left(\frac{(C_a+C_b)\boldsymbol{z}(k)}{C_a}\right) + C_c\sum_{k'=1,k'\neq k}^{K^2}\frac{\boldsymbol{z}(k')}{C_c}\right) \\
&\geq \frac{C_a}{C_a+C_b}\log\left(\frac{(C_a+C_b)\boldsymbol{z}(k)}{C_a}\right) + C_c\sum_{k'=1,k'\neq k}^{K^2}\log\left(\frac{\boldsymbol{z}(k')}{C_c}\right).
\end{aligned} \tag{39}$$

Plugging this inequality into Equation (38), we get

$$\begin{aligned}
-\log\left(\frac{\boldsymbol{z}(k)}{\sum_{k'=1}^{K^2}\boldsymbol{z}(k')}\right) &\geq -\log(\boldsymbol{z}(k)) + \frac{C_a}{C_a+C_b}\log\left(\frac{(C_a+C_b)\boldsymbol{z}(k)}{C_a}\right) + C_c\sum_{k'=1,k'\neq k}^{K^2}\log\left(\frac{\boldsymbol{z}(k')}{C_c}\right) \\
&= -\frac{C_a}{C_a+C_b}\left[\log(\boldsymbol{z}(k)) - \frac{1}{K^2-1}\sum_{k'=1,k'\neq k}^{K^2}\log(\boldsymbol{z}(k'))\right] + C_d,
\end{aligned}$$

where the constant $C_d := \frac{C_a}{C_a+C_b}\log\left(\frac{C_a+C_b}{C_a}\right) + \frac{C_b}{C_a+C_b}\log(1/C_c)$. Note that in Equation (38), C_a and C_b can be any positive numbers. To prove Theorem G.5, we set $C_a := \exp(\sqrt{E_H E_W})$ and $C_b := \exp(-\sqrt{E_H E_W}/(K^2-1))$, which shall lead to the tightest lower bound for the objective of Equation (13). Applying Equation (38) to the objective, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{W}^\lambda\boldsymbol{h}_{k,i}^\lambda, \boldsymbol{y}_k^\lambda) \\
&\geq \frac{C_b}{(C_a+C_b)N(K^2-1)}\sum_{i=1}^{n}\left[\left(\sum_{k\in\mathbb{K}^2}\boldsymbol{h}_{k,i}^\lambda\right)^\intercal\left(\sum_{k\in\mathbb{K}^2}\boldsymbol{w}_k^\lambda\right) - K^2\sum_{k\in\mathbb{K}^2}\boldsymbol{h}_{k,i}^{\lambda\intercal}\boldsymbol{w}_k^\lambda\right] + C_d.
\end{aligned} \tag{40}$$

Defining $\bar{\boldsymbol{h}}_i^\lambda := \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\boldsymbol{h}_{k,i}^\lambda$ for $i \in [n]$, it follows from the simple inequality $2ab \leq a^2+b^2$ that

$$\begin{aligned}
&\sum_{i=1}^{n}\left[\left(\sum_{k\in\mathbb{K}^2}\boldsymbol{h}_{k,i}^\lambda\right)^\intercal\left(\sum_{k\in\mathbb{K}^2}\boldsymbol{w}_k^\lambda\right) - K^2\sum_{k\in\mathbb{K}^2}\boldsymbol{h}_{k,i}^{\lambda\intercal}\boldsymbol{w}_k^\lambda\right] \\
&= K^2\sum_{i=1}^{n}\sum_{k\in\mathbb{K}^2}(\bar{\boldsymbol{h}}_i^\lambda - \boldsymbol{h}_{k,i}^\lambda)^\intercal\boldsymbol{w}_k^\lambda \\
&\geq -\frac{K^2}{2}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n}\|\bar{\boldsymbol{h}}_i^\lambda - \boldsymbol{h}_{k,i}^\lambda\|^2/C_e - \frac{C_e N}{2}\sum_{k\in\mathbb{K}^2}\|\boldsymbol{w}_k^\lambda\|^2,
\end{aligned} \tag{41}$$

where we pick $C_e := \sqrt{E_H/E_W}$. The two terms in the right hand side of Equation (41) can be bounded via the constrains of Equation (13). Specifically, we have

$$\frac{C_e N}{2} \sum_{k\in\mathbb{K}^2} \|\boldsymbol{w}_k^\lambda\|^2 \leq \frac{K^2 N\sqrt{E_H E_W}}{2}, \tag{42}$$

and

$$\begin{aligned}\frac{K^2}{2} \sum_{k\in\mathbb{K}^2} \sum_{i=1}^{n} \|\bar{\boldsymbol{h}}_i^\lambda - \boldsymbol{h}_{k,i}^\lambda\|^2/C_e \overset{a}{=} & \frac{K^4}{2C_e} \sum_{i=1}^{N} \left(\frac{1}{K^2} \sum_{k\in\mathbb{K}^2} \|\boldsymbol{h}_{k,i}^\lambda\|^2 - \|\bar{h}_i^\lambda\|^2 \right) \\ \leq & \frac{K^2}{2C_e} \sum_{k\in\mathbb{K}^2} \sum_{i=1}^{N} \|\boldsymbol{h}_{k,i}^\lambda\|^2 \leq \frac{K^2 N\sqrt{E_H E_W}}{2}, \end{aligned} \tag{43}$$

where $\overset{a}{=}$ uses the fact that $\mathbb{E}\|\boldsymbol{a} - \mathbb{E}[\boldsymbol{a}]\|^2 = \mathbb{E}\|\boldsymbol{a}\|^2 - \|\mathbb{E}[\boldsymbol{a}]\|^2$. Thus, plugging Equation (41), Equation (42), and Equation (43) into Equation (40), we have

$$\frac{1}{N} \sum_{k\in\mathbb{K}^2} \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{W}^\lambda \boldsymbol{h}_{k,i}^\lambda, \boldsymbol{y}_k^\lambda) \geq -\frac{C_b}{(C_a + C_b)} \frac{K^2\sqrt{E_H E_W}}{K^2 - 1} + C_d := L_0. \tag{44}$$

Now, we check the conditions that reduce Equation (44) to an equality.

By the strict concavity of $\log(\cdot)$, Equation (39) reduces to an equality only if

$$\frac{(C_a + C_b)\boldsymbol{z}(k)}{C_a} = \frac{\boldsymbol{z}(k')}{C_c}$$

for $k' \neq k$. Therefore, Equation (40) reduces to an equality only if

$$\frac{(C_a + C_b)\boldsymbol{h}_{k,i}^{\lambda\intercal} \boldsymbol{w}_k^\lambda}{C_a} = \frac{\boldsymbol{h}_{k,i}^{\lambda\intercal} \boldsymbol{w}_{k'}^\lambda}{C_c}.$$

Recognizing $C_c = \frac{C_b}{(C_a+C_b)(K^2-1)}$ and taking the logarithm of both sides of the above equation, we obtain

$$\boldsymbol{h}_{k,i}^{\lambda\intercal} \boldsymbol{w}_k^\lambda = \boldsymbol{h}_{k,i}^{\lambda\intercal} \boldsymbol{w}_{k'}^\lambda + \log\left(\frac{C_a(K^2-1)}{C_b}\right),$$

for all $(k, i, k') \in \{(k, i, k') : k \in \mathbb{K}^2, k' \in \mathbb{K}^2, k' \neq k, i \in [n]\}$. Equation (41) becomes equality if and only if

$$\bar{\boldsymbol{h}}_i^\lambda - \boldsymbol{h}_{k,i}^\lambda = -C_e \boldsymbol{w}_k^\lambda, \quad k \in \mathbb{K}^2,\ i \in [n].$$

By the definition of $\boldsymbol{w}_k^\lambda$ and $\boldsymbol{h}_{k,i}^\lambda$, $\bar{\boldsymbol{h}}_i^\lambda = \frac{1}{K^2}\sum_{k\in\mathbb{K}^2} \boldsymbol{h}_{k,i}^\lambda = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{h}_{k,i}$. Defining $\bar{\boldsymbol{h}}_i := \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{h}_{k,i}$ and plugging this equality into the above equation, we get

$$\bar{\boldsymbol{h}}_i - \lambda \boldsymbol{h}_{a,i} - (1-\lambda)\boldsymbol{h}_{b,i} = -C_e\left(\lambda \boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_b\right) \tag{45}$$

For $a \neq b \neq c$, we define

$$\bar{\boldsymbol{h}}_i - \lambda \boldsymbol{h}_{a,i} - (1-\lambda)\boldsymbol{h}_{c,i} = -C_e\left(\lambda \boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_c\right) \tag{46}$$

$$\bar{\boldsymbol{h}}_i - \lambda \boldsymbol{h}_{b,i} - (1-\lambda)\boldsymbol{h}_{c,i} = -C_e\left(\lambda \boldsymbol{w}_b + (1-\lambda)\boldsymbol{w}_c\right) \tag{47}$$

Subtracting Equation (46) from Equation (47), we get

$$\boldsymbol{h}_{a,i} = \boldsymbol{h}_{b,i} + C_e(\boldsymbol{w}_a - \boldsymbol{w}_b).$$

Plugging this equality to Equation (45), we get

$$\bar{\boldsymbol{h}}_i - \boldsymbol{h}_{b,i} = -C_e \boldsymbol{w}_b,$$

which is the same result of the balanced case where the number of classes is K. As a result, Equation (41) becomes equality if and only if

$$\bar{\boldsymbol{h}}_i - \boldsymbol{h}_{k,i} = -C_e \boldsymbol{w}_k, \quad k \in [K],\ i \in [n]. \tag{48}$$

The remainder of the proof is identical to that in [7]. However, for the sake of clarity, we present it here in full rather than omitting it.

Applying Lemma G.6 shown in the below, we have $(\boldsymbol{W}, \boldsymbol{H})$, which satisfies Equation (37).

Reversely, it is easy to verify that Equation (44) reduces to equality when $(\boldsymbol{W}, \boldsymbol{H})$ admits Equation (37). So, L_0 is the global minimum of Equation (13) and Equation (37) is the unique form for the minimizers. We complete the proof of Theorem G.5. □

Lemma G.6. *Suppose* $(\boldsymbol{W}, \boldsymbol{H})$ *satisfies*

$$\bar{\boldsymbol{h}}_i - \boldsymbol{h}_{k,i} = -\sqrt{\frac{E_H}{E_W}}\boldsymbol{w}_k, \quad k \in [K],\ i \in [n], \tag{49}$$

and

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{h}_{k,i}\|^2 = E_H,\ \ \frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2 = E_W,\ \ \bar{\boldsymbol{h}}_i = \boldsymbol{0}_p,\ i \in [n], \tag{50}$$

where $\bar{\boldsymbol{h}}_i := \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{h}_{k,i}$ *with* $i \in [n]$. *Moreover, there exists a constant* C *such that for all* $\{(k, i, k') : k \in [K], k' \in [K], k' \neq k, i \in [n]\}$, *we have*

$$\boldsymbol{h}_{k,i} \cdot \boldsymbol{w}_k = \boldsymbol{h}_{k,i} \cdot \boldsymbol{w}_{k'} + C. \tag{51}$$

Then, $(\boldsymbol{W}, \boldsymbol{H})$ *satisfies Equation* (37).

Proof. Combining Equation (49) with the last equality in Equation (50), we have

$$\boldsymbol{W} = \sqrt{\frac{E_W}{E_H}}\,[\boldsymbol{h}_1, ..., \boldsymbol{h}_K]^\intercal, \quad \boldsymbol{h}_{k,i} = \boldsymbol{h}_k,\ \ k \in [K],\ i \in [n].$$

Thus, it remains to show

$$\boldsymbol{W} = \sqrt{E_W}(\boldsymbol{M}^*)^\intercal, \tag{52}$$

where $\boldsymbol{M}^*$ is a K-simplex ETF.

Plugging $\boldsymbol{h}_k = \boldsymbol{h}_{k,i} = \sqrt{\frac{E_W}{E_H}}\boldsymbol{w}_k$ into Equation (51), we have, for all $(k, k') \in \{(k, k') : k \in [K], k' \in [K], k' \neq k\}$,

$$\sqrt{\frac{E_W}{E_H}}\|\boldsymbol{w}_k\|^2 = \boldsymbol{h}_{k,i} \cdot \boldsymbol{w}_k = \boldsymbol{h}_{k,i} \cdot \boldsymbol{w}_{k'} + C = \sqrt{\frac{E_W}{E_H}}\boldsymbol{w}_k\boldsymbol{w}_{k'} + C,$$

and

$$\sqrt{\frac{E_W}{E_H}}\|\boldsymbol{w}_{k'}\|^2 = \boldsymbol{h}_{k',i} \cdot \boldsymbol{w}_{k'} = \boldsymbol{h}_{k',i} \cdot \boldsymbol{w}_k + C = \sqrt{\frac{E_W}{E_H}}\boldsymbol{w}_{k'}\boldsymbol{w}_k + C.$$

Therefore, from $\frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2 = E_W$, we have $\|\boldsymbol{w}_k\| = \sqrt{E_W}$ and $\boldsymbol{h}_k\boldsymbol{w}_{k'} = C' := \sqrt{E_H E_W} - C$.

Furthermore, recalling that $\bar{\boldsymbol{h}}_i = \boldsymbol{0}_p$ for $i \in [n]$, we have $\sum_{k=1}^{K}\boldsymbol{h}_k = \boldsymbol{0}_p$, which further yields $\sum_{k=1}^{K}\boldsymbol{h}_k \cdot \boldsymbol{w}_{k'} = 0$ for $k' \in [K]$. Then, it follows from $\boldsymbol{h}_k\boldsymbol{w}_{k'} = C'$ and $\boldsymbol{h}_k\boldsymbol{w}_k = \sqrt{E_H E_W}$ that $\boldsymbol{h}_k\boldsymbol{w}_{k'} = -\sqrt{E_H E_W}/(K-1)$. Thus, we obtain

$$\boldsymbol{W}\boldsymbol{W}^\top = \sqrt{\frac{E_W}{E_H}}\boldsymbol{W}\,[\boldsymbol{h}_1, ..., \boldsymbol{h}_K] = E_W\left[\frac{K}{K-1}\left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{1}_K\boldsymbol{1}_K^\top\right)\right],$$

which implies Equation (52). We complete the proof. □

G.6 Proof of Proposition G.7

The proposed $a < b$ method is motivated by two questions: whether mixed labels (a, b) and (b, a) can be treated as equivalent, and whether training with only (a, b) without singleton labels still forms a simplex ETF. In this section, we demonstrate that even when only $a < b$ exists, the assumptions on all singleton classes (i.e., satisfying NC and forming a simplex ETF) continue to hold, and thus Theorem 4.2 and Proposition 4.3 remain satisfied. This thereby establishes that the $a < b$ setting is a special case of the general framework, as the following statement holds.

Proposition G.7. *Let $\mathbb{K}^<$ be the mixed label set where $a < b$ for all $(a, b) \in \mathbb{K}^2$ and $\boldsymbol{W}^\lambda_{\mathbb{K}^<}$ be the partial matrix of $\boldsymbol{W}^\lambda$ which has class vectors for mixed labels $(a, b) \in \mathbb{K}^<$.*

Then, $\boldsymbol{W}$ is a K-simplex ETF if $\boldsymbol{W}^\lambda_{\mathbb{K}^<}$ is a $|\mathbb{K}^<|$-simplex ETF.

For the simplicity, we remove the subscript $\mathbb{K}^<$ of $\boldsymbol{W}^\lambda_{\mathbb{K}^<}$ and $\boldsymbol{w}^\lambda_{\mathbb{K}^<}$ in the following proof.

Proof. Let $f(x; \alpha, \beta)$ be the probability density function of Beta distribution $D_\lambda(\alpha, \beta)$. For the mixup ratio λ sampled from $D_\lambda(\alpha, \alpha)$, we have

$$\begin{aligned}\boldsymbol{w}^\lambda_{(a,b)} =& \mathbb{E}_\lambda \left(\lambda \boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_b\right) \\ \overset{a}{=}& \frac{1}{2}\mathbb{E}_\lambda \left((\lambda \boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_b) + ((1-\lambda)\boldsymbol{w}_a + \lambda \boldsymbol{w}_b)\right) \\ =& \frac{1}{2}\left(\boldsymbol{w}_a + \boldsymbol{w}_b\right), \end{aligned} \tag{53}$$

where in $\overset{a}{=}$, we use $f(\lambda; \alpha, \alpha) = f(1-\lambda; \alpha, \alpha)$.

