Neural Collapse in Multi-label Learning with Pick-all-label Loss

Pengyu Li [*1] **Xiao Li** [*1] **Yutong Wang** [1] **Qing Qu** [1]

Abstract

We study deep neural networks for the multi-label classification (M-lab) task through the lens of neural collapse (NC). Previous works have been restricted to the multi-class classification setting and discovered a prevalent NC phenomenon comprising of the following properties for the last-layer features: (i) the variability of features within every class collapses to zero, (ii) the set of feature means form an equi-angular tight frame (ETF), and (iii) the last layer classifiers collapse to the feature mean upon some scaling. We generalize the study to multi-label learning, and prove for the first time that a generalized NC phenomenon holds with the "pick-all-label" formulation, which we term as M-lab NC. While the ETF geometry remains consistent for features with a single label, multi-label scenarios introduce a unique combinatorial aspect we term the "tag-wise average" property, where the means of features with multiple labels are the scaled averages of means for single-label instances. Theoretically, under proper assumptions on the features, we establish that the only global optimizer of the pick-all-label cross-entropy loss satisfy the multi-label NC. In practice, we demonstrate that our findings can lead to better test performance with more efficient training techniques for M-lab learning.

1. Introduction

In recent years, deep learning showed tremendous success in classifying problems (LeCun et al., 2015), thanks in part to its ability to extract salient features from data (Bengio et al., 2013). While the success extends to *multi-label* (M-lab) *classification*, the structures of the learned features in the M-lab regime is less well-understood. This work aims to fill this gap by understanding the geometric structures of features for M-lab learned via deep neural networks and utilize the structure for better training and prediction.

Recently, an intriguing phenomenon has been observed in the terminal phase of training overparameterized deep networks for the task of *multi-class* (M-clf) *classification* in which the last-layer features and classifiers collapse to simple but elegant mathematical structures: all training inputs are mapped to class-specific points in feature space, and the last-layer classifier converges to the dual of class means of the features while attaining the maximum possible margin with a simplex equiangular tight frame (Simplex ETF) structure (Papyan et al., 2020). See the top row of Figure 1 for an illustration. This phenomenon, termed *Neural Collapse* (NC), persists across a variety of different network architectures, datasets, and even the choices of losses (Han et al., 2022; Zhou et al., 2022b;a; Yaras et al., 2022). The NC phenomenon has been widely observed and analyzed theoretically (Papyan et al., 2020; Fang et al., 2021; Zhu et al., 2021; Wang et al., 2023a) in the context of M-clf learning problems. It is applied to understand transfer learning (Galanti et al., 2022b; Li et al., 2022), and robustness (Papyan et al., 2020; Ji et al., 2022), where the line of study has significantly advanced our understanding of representation structures for M-clf using deep networks.

Our contributions. We demonstrate a general version of the NC phenomenon in M-lab, and our study provides new insights into prediction and training for the M-lab problem. Specifically, our contributions can be highlighted as follows.

- **Multi-label neural collapse phenomenon.** We show that the last-layer features and classifier learned via overparameterized deep networks exhibit a more general version of NC which we term it as *multi-label neural collapse* (M-lab NC). Specifically, while features linked to single-label instances retain a Simplex ETF configuration and undergo collapse, the more complex features with higher label counts intriguingly represent a scaled "tag-wise average" of their single-label counterparts; see the bottom row of Figure 1 for an illustration. This new pattern, referred to as *multi-label ETF*, is consistently observed in the training of practical neural networks for M-lab tasks.
- **Global optimality of M-lab NC.** Theoretically, we study the global optimality of M-lab NC based upon a commonly used pick-all-label loss for M-lab learning. By treating the last-layer feature as free optimization variables (Fang et al., 2021; Zhu et al., 2021), we show that

[*]Equal contribution [1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, USA. Correspondence to: Qing Qu <qingqu@umich.edu>.