From the definition of a simplex ETF, we get

$$\sum_{(a,b)\in\mathbb{K}^<} \boldsymbol{w}^\lambda_{(a,b)} = 0 \tag{54}$$

Plugging the equality of Equation (53) into Equation (54), we have

$$\begin{aligned}\sum_{(a,b)\in\mathbb{K}^<} \boldsymbol{w}^\lambda_{(a,b)} =& \frac{K-1}{2}\sum_{i=1}^{K} \boldsymbol{w}_i = 0 \\ \therefore\ \boldsymbol{w}_i =& -\sum_{j\neq i}^{K} \boldsymbol{w}_j,\ \ \forall\, i \in [K] \end{aligned} \tag{55}$$

From the definition of $\mathbb{K}^<$, we can get $< i, j, k >$ for all $i \in [K]$, satisfying

$$\boldsymbol{w}_i = \boldsymbol{w}^\lambda_{\{i,j\}} - \boldsymbol{w}^\lambda_{\{j,k\}} + \boldsymbol{w}^\lambda_{\{i,k\}}, \tag{56}$$

where $\{a, b\} = (a, b)$ if $a < b$ otherwise (b, a) and $i \neq j \neq k$.

Now, we show that $\boldsymbol{w}_i^\top \boldsymbol{w}_{i'} = -\frac{1}{K-1}$ is true for all $i \neq i'$

$$\begin{aligned}\boldsymbol{w}_i^\top \boldsymbol{w}_{i'} \overset{Equation\ (56)}{=}& \left(\boldsymbol{w}^\lambda_{\{i,j\}} - \boldsymbol{w}^\lambda_{\{j,k\}} + \boldsymbol{w}^\lambda_{\{i,k\}}\right)^\top \left(\boldsymbol{w}^\lambda_{\{i',j'\}} - \boldsymbol{w}^\lambda_{\{j',k'\}} + \boldsymbol{w}^\lambda_{\{i',k'\}}\right) \\ \overset{a}{=}& -\frac{1}{K-1} \end{aligned} \tag{57}$$

where in $\overset{a}{=}$, we use the property of the simplex ETF, i.e., $\left(\boldsymbol{w}^\lambda_{(a,b)}\right)^\top \boldsymbol{w}^\lambda_{(a',b')} = -\frac{1}{K-1}$ for all $(a, b) \neq (a', b')$. We complete the proof. □

G.7 Proof of Theorem 5.1

We present a convex optimization program that serves as a relaxation of the non-convex UFM$_\lambda$-MS (Equation (13)), leveraging the established result that a quadratically constrained quadratic program can be transformed into the following semi-definite program [34]:

$$\begin{aligned}\min_{\boldsymbol{X}} & \sum_{k\in\mathbb{K}^2} \frac{n_k}{N} \mathcal{L}(\boldsymbol{W}^\lambda \boldsymbol{h}^\lambda, \boldsymbol{y}^\lambda_k) \\ \text{s.t. } & \frac{1}{K^2}\sum_{k=1}^{K^2} \boldsymbol{X}(k,k) \leq E_H, \\ & \frac{1}{K^2}\sum_{k=K^2+1}^{2K^2} \boldsymbol{X}(k,k) \leq E_W, \end{aligned} \tag{58}$$

where $\boldsymbol{X}^\lambda \succeq 0 \in \mathbb{R}^{2K^2 \times 2K^2}$ and $\boldsymbol{z}_k^\lambda = \left[\boldsymbol{X}^\lambda(k, K^2+1), \boldsymbol{X}^\lambda(k, K^2+2), \ldots, \boldsymbol{X}^\lambda(k, 2K^2)\right]^\top$ for all $1 \le k \le K^2$.

Lemma G.8. *Assume $p \ge 2K$ and the loss function $\mathcal{L}$ is convex in its first argument. Let $\boldsymbol{X}^{\lambda\star}$ be a minimizer of the convex program (Equation (58)). Define $\left(\boldsymbol{W}^{\lambda\star}, \boldsymbol{H}^{\lambda\star}\right)$ as*

$$\left[\boldsymbol{h}_{(1,1)}^{\lambda\star}, \boldsymbol{h}_{(1,2)}^{\lambda\star}, \ldots, \boldsymbol{h}_{(K,K)}^{\lambda\star},\ (\boldsymbol{W}^{\lambda\star})^\top\right] = \boldsymbol{P}(\boldsymbol{X}^{\lambda\star})^{1/2},$$

$$\boldsymbol{h}_{k,i}^\star = \boldsymbol{h}_k^\star,\ \ \textit{for all } i \in \mathcal{I}_k^\lambda, k \in \mathbb{K}^2,$$

where $(\boldsymbol{X}^{\lambda\star})^{1/2}$ denotes the positive square root of $\boldsymbol{X}^{\lambda\star}$ and $\boldsymbol{P} \in \mathbb{R}^{p \times 2K^2}$ is any partial orthogonal matrix such that $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}_{2K^2}$. Then, $(\boldsymbol{W}^{\lambda\star}, \boldsymbol{H}^{\lambda\star})$ is a minimizer of Equation (13). Moreover, if all $\boldsymbol{X}^{\lambda\star}$'s satisfy $\frac{1}{K^2}\sum_{k=1}^{K^2} \boldsymbol{X}^\star(k,k) = E_H$, then all the solutions of Equation (13) are in the form of Equation (58).

Proof. For any feasible solution $\left(\boldsymbol{W}^\lambda, \boldsymbol{H}^\lambda\right)$ for the original program Equation (13), we define

$$\boldsymbol{h}_k^\lambda := \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{h}_{k,i},\ k \in \mathbb{K}^2,$$

and

$$\boldsymbol{X}^\lambda := \left[\boldsymbol{h}_{(1,1)}^\lambda, \boldsymbol{h}_{(1,2)}^\lambda, \ldots, \boldsymbol{h}_{(K,K)}^\lambda, \boldsymbol{W}^{\lambda\top}\right]^\top \left[\boldsymbol{h}_{(1,1)}^\lambda, \boldsymbol{h}_{(1,2)}^\lambda, \ldots, \boldsymbol{h}_{(K,K)}^\lambda, \boldsymbol{W}^{\lambda\top}\right].$$

Clearly, $\boldsymbol{X}^\lambda \succeq 0$. For the other two constraints of Equation (58), we have

$$\frac{1}{K^2}\sum_{k=1}^{K^2} \boldsymbol{X}^\lambda(k,k) = \frac{1}{K^2}\sum_{k\in\mathbb{K}^2} \|\boldsymbol{h}_k^\lambda\|^2 \overset{a}{\le} \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}_{k,i}^\lambda\right\|^2 \overset{b}{\le} E_H$$

and

$$\frac{1}{K^2}\sum_{k=K^2+1}^{2K^2} \boldsymbol{X}^\lambda(k,k) = \frac{1}{K^2}\sum_{a=1}^{K}\sum_{b=1}^{K}\|\boldsymbol{w}_{(a,b)}^\lambda\|^2 \overset{c}{\le} E_W,$$

where $\overset{a}{\le}$ applies Jensen's inequality and $\overset{b}{\le}$ and $\overset{c}{\le}$ use that $\left(\boldsymbol{W}^\lambda, \boldsymbol{H}^\lambda\right)$ is a feasible solution. So $\boldsymbol{X}^\lambda$ is a feasible solution for the convex program Equation (58). Letting L_0 be the global minimum of Equation (58), for any feasible solution $\left(\boldsymbol{W}^\lambda, \boldsymbol{H}^\lambda\right)$, we obtain

$$\begin{aligned}\frac{1}{N}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n_k}\mathcal{L}(\boldsymbol{W}^\lambda\boldsymbol{h}_{k,i}^\lambda, \boldsymbol{y}_k^\lambda) &= \sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\left[\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{L}(\boldsymbol{W}^\lambda\boldsymbol{h}_{k,i}^\lambda, \boldsymbol{y}_k^\lambda)\right] \\ &\overset{a}{\ge} \sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\mathcal{L}(\boldsymbol{W}^\lambda\boldsymbol{h}_k^\lambda, \boldsymbol{y}_k^\lambda) = \sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\mathcal{L}(\boldsymbol{z}(k)^\lambda, \boldsymbol{y}_k^\lambda) \ge L_0, \end{aligned} \tag{59}$$

where in $\overset{a}{\ge}$, we use $\mathcal{L}$ is convex on the first argument, and so $\mathcal{L}(\boldsymbol{W}^\lambda\boldsymbol{h}^\lambda, \boldsymbol{y}_k^\lambda)$ is convex on $\boldsymbol{h}$ given $\boldsymbol{W}$ and $k \in \mathbb{K}^2$.

For the simplicity of our expressions, we hereafter remove the superscript λ of $\boldsymbol{X}^\lambda$, $\boldsymbol{W}^\lambda$, $\boldsymbol{H}^\lambda$, $\boldsymbol{h}^\lambda$ and $\boldsymbol{z}^\lambda$.

On the other hand, considering the solution $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ defined in Equation (58) with $\boldsymbol{X}^\star$ being a minimizer of Equation (58), we have $\left[\boldsymbol{h}_{(1,1)}^\star, \boldsymbol{h}_{(1,2)}^\star, \ldots, \boldsymbol{h}_{(K,K)}^\star, \boldsymbol{W}^\top\right]^\top \left[\boldsymbol{h}_{(1,1)}^\star, \boldsymbol{h}_{(1,2)}^\star, \ldots, \boldsymbol{h}_{(K,K)}^\star, \boldsymbol{W}^\top\right] = \boldsymbol{X}^\star$ ($p \ge 2K$ guarantees the existence of $\left[\boldsymbol{h}_{(1,1)}^\star, \boldsymbol{h}_{(1,2)}^\star, \ldots, \boldsymbol{h}_{(K,K)}^\star, (\boldsymbol{W}^\star)^\top\right]$). We can verify that $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ is a feasible solution for Equation (13) and have

$$\frac{1}{N}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n_k}\mathcal{L}(\boldsymbol{W}^\star\boldsymbol{h}_{k,i}^\star, \boldsymbol{y}_k^\lambda) = \sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\mathcal{L}(\boldsymbol{z}(k)^\star, \boldsymbol{y}_k^\lambda) = L_0, \tag{60}$$

where $\boldsymbol{z}(k)^\star = \left[\boldsymbol{X}^\star(k, K^2+1), \boldsymbol{X}^\star(k, K^2+2), \ldots, \boldsymbol{X}^\star(k, 2K^2)\right]^\top$ for $k \in \mathbb{K}^2$.

Combining Equation (59) and Equation (60), we conclude that L_0 is the global minimum of Equation (13) and $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ is a minimizer.

Suppose there is a minimizer $(\boldsymbol{W}', \boldsymbol{H}')$ that cannot be written as Equation (58). Let

$$\boldsymbol{h}'_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{h}'_{k,i},\ k \in \mathbb{K}^2,$$

and

$$\boldsymbol{X}' = \left[\boldsymbol{h}'_{(1,1)}, \boldsymbol{h}'_{(1,2)}, \ldots, \boldsymbol{h}'_{(K,K)}, (\boldsymbol{W}')^\top\right]^\top \left[\boldsymbol{h}'_{(1,1)}, \boldsymbol{h}'_{(1,2)}, \ldots, \boldsymbol{h}'_{(K,K)}, (\boldsymbol{W}')^\top\right].$$

Equation (59) implies that $\boldsymbol{X}'$ is a minimizer of Equation (58). As $(\boldsymbol{W}', \boldsymbol{H}')$ cannot be written as Equation (58) with $\boldsymbol{X}^\star = \boldsymbol{X}'$, then there is a $k' \in \mathbb{K}^2$, $i, j \in [n_{k'}]$ with $i \neq j$ such that $\boldsymbol{h}_{k',i} \neq \boldsymbol{h}_{k',j}$. We have

$$\begin{aligned}
&\frac{1}{K^2}\sum_{k=1}^{K^2} \boldsymbol{X}'(k,k) = \frac{1}{K^2}\sum_{k\in\mathbb{K}^2} \|\boldsymbol{h}'_k\|^2 \\
=&\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}'_{k,i}\right\|^2 - \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\|\boldsymbol{h}'_{k,i} - \boldsymbol{h}'_k\|^2 \\
\leq&\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}'_{k,i}\right\|^2 - \frac{1}{K^2}\frac{1}{n_{k'}}(\|\boldsymbol{h}'_{k',i} - \boldsymbol{h}'_{k'}\|^2 + \|\boldsymbol{h}'_{k',j} - \boldsymbol{h}'_{k'}\|^2) \\
\leq&\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}'_{k,i}\right\|^2 - \frac{1}{K^2}\frac{1}{2n_{k'}}\|\boldsymbol{h}'_{k',i} - \boldsymbol{h}'_{k',j}\|^2 \\
<&E_H.
\end{aligned}$$

By contraposition, if all $\boldsymbol{X}^\star$ satisfy that $\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}^\star(k,k) = E_H$, then all the solutions of Equation (13) are in the form of Equation (58). We complete the proof. □

Restated Theorem 5.1. Assume $p \geq 2K$ and the loss function $\mathcal{L}$ is convex in the first argument. Let $\boldsymbol{X}^\star$ be any minimizer of the convex program with $n_{(1,1)} = n_{(1,2)} = \cdots = n_{(K_A,K_A)} = n_A$ and $n_{(K_A+1,K_A+1)} = n_{(K_A+1,K_A+2)} = \cdots = n_{(K,K)} = w_r n_B$. Define $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ as

$$\begin{aligned}
\left[\boldsymbol{h}^\star_{(1,1)}, \boldsymbol{h}^\star_{(1,2)}, \ldots, \boldsymbol{h}^\star_{(K,K)},\ (\boldsymbol{W}^\star)^\top\right] &= \boldsymbol{P}(\boldsymbol{X}^\star)^{1/2} \\
\boldsymbol{h}^\star_{k,i} = \boldsymbol{h}^\star_k,&\ \text{ for all } i \in \mathcal{I}^\lambda_k, k \in \mathbb{K}^2_A, \\
\boldsymbol{h}^\star_{k,i} = \boldsymbol{h}^\star_k,&\ \text{ for all } i \in \mathcal{I}^\lambda_k, k \in \mathbb{K}^2_B,
\end{aligned}$$

where $\boldsymbol{P} \in \mathbb{R}^{p\times 2K^2}$ is any partial orthogonal matrix such that $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_{2K^2}$. Then $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ is a global minimizer of the mixed-label balanced UFM$_\lambda$-MS.

Proof of Theorem 5.1. For any feasible solution $\left(\boldsymbol{W}^\lambda, \boldsymbol{H}^\lambda\right)$ for the original program Equation (13), we define

$$\boldsymbol{h}^\lambda_{k_A} := \frac{1}{n_A}\sum_{i=1}^{n_A}\boldsymbol{h}_{k_A,i},\ k_A \in \mathbb{K}^2_A, \text{ and } \boldsymbol{h}^\lambda_{k_B} := \frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\boldsymbol{h}_{k_B,i},\ k_B \in \mathbb{K}^2_B,$$

and

$$\boldsymbol{X} := \left[\boldsymbol{h}^\lambda_{(1,1)}, \boldsymbol{h}^\lambda_{(1,2)}, \ldots, \boldsymbol{h}^\lambda_{(K,K)}, \boldsymbol{W}^{\lambda\top}\right]^\top \left[\boldsymbol{h}^\lambda_{(1,1)}, \boldsymbol{h}^\lambda_{(1,2)}, \ldots, \boldsymbol{h}^\lambda_{(K,K)}, \boldsymbol{W}^{\lambda\top}\right].$$

Clearly, $\boldsymbol{X} \succeq 0$. For the other two constraints of Equation (58), we have

$$\begin{aligned}\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}(k,k) &= \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\|\boldsymbol{h}_k^\lambda\|^2 \\ &\overset{a}{\leq} \frac{1}{K^2}\left(\sum_{k_A\in\mathbb{K}_A^2}\frac{1}{n_A}\sum_{i=1}^{n_A}\left\|\boldsymbol{h}_{k_A,i}^\lambda\right\|^2 + \sum_{k_B\in\mathbb{K}_B^2}\frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\left\|\boldsymbol{h}_{k_B,i}^\lambda\right\|^2\right) \\ &\overset{b}{\leq} E_H\end{aligned}$$

and

$$\frac{1}{K^2}\sum_{k=K^2+1}^{2K^2}\boldsymbol{X}(k,k) = \frac{1}{K^2}\sum_{a=1}^{K}\sum_{b=1}^{K}\|\boldsymbol{w}_{(a,b)}^\lambda\|^2 \overset{c}{\leq} E_W,$$

where $\overset{a}{\leq}$ applies Jensen's inequality and $\overset{b}{\leq}$ and $\overset{c}{\leq}$ use that $\left(\boldsymbol{W}^\lambda, \boldsymbol{H}^\lambda\right)$ is a feasible solution. So $\boldsymbol{X}$ is a feasible solution for the convex program Equation (58). Letting L_0 be the global minimum of Equation (58), for any feasible solution $\left(\boldsymbol{W}^\lambda, \boldsymbol{H}^\lambda\right)$, we obtain

$$\begin{aligned}&\frac{1}{N}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n_k}\mathcal{L}(\boldsymbol{W}^\lambda\boldsymbol{h}_{k,i}^\lambda, \boldsymbol{y}_k^\lambda) \\ &= \sum_{k_A\in\mathbb{K}_A^2}\frac{n_A}{N}\left[\frac{1}{n_A}\sum_{i=1}^{n_A}\mathcal{L}(\boldsymbol{W}^\lambda\boldsymbol{h}_{k_A,i}^\lambda, \boldsymbol{y}_{k_A}^\lambda)\right] + \sum_{k_B\in\mathbb{K}_B^2}\frac{w_r n_B}{N}\left[\frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\mathcal{L}(\boldsymbol{W}^\lambda\boldsymbol{h}_{k_B,i}^\lambda, \boldsymbol{y}_{k_B}^\lambda)\right] \\ &\overset{a}{\geq} \sum_{k_A\in\mathbb{K}_A^2}\frac{n_A}{N}\mathcal{L}(\boldsymbol{W}^\lambda\boldsymbol{h}_{k_A}^\lambda, \boldsymbol{y}_{k_A}^\lambda) + \sum_{k_B\in\mathbb{K}_B^2}\frac{w_r n_B}{N}\mathcal{L}(\boldsymbol{W}^\lambda\boldsymbol{h}_{k_B}^\lambda, \boldsymbol{y}_{k_B}^\lambda) \\ &= \sum_{k_A\in\mathbb{K}_A^2}\frac{n_A}{N}\mathcal{L}(\boldsymbol{z}(k_A)^\lambda, \boldsymbol{y}_{k_A}^\lambda) + \sum_{k_B\in\mathbb{K}_B^2}\frac{w_r n_B}{N}\mathcal{L}(\boldsymbol{z}(k_B)^\lambda, \boldsymbol{y}_{k_B}^\lambda) \geq L_0,\end{aligned} \tag{61}$$

where in $\overset{a}{\geq}$, we use $\mathcal{L}$ is convex on the first argument, and so $\mathcal{L}(\boldsymbol{W}^\lambda\boldsymbol{h}^\lambda, \boldsymbol{y}_k^\lambda)$ is convex on $\boldsymbol{h}$ given $\boldsymbol{W}$ and $k \in \mathbb{K}^2$.

For the simplicity of our expressions, we hereafter remove the superscript λ of $\boldsymbol{W}^\lambda$, $\boldsymbol{H}^\lambda$, $\boldsymbol{h}^\lambda$ and $\boldsymbol{z}^\lambda$.

On the other hand, considering the solution $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ defined in Equation (58) with $\boldsymbol{X}^\star$ being a minimizer of Equation (58), we have $\left[\boldsymbol{h}_{(1,1)}^\star, \boldsymbol{h}_{(1,2)}^\star, \ldots, \boldsymbol{h}_{(K,K)}^\star, \boldsymbol{W}^\top\right]^\top \left[\boldsymbol{h}_{(1,1)}^\star, \boldsymbol{h}_{(1,2)}^\star, \ldots, \boldsymbol{h}_{(K,K)}^\star, \boldsymbol{W}^\top\right] = \boldsymbol{X}^\star$ ($p \geq 2K$ guarantees the existence of $\left[\boldsymbol{h}_{(1,1)}^\star, \boldsymbol{h}_{(1,2)}^\star, \ldots, \boldsymbol{h}_{(K,K)}^\star, (\boldsymbol{W}^\star)^\top\right]$). We can verify that $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ is a feasible solution for Equation (13) and have

$$\frac{1}{N}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n_k}\mathcal{L}(\boldsymbol{W}^\star\boldsymbol{h}_{k,i}^\star, \boldsymbol{y}_k^\lambda) = \sum_{k_A\in\mathbb{K}_A^2}\frac{n_A}{N}\mathcal{L}(\boldsymbol{z}(k_A)^\star, \boldsymbol{y}_{k_A}^\lambda) + \sum_{k_B\in\mathbb{K}_B^2}\frac{w_r n_B}{N}\mathcal{L}(\boldsymbol{z}(k_B)^\star, \boldsymbol{y}_{k_B}^\lambda) = L_0, \tag{62}$$

where $\boldsymbol{z}(k_A)^\star = \left[\boldsymbol{X}^\star(k_A, K^2+1), \boldsymbol{X}^\star(k_A, K^2+2), \ldots, \boldsymbol{X}^\star(k_A, K^2+K_A)\right]^\top$ for $k_A \in \mathbb{K}_A^2$ and $\boldsymbol{z}(k_B)^\star = \left[\boldsymbol{X}^\star(k_B, K^2+K_A+1), \boldsymbol{X}^\star(k_B, K^2+K_A+2), \ldots, \boldsymbol{X}^\star(k_B, 2K^2)\right]^\top$ for $k_B \in \mathbb{K}_B^2$.

Combining Equation (61) and Equation (62), we conclude that L_0 is the global minimum of Equation (13) and $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ is a minimizer.

Suppose there is a minimizer $(\boldsymbol{W}', \boldsymbol{H}')$ that cannot be written as Equation (58). Let

$$\boldsymbol{h}'_{k_A} = \frac{1}{n_A}\sum_{i=1}^{n_A}\boldsymbol{h}'_{k_A,i},\ k_A \in \mathbb{K}_A^2, \text{ and } \boldsymbol{h}'_{k_B} = \frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\boldsymbol{h}'_{k_B,i}, k_B \in \mathbb{K}_B^2$$

and

$$\boldsymbol{X}' = \left[\boldsymbol{h}'_{(1,1)}, \boldsymbol{h}'_{(1,2)}, \ldots, \boldsymbol{h}'_{(K,K)}, (\boldsymbol{W}')^\top\right]^\top \left[\boldsymbol{h}'_{(1,1)}, \boldsymbol{h}'_{(1,2)}, \ldots, \boldsymbol{h}'_{(K,K)}, (\boldsymbol{W}')^\top\right].$$

Equation (61) implies that $\boldsymbol{X}'$ is a minimizer of Equation (58). As $(\boldsymbol{W}', \boldsymbol{H}')$ cannot be written as Equation (58) with $\boldsymbol{X}^\star = \boldsymbol{X}'$, then there is a $k' \in \mathbb{K}^2$, $i, j \in [n'_k]$ with $i \neq j$ such that $\boldsymbol{h}_{k',i} \neq \boldsymbol{h}_{k',j}$. We have

$$\begin{aligned}
&\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}'(k,k) = \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\|\boldsymbol{h}'_k\|^2 \\
=&\frac{1}{K^2}\sum_{k_A\in\mathbb{K}_A^2}\frac{1}{n_A}\sum_{i=1}^{n_A}\left\|\boldsymbol{h}'_{k_A,i}\right\|^2 - \frac{1}{K^2}\sum_{k_A\in\mathbb{K}_A^2}\frac{1}{n_A}\sum_{i=1}^{n_A}\|\boldsymbol{h}'_{k_A,i} - \boldsymbol{h}'_{k_A}\|^2 \\
&+\frac{1}{K^2}\sum_{k_B\in\mathbb{K}_B^2}\frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\left\|\boldsymbol{h}'_{k_B,i}\right\|^2 - \frac{1}{K^2}\sum_{k_B\in\mathbb{K}_B^2}\frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\|\boldsymbol{h}'_{k_B,i} - \boldsymbol{h}'_{k_B}\|^2 \\
\leq&\frac{1}{K^2}\sum_{k_A\in\mathbb{K}^2}\frac{1}{n_A}\sum_{i=1}^{n_A}\left\|\boldsymbol{h}'_{k_A,i}\right\|^2 - \frac{1}{K^2}\frac{1}{n_{k'_A}}(\|\boldsymbol{h}'_{k'_A,i} - \boldsymbol{h}'_{k'_A}\|^2 + \|\boldsymbol{h}'_{k'_A,j} - \boldsymbol{h}'_{k'_A}\|^2) \\
&+\frac{1}{K^2}\sum_{k_B\in\mathbb{K}_B^2}\frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\left\|\boldsymbol{h}'_{k_B,i}\right\|^2 - \frac{1}{K^2}\frac{1}{n_{k'_B}}(\|\boldsymbol{h}'_{k'_B,i} - \boldsymbol{h}'_{k'_B}\|^2 + \|\boldsymbol{h}'_{k'_B,j} - \boldsymbol{h}'_{k'_B}\|^2) \\
\leq&\frac{1}{K^2}\sum_{k_A\in\mathbb{K}_A^2}\frac{1}{n_A}\sum_{i=1}^{n_A}\left\|\boldsymbol{h}'_{k_A,i}\right\|^2 - \frac{1}{K^2}\frac{1}{2n_{k'_A}}\|\boldsymbol{h}'_{k'_A,i} - \boldsymbol{h}'_{k'_A,j}\|^2 \\
&+\frac{1}{K^2}\sum_{k_B\in\mathbb{K}_B^2}\frac{1}{w_r n_B}\sum_{i=1}^{w_r n_B}\left\|\boldsymbol{h}'_{k_B,i}\right\|^2 - \frac{1}{K^2}\frac{1}{2n_{k'_B}}\|\boldsymbol{h}'_{k'_B,i} - \boldsymbol{h}'_{k'_B,j}\|^2 \\
<&E_H.
\end{aligned}$$

By contraposition, if all $\boldsymbol{X}^\star$ satisfy that $\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}^\star(k,k) = E_H$, then all the solutions of Equation (13) are in the form of Equation (58). We complete the proof. □

H Experimental Setup

Implementation Details. Our experiments follow the setups of [11] and [48] for CIFAR10-LT, ImageNet-LT, Places-LT, and iNaturalist2018 and [21] for CIFAR100-LT. We employ ResNet32 for CIFAR10-LT, doubling the feature dimensions for CIFAR100-LT. For ImageNet-LT and iNaturalist2018, we use ResNet50, and for Places-LT, use ResNet152, respectively. To reproduce baseline comparisons, we adopt the same hyperparameter settings as in [11] and [48].

Datasets. Following [11, 48], we use the long-tailed variants of CIFAR10, CIFAR100 [83], ImageNet [97], Places365 [98], and iNaturalist2018 [99].

CIFAR10-LT. 10 imbalanced classes, subsampled at exponentially decreasing rates from CIFAR10 [83].

CIFAR100-LT. 100 imbalanced classes, constructed analogously to CIFAR10-LT.

ImageNet-LT. Derived from ImageNet for large-scale object classification. Class frequencies follow a Pareto distribution ($\alpha = 5$) with cardinalities from 5 to 1,280, totaling 115.8K images across 1,000 classes.

Places-LT. An extended version of Places, with class sizes ranging from 5 to 4,980, yielding 184.5K images from 365 classes.

iNaturalist2018. A large-scale real-world species classification dataset with extreme label imbalance, comprising 437,513 images from 8,142 categories.

Architectures. For CIFAR10-LT, we use ResNet32 [100] with three residual blocks, producing feature dimensions of 16, 32, and 64, respectively. CIFAR100-LT doubles these dimensions. Differing from the standard ResNet architecture used for ImageNet, the ResNet32's first convolutional layer has a kernel size, stride, and padding of 3, 1, and 1, respectively. ResNet50 and 152 follow [100].

Hyperparameters. For CIFAR10/100-LT, models are trained with mini-batch size 128 using SGD with momentum 0.9 and weight decay 2e-4 for 200 epochs. The learning rate is linearly warmed up from 0.02 and decayed by 0.1 at epochs 160 and 180. For ImageNet-LT and Places-LT, models are trained with SGD (momentum 0.9, weight decay 5e-4) and a cosine annealing scheduler. Mixup alpha is set per dataset: $\alpha = 1.0$ for CIFAR10/100-LT, $\alpha = 0.2$ for others.

ETF+DR [21]. In [21], it was proven that by fixing the classifier as a K-simplex ETF, NC is satisfied regardless of class balance, and that using this fixed ETF classifier along with a specialized loss (Dot-Regression; DR) improves model performance in imbalanced learning environments. Leveraging the advantages of the fixed ETF classifier, we hypothesized that our method could produce synergies with this approach, and we conducted experiments applying our method to this framework. However, a scale factor is necessary for the fixed ETF classifier, due to class vectors should be normalized. For this reason, we make a modified version of the fixed ETF classifier to apply our methods, named as *fixed Mixed-Singleton Weighted ETF classifier (MS-WETF)*.

The scale of class vectors is important for softmax cross-entropy loss. Thus, we remove the scale factor and add learnable parameter $\boldsymbol{s} \in \mathbb{R}^K$ to control the scale of each class vectors.

$$\boldsymbol{W}_{\text{WETF}} = \boldsymbol{s} \cdot \boldsymbol{W}_{\text{ETF}}$$

Then, we make $\boldsymbol{W}_{\text{WETF}}$ as Mixed-Singleton classifier

$$\boldsymbol{W}^{\lambda}_{\text{MS-WETF, (a,b)}} = [\lambda \boldsymbol{w}_{\text{WETF},a} + (1-\lambda)\boldsymbol{w}_{\text{WETF},b}]_{(a,b)\in\mathbb{K}^2}$$

Remix [38]. In [38], they pointed out that using the same mixing factor λ for mixed samples in both last-layer features and their respective labels does not make sense under the imbalanced learning environments. As a result, Remix has been proposed to disentangle λ as below:

$$\begin{aligned}\tilde{\boldsymbol{x}}^{\text{RM}} =& \lambda_x \boldsymbol{x}_i + (1-\lambda_x)\boldsymbol{x}_{\pi(i)},\\ \tilde{\boldsymbol{y}}^{\text{RM}} =& \lambda_y \boldsymbol{y}_{c_i} + (1-\lambda_y)\boldsymbol{y}_{c_{\pi(i)}}, \forall (i,\pi(i)) \in \mathcal{I}^{\lambda},\end{aligned}$$

where λ_x is sampled from the beta distribution and λ_y is defined as below:

$$\lambda_y = \begin{cases} 0 & n_i/n_{\pi(i)} \geq \kappa \text{ and } \lambda < \tau; \\ 1 & n_i/n_{\pi(i)} \leq 1/\kappa \text{ and } 1-\lambda < \tau; \\ \lambda & \text{otherwise} \end{cases}$$

Here n_i and $n_{\pi(i)}$ denote the number of samples in the corresponding class from $\boldsymbol{x}_i$ and $\boldsymbol{x}_{\pi(i)}$. κ and τ are two hyperparameters in Remix, and we used the same values as those employed in the original Remix implementation: $\kappa = 3$ and $\tau = 0.5$.

Unlike Remix, which controls the mixing factor λ, our method controls only the sample and label pairs. Owing to this independence from Remix, integrating our method with Remix simply requires replacing the original index pair set $\mathcal{I}^{\lambda}$ with the balanced mixed-label pair set $\tilde{\mathcal{I}}^{\lambda}$ obtained through BMLS. Also, when initializing the MS classifier, we used λ_y from the same way to Remix.

DBN-mix [39]. DBN-mix is a method that expands bilateral mixup (which is from BBN-mix [48]) to double branches while one input sample $\boldsymbol{x}_i$ comes from random sampler and the other $\boldsymbol{x}_j$ comes from class-balanced sampler. Therefore, there are two mixed samples generated in each mini-batch as below:

$$\begin{aligned}\tilde{\boldsymbol{x}}^{\text{cb}} =& \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_j,\\ \tilde{\boldsymbol{x}}^{\text{rb}} =& (1-\lambda)\boldsymbol{x}_i + \lambda \boldsymbol{x}_j,\\ \tilde{\boldsymbol{y}}^{\text{cb}} =& \lambda \boldsymbol{y}_i + (1-\lambda)\boldsymbol{y}_j,\\ \tilde{\boldsymbol{y}}^{\text{rb}} =& (1-\lambda)\boldsymbol{y}_i + \lambda \boldsymbol{y}_j,\end{aligned}$$

where the mixed samples $\tilde{\boldsymbol{x}}^{\text{cb}}$ and $\tilde{\boldsymbol{x}}^{\text{rb}}$ are trained by their respective different classifiers.

In this setting, the loss from each branch $\mathcal{L}$ is computed separately as shown below, and the final loss $\mathcal{L}_{\text{total}}$ is obtained by taking their weighted sum via a hyperparameter γ.

$$\mathcal{L}_{\text{total}} = \gamma \cdot \mathcal{L}\left(\tilde{\boldsymbol{p}}^{\text{cb}}, \tilde{\boldsymbol{y}}^{\text{cb}}\right) + (1-\gamma) \cdot \mathcal{L}\left(\tilde{\boldsymbol{p}}^{\text{rb}}, \tilde{\boldsymbol{y}}^{\text{rb}}\right),$$

where $\tilde{\boldsymbol{p}}$ is the logit of the mixed sample $\tilde{\boldsymbol{x}}$ and $\mathcal{L}$ denotes the cross-entropy loss. This loss is then used to train the classifiers of each branch and the shared backbone in an end-to-end manner.

In DBN-mix, two different samples, each drawn from a different sampler, are mixed together, which causes the mixed-label balance to break in both branches even if the class-balanced sampler is replaced with BMLS. To address this, we employ two samplers in parallel and configure each branch as follows:

$$\begin{aligned}
\tilde{\boldsymbol{x}}^{\text{cb}} &= \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_{\pi(i)}, \\
\tilde{\boldsymbol{x}}^{\text{rb}} &= \lambda \boldsymbol{x}_j + (1-\lambda)\boldsymbol{x}_{\pi(j)}, \\
\tilde{\boldsymbol{y}}^{\text{cb}} &= \lambda \boldsymbol{y}_i + (1-\lambda)\boldsymbol{y}_{\pi(i)}, \\
\tilde{\boldsymbol{y}}^{\text{rb}} &= \lambda \boldsymbol{y}_j + (1-\lambda)\boldsymbol{y}_{\pi(j)},
\end{aligned}$$

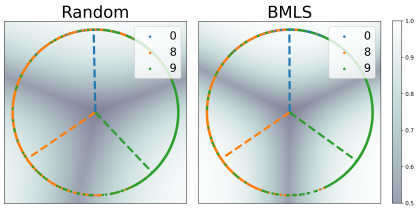
for all $(i, \pi(i)) \in \mathcal{I}^{\lambda}$ and $(j, \pi(j)) \in \tilde{\mathcal{I}}^{\lambda}$, which denote a random sampler and BMLS, respectively. In our experiments, we empirically identified appropriate values for the hyperparameter γ. As a result, we set $\gamma = 0.9$ for CIFAR10-LT and $\gamma = 0.5$ for CIFAR100-LT.

I Additional Experimental Results

I.1 Validation of BMLS

Table 9: Experimental results of mixed-label frequency control experiments on Many/Medium/Few Categories in the CIFAR10/100-LT dataset.

$\mathcal{D}$	Ctgy.	Coverage ratio					
		0.0	0.2	0.4	0.6	0.8	1.0
C10	few	58.74	58.56	58.50	58.51	60.06	65.60
	med	68.46	71.09	74.63	77.85	77.10	75.21
	many	86.93	92.13	91.79	91.49	91.91	89.59
	all	71.08	73.64	74.94	76.14	76.43	76.64
C100	few	7.72	10.28	12.17	12.66	13.16	13.20
	med	34.49	39.13	42.95	44.47	46.21	46.97
	many	61.26	65.64	66.49	67.85	68.93	67.60
	all	36.37	40.31	42.50	43.66	44.81	44.60



Sampler	Test Acc. (%) ↑	U_G ↑	U ↑
Random	72.91	14.6404	4.2325
CBS	75.86	15.2521	4.4848
CAS	76.60	15.3319	4.4999
BMLS	78.71	15.3379	4.5651

Figure 4: Experiments on CIFAR10-LT dataset for the effectiveness of BMLS to minority collapse. (figure) Visualization of 2D-projection of class vectors about Many class $\{0\}$ and Few classes $\{8, 9\}$. Dashed line indicates each class vector and contrast of background means the confidence value, i.e., a confidence close to 0.5 indicates that the model is confused between the two classes for the given sample, and this is represented by darker colors in the figure. (table) Quantitative comparison results. (U_G: Uniformity of all classes, U: Uniformity of $\{0, 8, 9\}$ classes)

Uniformity [101] prefers a feature distribution that preserves maximal information, i.e., the uniform distribution on the unit hypersphere. To assess the uniformity on different centroids of classes, we use the following metric $\mathbf{U}$:

$$\mathbf{U} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1, j\neq i}^{K} \|\bar{\boldsymbol{h}}_i - \bar{\boldsymbol{h}}_j\|_2 \tag{63}$$

where $\bar{\boldsymbol{h}}_k$ is the center of samples from class k on the hypersphere: $\bar{\boldsymbol{h}}_k = \frac{\sum_{\boldsymbol{h}_{k,i} \in \mathbb{S}_k} \boldsymbol{h}_{k,i}}{\|\sum_{\boldsymbol{h}_{k,i} \in \mathbb{S}_k} \boldsymbol{h}_{k,i}\|_2}$.

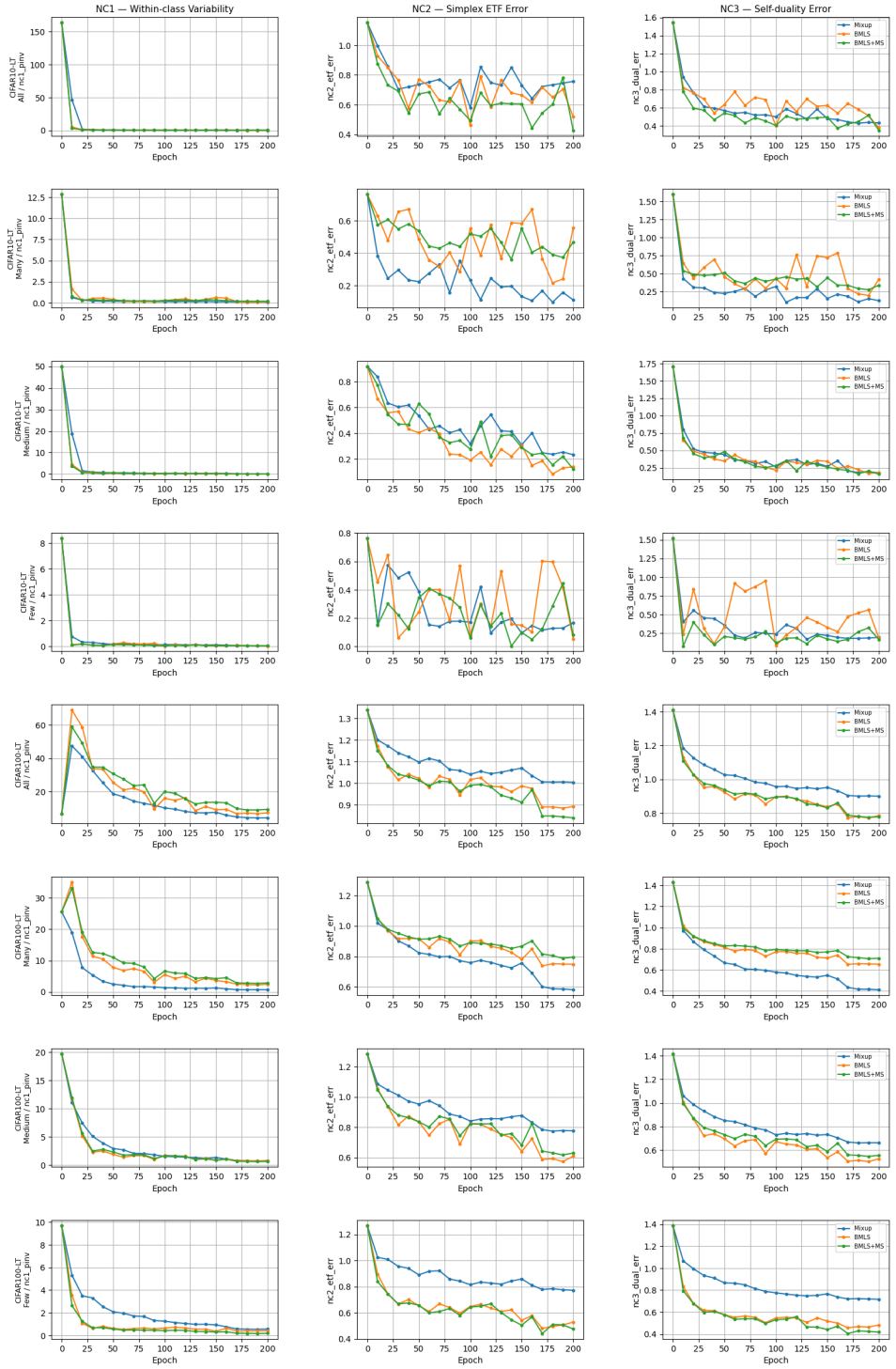


Figure 5: NC metric measurement on CIFAR10-LT (top four rows) and CIFAR100-LT (bottom four rows) with imbalance factor 100. ($\mathcal{NC}_1$: Within-class variability, $\mathcal{NC}_2$: Simplex ETF error, $\mathcal{NC}_3$: Self-duality error)

NC metric measurement. To furnish more direct empirical evidence, we incorporate traditional NC metrics ($\mathcal{NC}_1$:Within-class Variability, $\mathcal{NC}_2$: Simplex ETF Error, $\mathcal{NC}_3$: Self-duality Error) [20, 102] to quantitatively measure and compare explicit collapse behavior for Mixup vs. BMLS vs. BMLS+MS on CIFAR10/100-LT (imbalance factor: 100), as illustrated in Figure 5.

Table 10: NC metric measurement on CIFAR10/100-LT with imbalance factor 100. ($\mathcal{NC}_1$: Within-class variability, $\mathcal{NC}_2$: Simplex ETF error, $\mathcal{NC}_3$: Self-duality error)

Method	*Many*			*Med*			*Few*			*All*		
	$\mathcal{NC}_1$	$\mathcal{NC}_2$	$\mathcal{NC}_3$	$\mathcal{NC}_1$	$\mathcal{NC}_2$	$\mathcal{NC}_3$	$\mathcal{NC}_1$	$\mathcal{NC}_2$	$\mathcal{NC}_3$	$\mathcal{NC}_1$	$\mathcal{NC}_2$	$\mathcal{NC}_3$
CIFAR10-LT												
Mixup	**0.0826**	**0.1132**	**0.1235**	0.1072	0.2352	0.1793	0.0330	0.1668	0.1942	**0.1723**	0.7573	0.4301
BMLS	0.1474	0.5549	0.4192	**0.0640**	0.1408	0.1799	**0.0199**	**0.0541**	0.1877	0.2545	0.5200	0.3772
BMLS+MS	0.1798	0.4678	0.3378	0.1364	**0.1248**	**0.1598**	0.0350	0.0816	**0.1672**	0.3122	**0.4288**	**0.3487**
CIFAR100-LT												
Mixup	**0.6768**	**0.5816**	**0.4123**	0.7298	0.7768	0.6627	0.5633	0.7720	0.7165	**4.3097**	1.0042	0.9001
BMLS	2.4062	0.7486	0.6539	0.7870	**0.6094**	**0.5268**	0.4226	0.5283	0.4820	7.3132	0.8914	0.7843
BMLS+MS	2.8291	0.7950	0.7100	**0.6107**	0.6304	0.5565	**0.1823**	**0.4758**	**0.4169**	9.3861	**0.8390**	**0.7780**

$\mathcal{NC}_1$ is a within-class variability collapse, where the value closer to 0 indicates that the last-layer features become closer to its corresponding class mean:

$$\mathcal{NC}_1 := \frac{1}{K}\text{tr}\{\Sigma_W \Sigma_B^{\dagger}\}$$

$\mathcal{NC}_2$ is a simplex ETF error, where the value closer to 0 indicates that the class means become closer to a simplex ETF:

$$\mathcal{NC}_2 := \left\| \frac{\boldsymbol{M}\boldsymbol{M}^\top}{\|\boldsymbol{M}\boldsymbol{M}^\top\|_F} - \frac{1}{\sqrt{K-1}}\left(\boldsymbol{I}_k - \frac{1}{K}\mathbf{1}_k\mathbf{1}_k^\top\right)\right\|_F,$$

where $\boldsymbol{M} = [\frac{\mu_k - \mu_G}{\|\mu_k - \mu_G\|_2}]_{k=1}^K \in \mathbb{R}^{p\times K}$ and $\|\cdot\|_F$ is a Frobenius norm.

$\mathcal{NC}_3$ is a self-duality error, where the value closer to 0 indicates that the class mean and its corresponding class vector are aligned at the simplex ETF:

$$\mathcal{NC}_3 := \left\| \frac{\boldsymbol{A}\boldsymbol{M}^\top}{\|\boldsymbol{A}\boldsymbol{M}^\top\|_F} - \frac{1}{\sqrt{K-1}}\left(\boldsymbol{I}_k - \frac{1}{K}\mathbf{1}_k\mathbf{1}_k^\top\right)\right\|_F,$$

where $\boldsymbol{A}$ is the last-layer classifier.

As evidenced by the results in Figure 5 and Table 10, $\mathcal{NC}_2$ and $\mathcal{NC}_3$, excluding $\mathcal{NC}_1$, exhibit lower values under the proposed method, suggesting that the proposed method more effectively preserves the simplex ETF structure. The higher $\mathcal{NC}_1$ value suggests that our proposed method produces a stronger augmentation effect. This aligns with the observation that higher $\mathcal{NC}_1$ values are associated with greater representational diversity and improved expressiveness in OOD settings [103].

I.2 Collapse Geometry under Mixed-label Coverage Control

We measure the collapse geometry at each coverage level of the mixed-label frequency control experiment (Figure 3a), with five seeds per level. The *Few–few cosine* is the mean cosine similarity over all pairs of class vectors of *Few* classes; under a simplex ETF it equals $-1/(K-1)$, and positive values indicate minority collapse. Table 11 reports CIFAR10-LT (CIFAR100-LT is in Table 1), and Table 12 reports the correlations between coverage and each metric. On CIFAR10-LT the metrics improve up to coverage 0.6 and then plateau: with only 45 mixed labels under the $a < b$ restriction, a rotating subset covers the pair space over the course of training, whereas CIFAR100-LT has 4,950 mixed labels and the per-epoch budget never saturates.

Class-vector geometry across samplers. Table 13 compares the Few–few cosine of the four samplers. On CIFAR10-LT the separation is clear: the random sampler leaves the minority class vectors positively correlated, while BMLS_{MS} nearly reaches the ideal value. On CIFAR100-LT, we report the three balanced samplers as a tie: the gaps of CAS and BMLS_{MS} differ by 0.0004, well inside the per-seed standard deviations. This metric loses resolution as K grows, since the ideal value $-1/(K-1)$ approaches zero while independent high-dimensional vectors are also near-orthogonal in the mean; $\mathcal{NC}_2$ and $\mathcal{NC}_3$, which characterize the full frame, do separate the methods at $K = 100$ (Table 10).

Table 11: Collapse geometry under mixed-label coverage control on CIFAR10-LT (imbalance factor 100; mean_{std} over five seeds; ideal Few–few cosine $-1/9 \approx -0.1111$).

Coverage	$\mathcal{NC}_2$ few ↓	$\mathcal{NC}_3$ few ↓	Few–few cosine	F1 ↑	Acc. ↑
0.2	$0.4682_{0.0269}$	$0.3535_{0.0118}$	$0.0376_{0.0105}$	$73.68_{0.64}$	$73.69_{0.60}$
0.4	$0.2100_{0.0867}$	$0.1355_{0.0542}$	$-0.0202_{0.0151}$	$74.82_{0.45}$	$74.93_{0.39}$
0.6	$0.1497_{0.0451}$	$\mathbf{0.0975}_{0.0825}$	$\mathbf{-0.1045}_{0.0147}$	$\mathbf{76.71}_{0.73}$	$\mathbf{76.76}_{0.69}$
0.8	$0.1577_{0.0290}$	$0.1037_{0.0320}$	$-0.0771_{0.0119}$	$76.14_{0.47}$	$76.31_{0.40}$
1.0	$\mathbf{0.1235}_{0.0684}$	$0.1009_{0.0218}$	$-0.0786_{0.0122}$	$76.01_{0.47}$	$76.14_{0.42}$

Table 12: Correlation between mixed-label coverage and collapse metrics across five coverage levels and five seeds per dataset ($n = 25$).

Dataset	Metric	Pearson r	p	Spearman ρ	p
CIFAR10-LT	$\mathcal{NC}_2$ few	-0.758	1.12e−5	-0.686	1.51e−4
	$\mathcal{NC}_3$ few	-0.693	1.22e−4	-0.518	8.03e−3
	Few–few cosine	-0.776	5.26e−6	-0.675	2.17e−4
	F1	$+0.690$	1.33e−4	$+0.686$	1.51e−4
	Test accuracy	$+0.722$	4.63e−5	$+0.692$	1.25e−4
CIFAR100-LT	$\mathcal{NC}_2$ few	-0.920	7.76e−11	-0.941	2.44e−12
	$\mathcal{NC}_3$ few	-0.950	3.91e−13	-0.981	8.95e−18
	Few–few cosine	-0.935	7.98e−12	-0.934	9.96e−12
	F1	$+0.886$	4.00e−9	$+0.883$	5.37e−9
	Test accuracy	$+0.895$	1.53e−9	$+0.908$	3.59e−10

Table 13: Few–few class-vector cosine across samplers (imbalance factor 100; mean_{std} over five seeds). Gap: distance to the ideal simplex ETF value $-1/(K-1)$.

	CIFAR10-LT (ideal -0.1111)		CIFAR100-LT (ideal -0.0101)	
Method	Few–few cosine	Gap ↓	Few–few cosine	Gap ↓
Mixup + random	$0.0633_{0.0084}$	$0.1744_{0.0084}$	$0.0321_{0.0025}$	$0.0422_{0.0025}$
CBS	$-0.0834_{0.0055}$	$0.0277_{0.0055}$	$-0.0036_{0.0007}$	$0.0065_{0.0007}$
CAS	$-0.0788_{0.0185}$	$0.0323_{0.0185}$	$-0.0055_{0.0011}$	$\mathbf{0.0046}_{0.0011}$
BMLS_{MS}	$-0.1075_{0.0086}$	$\mathbf{0.0037}_{0.0086}$	$-0.0051_{0.0024}$	$0.0050_{0.0024}$

I.3 F1 Score and Balanced Accuracy

On the standard long-tailed benchmarks the *training* set is long-tailed but the *test* set is class-balanced, so top-1 accuracy coincides by construction with balanced accuracy (BAcc), i.e., mean per-class recall. Because recall-based metrics do not capture precision, a method that inflates minority recall by over-predicting tail classes would look artificially good under them. We therefore also report macro F1 in Table 14. The ordering of methods is preserved under F1.

Table 14: F1 / BAcc / top-1 accuracy on CIFAR10/100-LT (mean over five seeds).

$\mathcal{D}$	Method \ imb	200	100	50	10
C10	Mixup	65.47/66.77/66.77	72.52/72.94/72.94	78.66/78.64/78.64	88.13/88.05/88.05
	+CAS	69.48/69.90/69.90	76.34/76.43/76.43	81.54/81.42/81.42	**89.29/89.24/89.24**
	$+\text{BMLS}_{\text{MS}}$	**73.85/74.15/74.15**	**79.75/79.75/79.75**	**83.52/83.46/83.46**	88.46/88.48/88.48
C100	Mixup	33.14/39.06/39.06	37.88/42.88/42.88	44.50/48.31/48.31	61.95/63.03/63.03
	+CAS	35.11/40.28/40.28	40.24/44.65/44.65	46.95/50.07/50.07	62.67/63.57/63.57
	$+\text{BMLS}_{\text{MS}}$	**37.27/41.10/41.10**	**45.17/47.85/47.85**	**50.52/52.48/52.48**	**63.87/64.31/64.31**

I.4 Integration with a Superclass-aware Method

BMLS uniformly balances all mixed labels, including pairs that are semantically unrelated. We do not regard such pairs as a defect: Mixup performs vicinal risk minimization by interpolating randomly selected pairs, so most pairs it forms are semantically unrelated by construction, and interpolations between distant classes populate regions of input space that no training sample occupies. Structure-

aware long-tailed methods typically use class relationships to reshape the objective or the decision boundary rather than the sampling distribution, so they operate on a different axis from BMLS. As a preliminary test, we added BMLS_{MS} to the superclass part of SCAD [40], a superclass-aware logit adjustment method, because its superclasses are also imbalanced. We use its semi-supervised long-tailed setting on CIFAR100-LT (imbalance factor 100, $N = 150$, $M = 300$), since the representative implementations of this family are semi-supervised. The integration improves Top-1 accuracy from 41.57 to 42.23 ($+0.66$) and the balanced-head accuracy from 44.15 to 44.97 ($+0.82$), and neither method required modification to be applied jointly.

I.5 Integration with DBN-mix on iNaturalist2018

We compared DBN-mix, DBN-mix+BMLS, and DBN-mix+BMLS_{MS} on iNaturalist2018 with a two-branch ResNet-50, a global batch size of 768, and 15 training epochs (reduced from the original 200 because of the computational budget). All methods used SGD with momentum 0.9, an initial learning rate of 0.1, weight decay 10^{-4}, one epoch of learning-rate warm-up, and learning-rate reductions by a factor of 10 at epochs 9 and 12. Mixup used $\alpha = 0.2$, and the final prediction was obtained by equally averaging the logits of the conventional (CB) and re-balanced (RB) branches. DBN-mix used random instance sampling for the CB branch and reverse soft-balanced sampling for the RB branch, with ordinary linear classifiers for both branches. BMLS retained random instance sampling for CB but used a circular BMLS for RB, in which a pair is not repeated until the complete class-pair cycle is traversed; its RB classifier was optimized with $\lambda\mathcal{L}_{CE}(\boldsymbol{z}, y_a) + (1-\lambda)\mathcal{L}_{CE}(\boldsymbol{z}, y_b)$. BMLS_{MS} used the same sampling but replaced the RB objective with MS, constructing $\boldsymbol{w}_{(a,b)} = \lambda\boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_b$ for each observed pair. Following DBN-mix, the mixed and singleton class vectors and the RB features were ℓ_2-normalized, and cosine logits were scaled by 30. To introduce BMLS and MS gradually, their RB loss was disabled for two CB-only epochs, ramped to weights of 0.167 and 0.333 at epochs 2 and 3, and combined equally with CB from epoch 4 onward.

The results are in Table 4. The DBN-mix baseline is strongly inverted (*many* 21.42 vs. *few* 42.78), the signature of over-correction by reverse-balanced sampling, and our components pull the classifier back toward a balanced configuration. All three methods share the same 15-epoch budget, backbone, and optimization settings, so the comparison is controlled even though none of them has converged. We do not claim improved tail recognition from this table, which is the profile we observe on CIFAR but do not reproduce here. This result also refines the applicability criterion of MS: the coverage condition (per-epoch sample budget exceeding the number of mixed labels) applies to the standalone configuration, while in integration MS remained beneficial below it (Remark 5.2).

I.6 Category-wise Accuracy

According to [5], we also calculate top-1 test accuracy of three disjoint set: many, medium, and few categories. The classes included in each set for the respective datasets are described in Table 15. In the tables of experimental results about many, medium, and few categories, we report the mean and std of top-1 test accuracies as $mean_{std}$ over random seeds. For brevity, we report solely the $mean$ of top-1 test accuracy for the *All* category in the main paper. Full results including other categories are provided in the following supplementary materials.

Table 15: The classes in Many/Medium/Few class sets.

	CIFAR10-LT	CIFAR100-LT	Places-LT	ImageNet-LT	iNaturalist2018
Many	[0,2]	[0,35]	[0, 130]	[0, 389]	[0, 841]
Medium	[3,6]	[36,70]	[131, 287]	[390, 834]	[842, 4542]
Few	[7,9]	[71,99]	[288, 364]	[835, 999]	[4543, 8141]
All	[0,9]	[0,99]	[0, 364]	[0, 999]	[0, 8141]

Table 16: Comparison experiments of samplers on the CIFAR10/100-LT dataset with various imbalance factors. The results are the mean of five repeated experiments with random seeds. Best in bold. (LOM: Label-Occurrence Mixup, CAS: Class-Aware Sampler, BMLS: Balanced Mixed Label Sampler)

Method	CIFAR10-LT				CIFAR100-LT			
	imbalance factor				imbalance factor			
	200	100	50	10	200	100	50	10
mixup								
Mixup [8]	67.30	72.80	78.60	87.70	38.70	43.00	48.10	58.20
Remix [38]	N/A	73.00	N/A	88.50	N/A	41.40	N/A	59.50
Remix+RS [38]	N/A	76.23	N/A	87.70	N/A	41.13	N/A	58.62
CMO [46]	N/A	N/A	N/A	N/A	N/A	43.90	48.30	59.50
SBN-mix [39]	69.87	76.33	81.04	89.84	40.30	45.07	50.39	62.37
OTMix [47]	N/A	78.30	83.40	90.20	N/A	46.40	50.70	61.60
ETF+CE [21]	60.06	67.00	77.20	87.00	N/A	N/A	N/A	N/A
ETF+DR [21]	71.90	76.50	81.00	87.70	40.90	45.30	50.40	N/A
2-stage or extra network								
BBN-mix [48]	N/A	79.82	82.18	88.32	N/A	42.56	47.02	59.12
DBN-mix [39]	79.58	83.47	86.82	90.87	**46.21**	**51.04**	54.93	64.98
UniMix [25]	78.48	82.75	84.32	89.66	42.07	45.45	51.11	61.25
MiSLAS [11]	N/A	82.10	85.70	90.00	N/A	47.00	52.30	63.20
CP-Mix [41]	78.34	82.44	85.08	89.87	43.56	48.20	52.12	61.91
class-balance loss								
CB+RS [15]	N/A	70.55	N/A	86.79	N/A	33.44	N/A	55.06
CB+RW [16]	N/A	72.37	N/A	86.54	N/A	33.99	N/A	57.12
CB+Focal [16]	N/A	74.57	N/A	87.10	N/A	36.02	N/A	57.99
LDAM [15]	N/A	73.35	N/A	86.96	N/A	39.60	N/A	56.91
LDAM+DRW [15]	N/A	77.03	N/A	88.16	N/A	42.04	N/A	58.71
class-balance sampling								
CAS [19]	N/A	68.40	N/A	86.90	N/A	31.90	N/A	55.00
LOM [18]	N/A	74.20	N/A	89.40	N/A	41.50	N/A	59.90
CAS+DRW [19]	N/A	73.50	N/A	87.70	N/A	41.50	N/A	57.60
LOM+DRW [18]	N/A	78.70	N/A	89.60	N/A	46.20	N/A	61.10
reproduced results and our method								
Mixup	66.77	72.94	78.64	88.05	39.06	42.88	48.31	63.03
+LOM	70.17	76.63	81.15	89.24	39.61	44.24	49.99	63.90
+CAS	69.90	76.43	81.42	89.24	40.28	44.65	50.07	63.57
+BMLS_{MS}	74.70	79.67	83.46	88.51	41.71	47.62	52.74	64.47
diff.	+7.93	+6.73	+4.82	+0.46	+2.65	+4.74	+4.43	+1.44
ETF+DR	71.58	76.82	81.25	87.59	41.20	45.07	50.71	63.08
BMLS+WETF_{MS}+CE	77.73	80.31	84.22	88.26	42.73	47.10	52.44	64.10
diff.	+6.15	+3.49	+2.97	+0.67	+1.53	+2.03	+1.73	+1.02
Remix	69.58	75.15	80.41	88.61	41.03	44.95	50.19	63.45
+BMLS	73.95	80.10	83.92	88.62	39.95	46.34	51.53	64.42
+BMLS_{MS}	73.18	78.00	83.70	88.20	40.25	46.82	49.78	63.54
diff.	+3.60	+2.85	+3.29	-0.41	-0.78	+1.87	-0.41	+0.09
DBN-mix	77.40	82.40	86.05	**91.01**	40.71	45.52	50.47	62.68
+BMLS_{MS}	**79.73**	**84.30**	**87.28**	90.93	44.42	49.08	**55.41**	**65.42**
diff.	+2.33	+1.90	+1.23	-0.08	+3.71	+3.56	+4.94	+2.74

Table 17: Experimental results on Many/Medium/Few categories in the CIFAR10/100-LT datasets.

	Method	Clf.	CIFAR10-LT				CIFAR100-LT			
			many	med	few	all	many	med	few	all
imb 200	random	FC	$91.17_{3.65}$	$69.99_{2.25}$	$38.09_{6.32}$	$66.77_{0.76}$	$71.16_{0.52}$	$35.22_{0.20}$	$3.85_{0.47}$	$39.06_{0.23}$
	LOM	FC	$82.63_{2.72}$	$69.29_{3.79}$	$58.89_{3.94}$	$70.17_{0.51}$	$65.92_{0.51}$	$39.44_{0.96}$	$7.15_{0.50}$	$39.61_{0.50}$
	CAS	FC	$85.65_{3.74}$	$67.67_{3.86}$	$57.14_{4.43}$	$69.90_{0.77}$	$66.32_{0.55}$	$40.54_{0.59}$	$7.62_{0.28}$	$40.28_{0.29}$
	BMLS	FC	$90.49_{0.26}$	$74.12_{1.21}$	$54.43_{2.25}$	$73.13_{0.67}$	$65.29_{0.45}$	$41.33_{0.76}$	$7.09_{0.37}$	$40.03_{0.38}$
	BMLS	MS	$88.94_{0.32}$	$72.97_{0.83}$	$62.77_{1.30}$	$74.70_{0.45}$	$63.24_{0.57}$	$44.86_{0.42}$	$11.19_{0.76}$	$41.71_{0.36}$
imb 100	random	FC	$93.39_{2.42}$	$74.05_{2.03}$	$50.99_{5.40}$	$72.94_{0.68}$	$72.09_{0.21}$	$41.10_{0.40}$	$8.77_{0.42}$	$42.88_{0.15}$
	LOM	FC	$90.89_{2.45}$	$74.31_{2.99}$	$65.46_{5.76}$	$76.63_{0.41}$	$67.07_{0.74}$	$46.29_{0.72}$	$13.43_{0.37}$	$44.24_{0.14}$
	CAS	FC	$90.54_{2.86}$	$75.54_{1.95}$	$63.51_{6.26}$	$76.43_{0.60}$	$68.28_{0.35}$	$46.47_{0.34}$	$13.12_{0.25}$	$44.65_{0.26}$
	BMLS	FC	$88.53_{1.01}$	$77.84_{0.25}$	$70.53_{1.27}$	$78.85_{0.34}$	$68.38_{0.27}$	$46.89_{0.33}$	$14.37_{0.88}$	$45.20_{0.33}$
	BMLS	MS	$89.14_{0.63}$	$76.34_{0.62}$	$74.63_{0.69}$	$79.67_{0.21}$	$66.31_{0.26}$	$49.80_{0.57}$	$21.80_{0.40}$	$47.62_{0.25}$
imb 50	random	FC	$95.25_{0.23}$	$78.52_{0.54}$	$62.19_{1.04}$	$78.64_{0.57}$	$73.72_{0.18}$	$48.62_{0.23}$	$16.40_{0.93}$	$48.31_{0.28}$
	LOM	FC	$91.57_{3.17}$	$79.62_{1.58}$	$72.78_{4.05}$	$81.15_{0.48}$	$68.80_{0.46}$	$52.97_{0.28}$	$23.06_{0.56}$	$49.99_{0.13}$
	CAS	FC	$92.78_{0.43}$	$79.28_{0.46}$	$72.89_{0.96}$	$81.42_{0.27}$	$69.16_{0.61}$	$52.71_{0.34}$	$23.18_{0.41}$	$50.07_{0.27}$
	BMLS	FC	$91.86_{0.40}$	$81.32_{0.36}$	$76.63_{1.05}$	$83.07_{0.43}$	$69.77_{0.55}$	$54.55_{0.28}$	$26.83_{0.67}$	$51.99_{0.26}$
	BMLS	MS	$89.45_{0.15}$	$79.29_{0.54}$	$83.03_{0.50}$	$83.46_{0.36}$	$67.06_{0.51}$	$55.30_{0.84}$	$31.88_{1.39}$	$52.74_{0.55}$
imb 10	random	FC	$94.79_{0.55}$	$85.38_{0.27}$	$84.86_{1.23}$	$88.05_{0.27}$	$76.06_{0.32}$	$64.10_{0.63}$	$45.56_{0.57}$	$63.03_{0.17}$
	LOM	FC	$93.95_{0.78}$	$86.04_{0.57}$	$88.81_{0.28}$	$89.24_{0.37}$	$72.42_{0.64}$	$65.76_{0.45}$	$51.08_{0.69}$	$63.90_{0.37}$
	CAS	FC	$94.14_{0.23}$	$86.34_{0.24}$	$88.21_{0.43}$	$89.24_{0.18}$	$72.59_{0.49}$	$65.20_{0.47}$	$50.40_{0.66}$	$63.57_{0.26}$
	BMLS	FC	$91.17_{0.40}$	$87.04_{0.26}$	$90.98_{0.59}$	$89.46_{0.19}$	$71.05_{0.70}$	$68.93_{0.66}$	$55.24_{0.47}$	$65.72_{0.29}$
	BMLS	MS	$91.63_{0.60}$	$84.92_{0.63}$	$90.18_{0.56}$	$88.51_{0.19}$	$71.61_{0.21}$	$65.67_{1.20}$	$54.17_{1.49}$	$64.47_{0.24}$

Table 18: Experimental results on Many/Medium/Few categories in the Places-LT datasets.

Method	Clf.	Places-LT				Places-LT (FT)			
		many	med	few	all	many	med	few	all
random	FC	$42.02_{0.76}$	$15.79_{0.54}$	$0.86_{0.12}$	$22.06_{0.50}$	$43.79_{0.29}$	$20.45_{0.27}$	$6.59_{0.26}$	$25.90_{0.06}$
LOM	FC	$38.65_{1.97}$	$22.60_{1.20}$	$5.69_{0.52}$	$24.79_{0.13}$	$41.31_{0.09}$	$39.98_{0.17}$	$25.11_{0.11}$	$37.32_{0.07}$
CAS	FC	$40.68_{0.33}$	$20.08_{0.53}$	$4.86_{0.50}$	$24.26_{0.22}$	$41.35_{0.08}$	$40.06_{0.06}$	$25.46_{0.17}$	$37.44_{0.04}$
BMLS	FC	$38.43_{0.21}$	$27.80_{0.12}$	$7.47_{0.26}$	$27.33_{0.17}$	$34.65_{0.04}$	$43.79_{0.05}$	$29.00_{0.08}$	$37.39_{0.01}$
BMLS	MS	$39.39_{0.32}$	$27.01_{0.40}$	$10.39_{0.12}$	$27.95_{0.26}$	$41.33_{0.09}$	$40.14_{0.00}$	$27.05_{0.15}$	$37.81_{0.01}$

Table 19: Experimental results on Many/Medium/Few categories in the ImageNet-LT and iNaturalist2018 datasets.

Method	Clf.	ImageNet-LT				iNaturalist2018			
		many	med	few	all	many	med	few	all
random	FC	$67.76_{0.43}$	$38.72_{0.50}$	$9.33_{0.28}$	$45.19_{0.43}$	$77.55_{0.39}$	$66.66_{0.38}$	$59.49_{0.38}$	$64.62_{0.31}$
LOM	FC	$62.46_{0.91}$	$44.55_{1.10}$	$20.00_{0.92}$	$47.49_{0.99}$	$63.25_{0.22}$	$68.36_{0.15}$	$66.63_{0.18}$	$67.06_{0.04}$
CAS	FC	$63.04_{0.31}$	$43.83_{0.34}$	$19.53_{0.40}$	$47.31_{0.33}$	$63.99_{0.63}$	$68.80_{0.02}$	$67.10_{0.08}$	$67.55_{0.09}$
BMLS	FC	$62.35_{0.69}$	$46.53_{0.43}$	$23.08_{0.54}$	$48.83_{0.55}$	$64.44_{2.52}$	$68.33_{0.37}$	$66.19_{0.87}$	$66.98_{0.19}$
BMLS	MS	$59.03_{0.89}$	$45.87_{1.01}$	$24.86_{0.92}$	$47.54_{0.94}$	$51.73_{0.83}$	$57.15_{0.14}$	$57.18_{0.28}$	$56.60_{0.18}$

Table 20: Experiments on CIFAR10/100-LT datasets with various imbalance factors for effectiveness of Mixed-Singleton Classifier. All experiments were conducted with BMLS sampler.

$\mathcal{D}$	Clf.	imbalance factor			
		200	100	50	10
C10	FC	73.13	78.85	83.07	89.46
	MS	74.70	79.67	83.46	88.51
	diff.	+1.57	+0.82	+0.39	-0.95
C100	FC	40.03	45.20	51.99	65.72
	MS	41.71	47.62	52.74	64.47
	diff.	+1.68	+2.42	+0.75	-1.25

Table 21: Ablation study including K^2 classifier (notated as K^2) on Many/Medium/Few categories in the CIFAR10/100-LT datasets (imbalance factor: 100). The results are divided into samplers (*Smpl.*), classifier (*Clf.*), and Mixed-Singleton classifier (*MS*) for comparison.

	Method	Clf.	CIFAR10-LT				CIFAR100-LT			
			many	med	few	all	many	med	few	all
Smpl.	random	FC	**93.39**	74.05	50.99	72.94	**72.09**	41.10	8.77	42.88
	BMLS	FC	88.53	**77.84**	**70.53**	**78.85**	68.38	**46.89**	**14.37**	**45.20**
		diff.	-4.86	+3.79	+19.54	+5.91	-3.71	+5.79	+5.60	+2.32
Clf.	random	MS	**89.47**	62.24	41.15	64.08	**67.29**	33.84	2.78	36.87
	BMLS	K^2	58.35	34.31	25.93	39.01	14.15	9.36	1.21	8.72
	BMLS	MS	89.14	**76.34**	**74.63**	**79.67**	66.31	**49.80**	**21.80**	**47.62**
		diff.	-0.33	+14.10	+33.48	+15.59	-0.98	+15.96	+19.02	+10.75
MS	diff. (BMLS_{MS} - BMLS)		+0.61	-1.50	+4.10	+0.82	-2.07	+2.91	+7.43	+2.42

Table 22: Experimental results of the ablation study on Many/Medium/Few categories in the CIFAR10/100-LT datasets. The results are the mean of five repeated experiments with random seeds.

	Method	Clf.	CIFAR10-LT				CIFAR100-LT			
			many	med	few	all	many	med	few	all
imb 200	random	FC	$91.17_{3.65}$	$69.99_{2.25}$	$38.09_{6.32}$	$66.77_{0.76}$	$71.16_{0.52}$	$35.22_{0.20}$	$3.85_{0.47}$	$39.06_{0.23}$
	random	MS	$88.59_{0.19}$	$53.77_{1.07}$	$16.74_{1.04}$	$53.11_{0.58}$	$64.59_{0.75}$	$28.43_{0.49}$	$0.75_{0.15}$	$33.42_{0.37}$
	BMLS	FC	$90.49_{0.26}$	$74.12_{1.21}$	$54.43_{2.25}$	$73.13_{0.67}$	$65.29_{0.45}$	$41.33_{0.76}$	$7.09_{0.37}$	$40.03_{0.38}$
	BMLS	MS	$88.94_{0.32}$	$72.97_{0.83}$	$62.77_{1.30}$	$74.70_{0.45}$	$63.24_{0.57}$	$44.86_{0.42}$	$11.19_{0.76}$	$41.71_{0.36}$
imb 100	random	FC	$93.39_{2.42}$	$74.05_{2.03}$	$50.99_{5.40}$	$72.94_{0.68}$	$72.09_{0.21}$	$41.10_{0.40}$	$8.77_{0.42}$	$42.88_{0.15}$
	random	MS	$89.47_{0.46}$	$62.24_{2.21}$	$41.15_{3.22}$	$64.08_{1.59}$	$67.29_{0.31}$	$33.84_{0.61}$	$2.78_{0.32}$	$36.87_{0.24}$
	BMLS	FC	$88.53_{1.01}$	$77.84_{0.25}$	$70.53_{1.27}$	$78.85_{0.34}$	$68.38_{0.27}$	$46.89_{0.33}$	$14.37_{0.88}$	$45.20_{0.33}$
	BMLS	MS	$89.14_{0.63}$	$76.34_{0.62}$	$74.63_{0.69}$	$79.67_{0.21}$	$66.31_{0.26}$	$49.80_{0.57}$	$21.80_{0.40}$	$47.62_{0.25}$
imb 50	random	FC	$95.25_{0.23}$	$78.52_{0.54}$	$62.19_{1.04}$	$78.64_{0.57}$	$73.72_{0.18}$	$48.62_{0.23}$	$16.40_{0.93}$	$48.31_{0.28}$
	random	MS	$90.04_{0.63}$	$64.80_{1.76}$	$52.11_{1.06}$	$68.56_{0.50}$	$68.28_{0.69}$	$42.17_{0.89}$	$8.00_{0.52}$	$41.66_{0.42}$
	BMLS	FC	$91.86_{0.40}$	$81.32_{0.36}$	$76.63_{1.05}$	$83.07_{0.43}$	$69.77_{0.55}$	$54.55_{0.28}$	$26.83_{0.67}$	$51.99_{0.26}$
	BMLS	MS	$89.45_{0.15}$	$79.29_{0.54}$	$83.03_{0.50}$	$83.46_{0.36}$	$67.06_{0.51}$	$55.30_{0.84}$	$31.88_{1.39}$	$52.74_{0.55}$
imb 10	random	FC	$94.79_{0.55}$	$85.38_{0.27}$	$84.86_{1.23}$	$88.05_{0.27}$	$76.06_{0.32}$	$64.10_{0.63}$	$45.56_{0.57}$	$63.03_{0.17}$
	random	MS	$91.54_{0.43}$	$76.31_{1.02}$	$75.25_{1.44}$	$80.56_{0.76}$	$71.91_{0.35}$	$57.38_{0.90}$	$37.03_{0.72}$	$56.71_{0.48}$
	BMLS	FC	$91.17_{0.40}$	$87.04_{0.26}$	$90.98_{0.59}$	$89.46_{0.19}$	$71.05_{0.70}$	$68.93_{0.66}$	$55.24_{0.47}$	$65.72_{0.29}$
	BMLS	MS	$91.63_{0.60}$	$84.92_{0.63}$	$90.18_{0.56}$	$88.51_{0.19}$	$71.61_{0.21}$	$65.67_{1.20}$	$54.17_{1.49}$	$64.47_{0.24}$

Table 23: Extension to the fixed ETF classifier on CIFAR10/100-LT datasets with various imbalance factors. The results are the mean of five repeated experiments with random seeds. Best in bold (†: the reported values are taken from [21])

Sampler	Clf.	$\mathcal{L}$	CIFAR10-LT				CIFAR100-LT			
			imbalance factor				imbalance factor			
			200	100	50	10	200	100	50	10
random	ETF	CE†	60.06	67.00	77.20	87.00	N/A	N/A	N/A	N/A
random	ETF	DR†	71.90	76.50	81.00	87.70	40.90	45.30	50.40	N/A
random	ETF	DR	71.58	76.82	81.25	87.59	41.20	45.07	50.71	63.08
LOM	ETF	DR	69.35	75.46	81.15	88.38	38.78	42.96	48.84	62.01
CAS	ETF	DR	69.17	76.16	80.81	**88.61**	38.91	43.18	49.05	62.50
BMLS	ETF	DR	**77.77**	**80.38**	**84.30**	87.91	39.54	43.60	49.54	62.06
		diff.	+6.19	+3.56	+3.05	+0.32	-1.66	-1.47	-1.17	-1.02
BMLS	MS-WETF	CE	77.73	80.31	84.22	88.26	**42.73**	**47.10**	**52.44**	**64.10**
		diff.	+6.15	+3.49	+2.97	+0.67	+1.53	+2.03	+1.73	+1.02

I.7 Comparison Experiments for ETF+DR

Table 24: Experimental results of extension to the fixed ETF Classifier on Many/Medium/Few categories in the CIFAR10-LT dataset. The results are the mean of five repeated experiments with random seeds.

	Method	Clf.	$\mathcal{L}$	CIFAR10-LT			
				many	med	few	all
imb 200	random	ETF	DR	$84.13_{0.64}$	$73.89_{0.92}$	$55.94_{1.24}$	$71.58_{0.39}$
	LOM	ETF	DR	$81.05_{3.12}$	$69.26_{2.29}$	$57.77_{4.75}$	$69.35_{0.38}$
	CAS	ETF	DR	$87.67_{6.09}$	$72.17_{0.94}$	$46.67_{6.61}$	$69.17_{0.67}$
	BMLS	ETF	DR	$84.52_{0.47}$	$74.15_{0.36}$	$75.85_{0.66}$	$77.77_{0.13}$
	BMLS	MS-WETF	CE	$85.41_{0.71}$	$74.96_{0.45}$	$73.74_{0.72}$	$77.73_{0.32}$
imb 100	random	ETF	DR	$83.75_{0.92}$	$75.42_{0.30}$	$71.75_{0.95}$	$76.82_{0.20}$
	LOM	ETF	DR	$88.89_{3.19}$	$74.46_{2.41}$	$63.37_{6.15}$	$75.46_{0.37}$
	CAS	ETF	DR	$91.03_{0.54}$	$75.97_{0.44}$	$61.55_{2.15}$	$76.16_{0.56}$
	BMLS	ETF	DR	$88.85_{0.16}$	$77.51_{0.39}$	$75.74_{0.42}$	$80.38_{0.23}$
	BMLS	MS-WETF	CE	$86.71_{0.88}$	$76.28_{0.69}$	$79.27_{1.39}$	$80.31_{0.43}$
imb 50	random	ETF	DR	$85.45_{0.50}$	$78.60_{0.28}$	$80.59_{0.42}$	$81.25_{0.18}$
	LOM	ETF	DR	$91.41_{1.07}$	$79.15_{1.05}$	$73.57_{1.93}$	$81.15_{0.37}$
	CAS	ETF	DR	$91.02_{1.68}$	$79.26_{1.09}$	$72.68_{2.07}$	$80.81_{0.22}$
	BMLS	ETF	DR	$88.17_{0.24}$	$80.21_{0.19}$	$85.87_{0.20}$	$84.30_{0.07}$
	BMLS	MS-WETF	CE	$87.01_{0.89}$	$80.36_{0.67}$	$86.59_{0.29}$	$84.22_{0.43}$
imb 10	random	ETF	DR	$89.67_{0.52}$	$83.81_{0.28}$	$90.54_{0.39}$	$87.59_{0.18}$
	LOM	ETF	DR	$92.79_{0.23}$	$85.14_{0.38}$	$88.28_{0.41}$	$88.38_{0.25}$
	CAS	ETF	DR	$92.87_{0.28}$	$85.33_{0.60}$	$88.72_{0.22}$	$88.61_{0.21}$
	BMLS	ETF	DR	$88.76_{0.93}$	$85.08_{1.00}$	$90.83_{0.79}$	$87.91_{0.24}$
	BMLS	MS-WETF	CE	$91.27_{0.32}$	$85.89_{0.20}$	$88.40_{0.42}$	$88.26_{0.04}$

Table 25: Experimental results of extension to the fixed ETF Classifier on Many/Medium/Few categories in the CIFAR100-LT dataset. The results are the mean of five repeated experiments with random seeds.

	Method	Clf.	$\mathcal{L}$	CIFAR100-LT			
				many	med	few	all
imb 200	random	ETF	DR	$68.23_{0.59}$	$42.05_{0.52}$	$6.63_{0.29}$	$41.20_{0.18}$
	LOM	ETF	DR	$63.90_{1.17}$	$38.98_{0.81}$	$7.36_{0.77}$	$38.78_{0.25}$
	CAS	ETF	DR	$64.10_{0.66}$	$38.86_{0.68}$	$7.68_{0.31}$	$38.91_{0.43}$
	BMLS	ETF	DR	$63.81_{0.48}$	$39.09_{0.69}$	$9.94_{0.54}$	$39.54_{0.45}$
	BMLS	MS-WETF	CE	$65.58_{0.70}$	$45.26_{0.54}$	$11.32_{0.52}$	$42.73_{0.41}$
imb 100	random	ETF	DR	$69.85_{0.40}$	$47.22_{0.35}$	$11.72_{0.81}$	$45.07_{0.25}$
	LOM	ETF	DR	$65.43_{0.88}$	$44.78_{0.94}$	$12.88_{0.91}$	$42.96_{0.25}$
	CAS	ETF	DR	$66.04_{0.40}$	$44.73_{0.34}$	$12.93_{0.35}$	$43.18_{0.18}$
	BMLS	ETF	DR	$65.59_{0.18}$	$44.49_{0.45}$	$15.21_{0.49}$	$43.60_{0.22}$
	BMLS	MS-WETF	CE	$63.44_{0.32}$	$51.15_{0.87}$	$21.92_{0.72}$	$47.10_{0.47}$
imb 50	random	ETF	DR	$70.56_{0.39}$	$53.52_{0.65}$	$22.69_{0.70}$	$50.71_{0.24}$
	LOM	ETF	DR	$67.73_{0.54}$	$51.15_{0.13}$	$22.59_{0.50}$	$48.84_{0.16}$
	CAS	ETF	DR	$67.87_{0.55}$	$51.58_{0.62}$	$22.63_{0.78}$	$49.05_{0.36}$
	BMLS	ETF	DR	$66.21_{0.58}$	$51.02_{0.49}$	$27.06_{0.59}$	$49.54_{0.39}$
	BMLS	MS-WETF	CE	$67.02_{0.90}$	$54.66_{0.62}$	$31.66_{0.41}$	$52.44_{0.40}$
imb 10	random	ETF	DR	$72.76_{0.29}$	$64.48_{0.50}$	$49.39_{0.36}$	$63.08_{0.21}$
	LOM	ETF	DR	$70.89_{0.43}$	$63.73_{0.42}$	$48.90_{0.49}$	$62.01_{0.19}$
	CAS	ETF	DR	$71.13_{0.45}$	$63.89_{0.34}$	$50.12_{0.45}$	$62.50_{0.27}$
	BMLS	ETF	DR	$68.95_{1.20}$	$64.83_{0.71}$	$50.18_{1.26}$	$62.06_{0.22}$
	BMLS	MS-WETF	CE	$68.81_{0.40}$	$64.95_{0.46}$	$57.24_{0.28}$	$64.10_{0.25}$

I.8 Comparison Experiments for Remix

Table 26: Comparison experiments of Remix on CIFAR10/100-LT datasets with various imbalance factors. The results are the mean of five repeated experiments with random seeds. Best in bold (LOM: Label-Occurrence Mixup, CAS: Class-Aware Sampler, BMLS: Balanced Mixed Label Sampler, †: the reported values are taken from [38], which used different experimental settings. ∗: the reproduced result of Remix on our experimental settings.)

Method	CIFAR10-LT				CIFAR100-LT			
	imbalance factor				imbalance factor			
	200	100	50	10	200	100	50	10
Remix†	N/A	75.36	N/A	88.15	N/A	41.94	N/A	59.36
$\text{Remix}^{\dagger}_{\text{RS}}$	N/A	76.23	N/A	87.70	N/A	41.13	N/A	58.62
Remix*	69.58	75.15	80.41	88.61	**41.03**	44.95	50.19	63.45
+LOM	71.39	76.72	82.03	**89.39**	39.95	43.72	49.46	63.49
+CAS	71.36	77.28	82.00	89.37	40.21	44.91	49.83	63.26
+BMLS	**73.95**	**80.10**	**83.92**	88.62	39.95	46.34	**51.53**	**64.42**
$+\text{BMLS}_{\text{MS}}$	73.18	78.00	83.70	88.20	40.25	**46.82**	49.78	63.54

Table 27: Experimental results of Remix on Many/Medium/Few categories in the CIFAR10/100-LT datasets. The results are the mean of five repeated experiments with random seeds. (†: the reported values are taken from [38], which used different experimental settings. ∗: the reproduced result of Remix on our experimental settings.)

	Method	CIFAR10-LT				CIFAR100-LT			
		many	med	few	all	many	med	few	all
imb 200	Remix†	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	$\text{Remix}^{\dagger}_{\text{RS}}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Remix*	$\mathbf{92.31}_{3.87}$	$71.40_{1.23}$	$44.43_{7.91}$	$69.58_{0.99}$	$\mathbf{70.43}_{0.23}$	$39.83_{1.02}$	$5.99_{0.42}$	$\mathbf{41.03}_{0.31}$
	+LOM	$90.15_{1.32}$	$72.19_{1.88}$	$51.56_{5.63}$	$71.39_{0.87}$	$63.48_{1.25}$	$41.62_{0.83}$	$8.70_{1.00}$	$39.95_{0.17}$
	+CAS	$88.35_{4.44}$	$71.63_{1.52}$	$54.02_{8.11}$	$71.36_{0.91}$	$64.44_{0.36}$	$41.06_{0.71}$	$9.11_{0.24}$	$40.21_{0.35}$
	+BMLS	$80.43_{7.52}$	$\mathbf{73.66}_{1.38}$	$\mathbf{67.86}_{7.14}$	$\mathbf{73.95}_{0.48}$	$61.38_{3.17}$	$\mathbf{42.97}_{0.88}$	$\mathbf{9.70}_{4.35}$	$39.95_{0.55}$
	$+\text{BMLS}_{\text{MS}}$	$89.47_{0.44}$	$72.17_{0.54}$	$58.23_{1.19}$	$73.18_{0.22}$	$64.88_{0.30}$	$40.37_{0.77}$	$9.55_{0.46}$	$40.25_{0.32}$
imb 100	Remix†	N/A	N/A	N/A	75.36	N/A	N/A	N/A	41.94
	$\text{Remix}^{\dagger}_{\text{RS}}$	N/A	N/A	N/A	76.23	N/A	N/A	N/A	41.13
	Remix*	$\mathbf{93.70}_{0.60}$	$76.23_{0.67}$	$55.17_{1.41}$	$75.15_{0.23}$	$\mathbf{71.16}_{0.41}$	$45.58_{0.85}$	$11.67_{0.80}$	$44.95_{0.37}$
	+LOM	$91.31_{0.63}$	$76.54_{1.27}$	$62.37_{1.84}$	$76.72_{0.62}$	$64.42_{0.30}$	$46.74_{0.62}$	$14.38_{0.32}$	$43.72_{0.29}$
	+CAS	$90.76_{0.81}$	$76.72_{0.85}$	$64.55_{1.22}$	$77.28_{0.43}$	$66.21_{0.43}$	$47.48_{0.34}$	$15.36_{0.80}$	$44.91_{0.20}$
	+BMLS	$89.23_{2.64}$	$\mathbf{77.78}_{1.46}$	$\mathbf{74.05}_{1.18}$	$\mathbf{80.10}_{0.36}$	$64.88_{0.47}$	$48.86_{0.46}$	$\mathbf{20.28}_{0.46}$	$46.34_{0.29}$
	$+\text{BMLS}_{\text{MS}}$	$91.25_{0.64}$	$74.78_{0.75}$	$69.05_{1.84}$	$78.00_{0.36}$	$67.21_{0.47}$	$\mathbf{48.92}_{0.20}$	$18.97_{0.77}$	$\mathbf{46.82}_{0.30}$
imb 50	Remix†	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	$\text{Remix}^{\dagger}_{\text{RS}}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Remix*	$\mathbf{94.25}_{0.49}$	$79.20_{0.31}$	$68.16_{1.33}$	$80.41_{0.25}$	$\mathbf{72.03}_{0.51}$	$51.73_{0.15}$	$21.22_{0.84}$	$50.19_{0.24}$
	+LOM	$91.47_{0.71}$	$79.68_{0.52}$	$75.73_{1.57}$	$82.03_{0.34}$	$67.05_{0.30}$	$52.47_{0.59}$	$23.98_{0.31}$	$49.46_{0.26}$
	+CAS	$92.08_{0.32}$	$79.89_{0.64}$	$74.72_{1.14}$	$82.00_{0.48}$	$67.88_{0.63}$	$52.42_{0.24}$	$24.28_{0.33}$	$49.83_{0.20}$
	+BMLS	$90.63_{0.44}$	$\mathbf{81.30}_{0.33}$	$80.70_{1.15}$	$\mathbf{83.92}_{0.38}$	$64.89_{0.40}$	$\mathbf{53.63}_{0.57}$	$\mathbf{32.41}_{0.53}$	$\mathbf{51.53}_{0.40}$
	$+\text{BMLS}_{\text{MS}}$	$89.71_{0.49}$	$80.09_{0.57}$	$\mathbf{82.49}_{0.87}$	$83.70_{0.16}$	$67.16_{1.86}$	$51.46_{0.82}$	$26.17_{1.78}$	$49.78_{0.43}$
imb 10	Remix†	N/A	N/A	N/A	88.15	N/A	N/A	N/A	59.36
	$\text{Remix}^{\dagger}_{\text{RS}}$	N/A	N/A	N/A	87.70	N/A	N/A	N/A	58.62
	Remix*	$\mathbf{94.85}_{0.65}$	$85.44_{0.43}$	$86.59_{0.42}$	$88.61_{0.18}$	$\mathbf{75.03}_{0.58}$	$63.77_{0.40}$	$48.70_{0.85}$	$63.45_{0.40}$
	+LOM	$93.75_{0.19}$	$85.79_{0.55}$	$89.83_{0.43}$	$\mathbf{89.39}_{0.17}$	$70.47_{0.50}$	$65.90_{0.38}$	$51.91_{0.57}$	$63.49_{0.16}$
	+CAS	$93.64_{0.72}$	$\mathbf{86.18}_{0.61}$	$89.36_{0.84}$	$89.37_{0.24}$	$70.72_{0.43}$	$65.45_{0.41}$	$51.37_{0.63}$	$63.26_{0.10}$
	+BMLS	$89.75_{0.43}$	$85.66_{0.17}$	$\mathbf{91.45}_{0.22}$	$88.62_{0.11}$	$69.28_{1.67}$	$\mathbf{68.32}_{1.15}$	$\mathbf{53.68}_{1.36}$	$\mathbf{64.42}_{0.30}$
	$+\text{BMLS}_{\text{MS}}$	$92.26_{0.17}$	$84.59_{0.34}$	$88.95_{0.21}$	$88.20_{0.15}$	$70.59_{0.42}$	$65.11_{0.32}$	$52.88_{0.61}$	$63.54_{0.26}$

I.9 Ablation Study including K2 Classifier

Table 28: Ablation study on CIFAR10/100-LT datasets with various imbalance factors including K^2 classifier (notated as K^2 on the table). The results are the mean of five repeated experiments with random seeds. Best in bold

Sampler	Clf.	CIFAR10-LT				CIFAR100-LT			
		imbalance factor				imbalance factor			
		200	100	50	10	200	100	50	10
Sampler									
random	FC	66.77	72.94	78.64	88.05	39.06	42.88	48.31	63.03
BMLS	FC	**73.13**	**78.85**	**83.07**	**89.46**	**40.03**	**45.20**	**51.99**	**65.72**
Classifier									
random	MS	53.11	64.08	68.56	80.56	33.42	36.87	41.66	56.71
BMLS	K^2	34.86	39.01	42.20	51.60	7.90	8.72	9.22	16.41
BMLS	MS	**74.70**	**79.67**	**83.46**	**88.51**	**41.71**	**47.62**	**52.74**	**64.47**

Table 29: Experimental results of the ablation study including K^2 classifier (notated as K^2 on the table) on Many/Medium/Few categories in the CIFAR10/100-LT datasets. The results are the mean of five repeated experiments with random seeds.

	Method	Clf.	CIFAR10-LT				CIFAR100-LT			
			many	med	few	all	many	med	few	all
imb 200	*Sampler*									
	random	FC	**91.17**$_{3.65}$	69.99$_{2.25}$	38.09$_{6.32}$	66.77$_{0.76}$	**71.16**$_{0.52}$	35.22$_{0.20}$	3.85$_{0.47}$	39.06$_{0.23}$
	BMLS	FC	90.49$_{0.26}$	**74.12**$_{1.21}$	**54.43**$_{2.25}$	**73.13**$_{0.67}$	65.29$_{0.45}$	**41.33**$_{0.76}$	**7.09**$_{0.37}$	**40.03**$_{0.38}$
	Classifier									
	random	MS	88.59$_{0.19}$	53.77$_{1.07}$	16.74$_{1.04}$	53.11$_{0.58}$	**64.59**$_{0.75}$	28.43$_{0.49}$	0.75$_{0.15}$	33.42$_{0.37}$
	BMLS	K^2	67.94$_{11.93}$	29.79$_{6.92}$	8.55$_{5.98}$	34.86$_{0.92}$	14.69$_{0.75}$	6.91$_{0.63}$	0.67$_{0.17}$	7.90$_{0.13}$
	BMLS	MS	**88.94**$_{0.32}$	**72.97**$_{0.83}$	**62.77**$_{1.30}$	**74.70**$_{0.45}$	63.24$_{0.57}$	**44.86**$_{0.42}$	**11.19**$_{0.76}$	**41.71**$_{0.36}$
imb 100	*Sampler*									
	random	FC	**93.39**$_{2.42}$	74.05$_{2.03}$	50.99$_{5.40}$	72.94$_{0.68}$	**72.09**$_{0.21}$	41.10$_{0.40}$	8.77$_{0.42}$	42.88$_{0.15}$
	BMLS	FC	88.53$_{1.01}$	**77.84**$_{0.25}$	**70.53**$_{1.27}$	**78.85**$_{0.34}$	68.38$_{0.27}$	**46.89**$_{0.33}$	**14.37**$_{0.88}$	**45.20**$_{0.33}$
	Classifier									
	random	MS	**89.47**$_{0.46}$	62.24$_{2.21}$	41.15$_{3.22}$	64.08$_{1.59}$	**67.29**$_{0.31}$	33.84$_{0.61}$	2.78$_{0.32}$	36.87$_{0.24}$
	BMLS	K^2	58.35$_{3.36}$	34.31$_{5.73}$	25.93$_{5.20}$	39.01$_{0.90}$	14.15$_{0.60}$	9.36$_{0.75}$	1.21$_{0.12}$	8.72$_{0.43}$
	BMLS	MS	89.14$_{0.63}$	**76.34**$_{0.62}$	**74.63**$_{0.69}$	**79.67**$_{0.21}$	66.31$_{0.26}$	**49.80**$_{0.57}$	**21.80**$_{0.40}$	**47.62**$_{0.25}$
imb 50	*Sampler*									
	random	FC	**95.25**$_{0.23}$	78.52$_{0.54}$	62.19$_{1.04}$	78.64$_{0.57}$	**73.72**$_{0.18}$	48.62$_{0.23}$	16.40$_{0.93}$	48.31$_{0.28}$
	BMLS	FC	91.86$_{0.40}$	**81.32**$_{0.36}$	**76.63**$_{1.05}$	**83.07**$_{0.43}$	69.77$_{0.55}$	**54.55**$_{0.28}$	**26.83**$_{0.67}$	**51.99**$_{0.26}$
	Classifier									
	random	MS	**90.04**$_{0.63}$	64.80$_{1.76}$	52.11$_{1.06}$	68.56$_{0.50}$	**68.28**$_{0.69}$	42.17$_{0.89}$	8.00$_{0.52}$	41.66$_{0.42}$
	BMLS	K^2	59.75$_{7.70}$	37.92$_{0.59}$	30.37$_{7.19}$	42.20$_{0.89}$	13.25$_{1.08}$	10.40$_{0.49}$	2.81$_{0.96}$	9.22$_{0.39}$
	BMLS	MS	89.45$_{0.15}$	**79.29**$_{0.54}$	**83.03**$_{0.50}$	**83.46**$_{0.36}$	67.06$_{0.51}$	**55.30**$_{0.84}$	**31.88**$_{1.39}$	**52.74**$_{0.55}$
imb 10	*Sampler*									
	random	FC	**94.79**$_{0.55}$	85.38$_{0.27}$	84.86$_{1.23}$	88.05$_{0.27}$	**76.06**$_{0.32}$	64.10$_{0.63}$	45.56$_{0.57}$	63.03$_{0.17}$
	BMLS	FC	91.17$_{0.40}$	**87.04**$_{0.26}$	**90.98**$_{0.59}$	**89.46**$_{0.19}$	71.05$_{0.70}$	**68.93**$_{0.66}$	**55.24**$_{0.47}$	**65.72**$_{0.29}$
	Classifier									
	random	MS	91.54$_{0.43}$	76.31$_{1.02}$	75.25$_{1.44}$	80.56$_{0.76}$	**71.91**$_{0.35}$	57.38$_{0.90}$	37.03$_{0.72}$	56.71$_{0.48}$
	BMLS	K^2	57.78$_{1.43}$	46.29$_{2.03}$	52.52$_{1.76}$	51.60$_{0.99}$	17.09$_{1.33}$	17.71$_{0.71}$	13.97$_{0.83}$	16.41$_{0.51}$
	BMLS	MS	**91.63**$_{0.60}$	**84.92**$_{0.63}$	**90.18**$_{0.56}$	**88.51**$_{0.19}$	71.61$_{0.21}$	**65.67**$_{1.20}$	**54.17**$_{1.49}$	**64.47**$_{0.24}$

I.10 An Empirical Study on Mixup Alpha

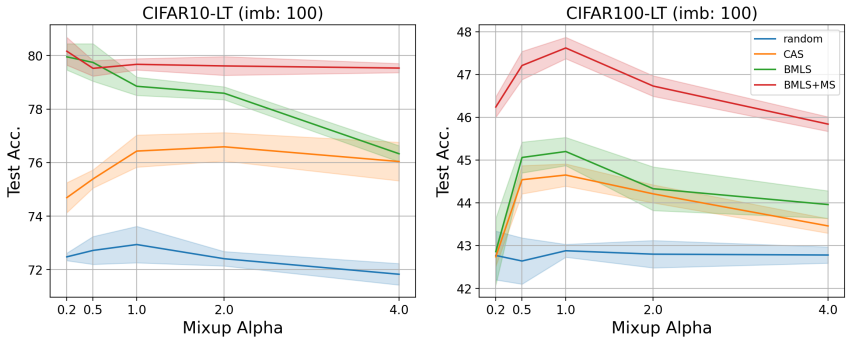


Figure 6: The change of test accuracy of each sampler on CIFAR10/100-LT (imb: 100)

Table 30: Ablation study on CIFAR10/100-LT datasets (imbalance factor: 100) with various mixup alpha values. The results are the mean of five repeated experiments with random seeds. Best in bold (CAS: Class-Aware Sampler, BMLS: Balanced Mixed Label Sampler)

Method	CIFAR10-LT					CIFAR100-LT				
	mixup alpha					mixup alpha				
	0.2	0.5	1.0	2.0	4.0	0.2	0.5	1.0	2.0	4.0
random	72.48	72.72	72.94	72.41	71.83	42.77	42.64	42.88	42.80	42.78
CAS	74.69	75.39	76.43	76.59	76.04	42.74	44.54	44.65	44.21	43.46
BMLS	79.95	**79.74**	78.85	78.59	76.33	42.86	45.06	45.20	44.33	43.96
BMLS_{MS}	**80.16**	79.52	**79.67**	**79.61**	**79.53**	**46.24**	**47.21**	**47.62**	**46.73**	**45.84**

Table 31: Experimental results of an empirical study about Mixup alpha on Many/Medium/Few categories in the CIFAR10/100-LT datasets. The results are the mean of five repeated experiments with random seeds.

	Method	CIFAR10-LT				CIFAR100-LT			
		many	med	few	all	many	med	few	all
$\alpha = 0.2$	random	$\mathbf{92.90}_{1.54}$	$72.28_{2.47}$	$52.34_{4.80}$	$72.48_{0.14}$	$\mathbf{71.69}_{0.65}$	$41.39_{0.71}$	$8.51_{0.68}$	$42.77_{0.57}$
	CAS	$89.05_{4.00}$	$72.29_{3.03}$	$63.53_{7.84}$	$74.69_{0.56}$	$66.31_{0.40}$	$43.65_{0.57}$	$12.39_{0.45}$	$42.74_{0.22}$
	BMLS	$89.73_{0.57}$	$\mathbf{76.84}_{0.54}$	$74.33_{1.83}$	$79.95_{0.49}$	$65.51_{1.01}$	$43.53_{0.93}$	$13.92_{0.71}$	$42.86_{0.78}$
	BMLS_{MS}	$87.15_{0.79}$	$75.39_{0.72}$	$\mathbf{79.53}_{1.80}$	$\mathbf{80.16}_{0.52}$	$64.73_{0.50}$	$\mathbf{48.93}_{0.62}$	$\mathbf{20.06}_{0.24}$	$\mathbf{46.24}_{0.24}$
$\alpha = 0.5$	random	$91.53_{3.38}$	$73.06_{3.07}$	$53.47_{6.92}$	$72.72_{0.52}$	$\mathbf{72.42}_{0.10}$	$40.62_{0.90}$	$8.12_{0.78}$	$42.64_{0.54}$
	CAS	$\mathbf{91.74}_{2.24}$	$74.91_{2.60}$	$59.66_{4.84}$	$75.39_{0.34}$	$68.63_{0.25}$	$45.83_{0.33}$	$13.08_{0.71}$	$44.54_{0.33}$
	BMLS	$90.96_{0.58}$	$\mathbf{78.39}_{0.40}$	$70.34_{2.05}$	$\mathbf{79.74}_{0.70}$	$67.86_{0.68}$	$46.46_{0.31}$	$15.06_{1.23}$	$45.06_{0.36}$
	BMLS_{MS}	$88.88_{1.02}$	$75.65_{0.56}$	$\mathbf{75.31}_{1.21}$	$79.52_{0.29}$	$64.13_{1.12}$	$\mathbf{50.41}_{0.82}$	$\mathbf{22.34}_{0.45}$	$\mathbf{47.21}_{0.33}$
$\alpha = 1.0$	random	$\mathbf{93.39}_{2.42}$	$74.05_{2.03}$	$50.99_{5.40}$	$72.94_{0.68}$	$\mathbf{72.09}_{0.21}$	$41.10_{0.40}$	$8.77_{0.42}$	$42.88_{0.15}$
	CAS	$90.54_{2.86}$	$75.54_{1.95}$	$63.51_{6.26}$	$76.43_{0.60}$	$68.28_{0.35}$	$46.47_{0.34}$	$13.12_{0.25}$	$44.65_{0.26}$
	BMLS	$88.53_{1.01}$	$\mathbf{77.84}_{0.25}$	$70.53_{1.27}$	$78.85_{0.34}$	$68.38_{0.27}$	$46.89_{0.33}$	$14.37_{0.88}$	$45.20_{0.33}$
	BMLS_{MS}	$89.14_{0.63}$	$76.34_{0.62}$	$\mathbf{74.63}_{0.69}$	$\mathbf{79.67}_{0.21}$	$66.31_{0.26}$	$\mathbf{49.80}_{0.57}$	$\mathbf{21.80}_{0.40}$	$\mathbf{47.62}_{0.25}$
$\alpha = 2.0$	random	$\mathbf{93.97}_{0.29}$	$73.29_{0.41}$	$49.65_{1.54}$	$72.41_{0.27}$	$\mathbf{71.92}_{0.39}$	$41.45_{0.71}$	$8.30_{0.23}$	$42.80_{0.32}$
	CAS	$88.30_{3.08}$	$75.95_{2.20}$	$65.74_{5.88}$	$76.59_{0.53}$	$66.38_{0.68}$	$47.04_{0.27}$	$13.27_{0.46}$	$44.21_{0.21}$
	BMLS	$88.17_{0.57}$	$\mathbf{77.24}_{1.02}$	$70.80_{1.53}$	$78.59_{0.24}$	$66.73_{1.07}$	$47.30_{0.71}$	$12.92_{0.89}$	$44.33_{0.51}$
	BMLS_{MS}	$84.22_{0.41}$	$74.66_{0.72}$	$\mathbf{81.59}_{0.31}$	$\mathbf{79.61}_{0.35}$	$65.15_{1.60}$	$\mathbf{49.53}_{1.26}$	$\mathbf{20.47}_{0.41}$	$\mathbf{46.73}_{0.24}$
$\alpha = 4.0$	random	$\mathbf{93.18}_{0.30}$	$71.64_{0.97}$	$50.75_{0.43}$	$71.83_{0.40}$	$\mathbf{71.84}_{0.31}$	$41.72_{0.37}$	$7.99_{0.76}$	$42.78_{0.19}$
	CAS	$87.37_{3.19}$	$75.20_{2.05}$	$65.84_{6.09}$	$76.04_{0.72}$	$64.18_{0.54}$	$47.25_{0.25}$	$13.15_{1.00}$	$43.46_{0.17}$
	BMLS	$86.59_{2.05}$	$\mathbf{77.49}_{0.78}$	$64.52_{2.86}$	$76.33_{0.30}$	$64.02_{0.24}$	$47.23_{0.53}$	$15.09_{0.48}$	$43.96_{0.32}$
	BMLS_{MS}	$86.57_{0.79}$	$73.88_{0.86}$	$\mathbf{80.03}_{0.89}$	$\mathbf{79.53}_{0.17}$	$61.20_{1.02}$	$\mathbf{50.70}_{1.28}$	$\mathbf{20.89}_{0.55}$	$\mathbf{45.84}_{0.17}$

I.11 Ablation study on mini-batch size

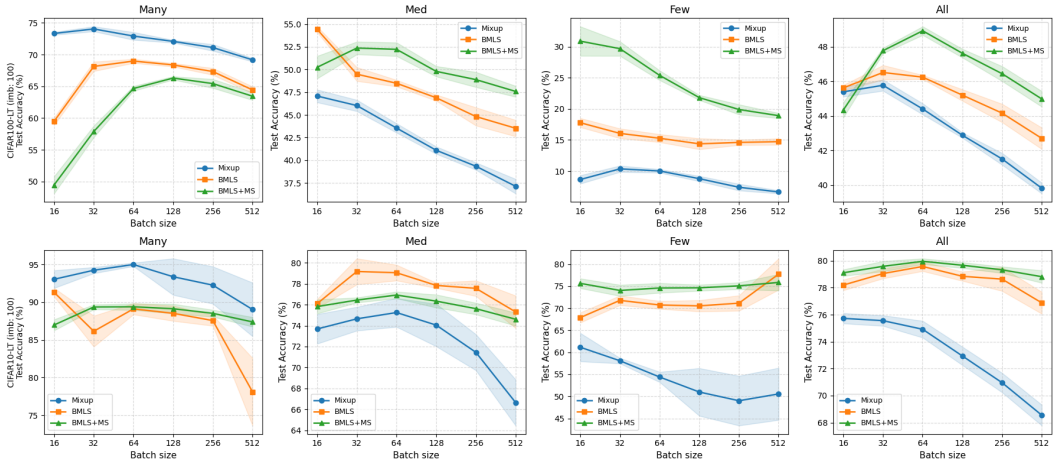


Figure 7: Experiments on various mini-batch size settings in CIFAR100-LT (top row) and CIFAR10-LT (bottom row).

We compared the performance of three methods, Mixup with random sampler, BMLS, and BMLS+MS, on CIFAR10/100-LT (imbalance factor: 100) with mini-batch sizes set to [16, 32, 64, 128, 256, 512] using five random seeds. As illustrated in Figure 7, while performance is affected by the mini-batch size, BMLS and BMLS+MS consistently outperform Mixup across all mini-batch sizes, except when bsz is 16 in CIFAR100-LT. In particular, the performance on the *Few* category is notably higher at a batch size of 16, which suggests that our proposed method further alleviates minority collapse.

I.12 Computational cost

Table 32: Experiments for computational cost on CIFAR10/100-LT (imbalance factor: 100). (t: Elapsed time (sec/epoch), M: Peak GPU Memory (MB), # params: number of learnable parameters)

CIFAR10-LT				CIFAR100-LT			
Method	t	M	# params	Method	t	M	# params
Mixup	5.2023	338.79	464,154	Mixup	5.7773	670.82	1,861,508
BMLS	6.6693	338.79	464,154	BMLS	6.1372	670.82	1,861,508
BMLS+MS	6.7795	338.81	464,154	BMLS+MS	6.2828	670.90	1,861,508

To demonstrate the effect of our proposed methods on the training time (or compute) and parametrization, we conducted experiments on CIFAR10/100-LT with imbalance factor 100. As shown in Table 32, our proposed methods increase the training time about x1.30 in CIFAR10-LT and x1.09 in CIFAR100-LT, while other factors such as Peak GPU memory usage and the number of parameters are no significant difference.

Broader Impacts

As discussed earlier, our proposed methods have the potential to serve as effective samplers and classifiers, thereby facilitating the development of new state-of-the-art approaches in imbalanced learning. Beyond these contributions, our work may also entail a range of potential societal impacts without harmfulness because our work does not include any potential malicious or unintended uses, fairness considerations, privacy considerations, and security considerations.

NeurIPS Paper Checklist

(1) **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We stated our main claims (i) mixed labels should be balanced and (ii) treating mixed labels as singletons is beneficial in both the abstract and introduction (see 3rd paragraph).

(2) **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We summarize the limitations in Section 7 and discuss them in detail in Section B; the assumptions of the theoretical model are discussed in Section 4.3.

(3) **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Our paper includes theoretical results by providing proofs for them in Section G. All the theorems, formulas, and proofs in the paper are appropriately numbered and cross-referenced. All assumptions are also clearly stated in the statement of any theorems. The proofs appear in the supplemental material (see Section G), and we provide a short proof sketch to provide intuition in Section 4 (More detailed proof sketch is also provided in Section G.) Furthermore, we provided explanations of the intuition behind each statement before and after it in the main paper.

(4) **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our code is publicly available at *link*, and this is stated at Abstract. Furthermore, we provided the detailed experimental setup in Section H.

(5) **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We used only publicly available datasets, such as CIFAR10/100, which are open access. Also, we provide sufficient instructions to faithfully reproduce the main experimental results through *link* and Section H.

(6) **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The full details are provided with the code (*link*) in Section H.

(7) **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We conducted all experiments with different random seeds and reported the mean and standard deviation of top-1 test accuracy (%) on them.

(8) **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide the information on the computer resources in the code (*link*). Furthermore, we conducted experiments about computational cost in Table 32.

(9) **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We preserve anonymity.

(10) **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See the Broader Impacts at the end of Appendix.

(11) **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper poses no such risk

(12) **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We acknowledge the creators or original owners of assets (e.g., code, data, models) used in the paper in our publicly available code and Section H by citing them.

(13) **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: See our code and Section H.

(14) **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

(15) **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: N/A

(16) **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We only used LLM for writing, editing, and formatting purposes.