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

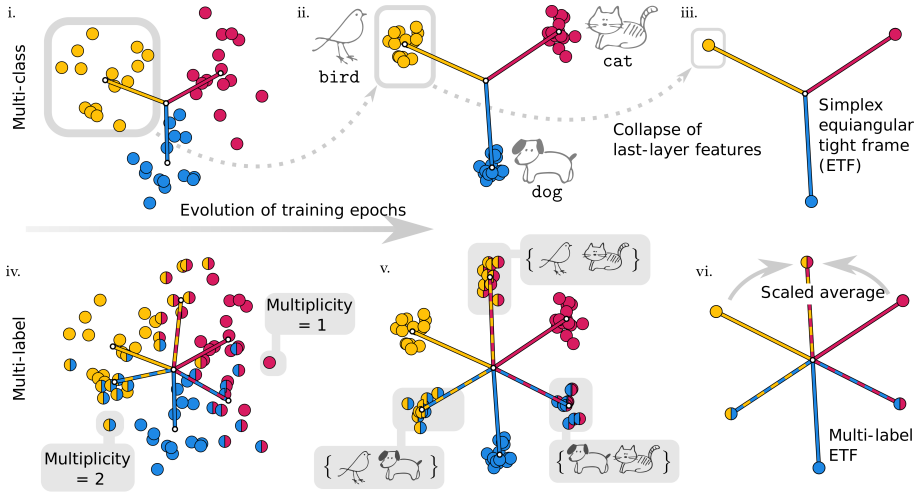


Figure 1. **An illustration of neural collapse for M-clf (top row) vs. M-lab (bottom row) learning**. For illustrative purposes, we consider a simple setting with the number of classes $K = 3$. The individual panels are scatterplots showing the top two singular vectors of the last-layer features $\boldsymbol{H}$ at the beginning (left) and end (right) stages of training. The solid (resp. dashed) line segments represent the mean of the multiplicity $= 1$ (resp. $= 2$) features with the same labels. *Panel i-iii*. As the training progresses, the last-layer features of samples corresponding a single label, e.g., `bird`, collapse tightly around its mean. *Panel iv-vi*. The analogous phenomenon holds in the multi-label setting. *Panel iv*. A training sample has multiplicity $= 1$ (resp. $= 2$) if it has one tag (resp. two tags). *Panel vi*. At the end stage of training, the feature mean of Multiplicity-2 {`bird`, `cat`} is a scaled tag-wise average of feature means of its associated multiplicity-1 samples, i.e., {`bird`} and {`cat`}.

all global solutions exhibit the properties of M-lab NC with benign global landscape. Moreover, we show that multi-label ETF only requires balanced training samples in each class within *the same* multiplicity, and *allows class imbalanced-ness* across different multiplicities.[1]

- **M-lab NC guided prediction and training.** In practice, we show that our findings lead to improved prediction and training for M-lab learning. For prediction, we propose a new one-nearest-neighbor (ONN) approach: for the given test sample, we assign its label to the tag of the nearest neighboring multi-label ETF in the feature space. Compared to the classical one-vs-all (OvA) approach for M-lab learning, ONN is much more efficient with higher test accuracy. For training, by fixing the last later to be ETF and reducing the feature dimension, our experimental results demonstrate that we can sufficiently reduce parameters without compromising overall performance.

Related works on multi-label learning. For M-lab learning, a commonly employed strategy is to decompose a given multi-label problem into several binary classification tasks (Menon et al., 2019). The "one-versus-all" (OvA) method, also known as *binary relevance* (BR), involves splitting the task into several binary classification "subtasks". Each of these subtasks requires training an separate binary classifier, one for each label (Moyano et al., 2018; Brinker et al., 2006). During testing, thresholding is used to convert the (real-valued) outputs of the model (i.e., the logits) into tags.

Although BR is simple to implement and performs well, a notable limitation is the lack of handling of dependency between labels (Cheng et al., 2010). The Pick-All-Label (PAL) approach (Reddi et al., 2019; Menon et al., 2019) can formulates the multi-label classification task in an "all-in-one" manner. Compared to OvA, the PAL approach has the advantage of not needing to train separate classifiers. However, its disadvantage is that during prediction, there is no natural thresholding rule for converting the logits to tags. In this work, we deal with this challenge by proposing the ONN technique as a principled approach, guided by our geometric analysis of M-lab, for performing prediction.

To the best of our knowledge, no work has previously analyzed the geometric structure in multi-label deep learning. Our work closes this gap by providing a generalization of the NC phenomenon to M-lab learning, and furthermore develops efficient prediction and training technique via our theoretical understandings. More discussion on related works for M-lab learning and NC could be found in Appendix A.

Basic notations. Throughout the paper, we use bold lowercase and upper letters, such as $\boldsymbol{a}$ and $\boldsymbol{A}$, to denote vectors and matrices, respectively. Non-bold letters are reserved for scalars. For any matrix $\boldsymbol{A} \in \mathbb{R}^{n_1 \times n_2}$, we write $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \dots & \boldsymbol{a}_{n_2} \end{bmatrix}$, so that $\boldsymbol{a}_i$ ($i \in \{1, \dots, n_2\}$) denotes the i-th column of $\boldsymbol{A}$. Analogously, we use the superscript notation to denote rows, i.e., $(\boldsymbol{a}^j)^\top$ is the j-th row of $\boldsymbol{A}$ for each $j \in \{1, \dots, n_1\}$ with $\boldsymbol{A}^\top = \begin{bmatrix} \boldsymbol{a}^1 & \dots & \boldsymbol{a}^{n_1} \end{bmatrix}$. For an integer $K > 0$, we use $\boldsymbol{I}_K$ to denote an identity

[1] We also empirically demonstrate that M-lab NC occurs even when only multiplicity-1 data are balanced (see Figure 2 for an illustration).

matrix of size $K \times K$, and we use $\mathbf{1}_K$ to denote an all-ones vector of length K.

2. Problem Formulation

We start by reviewing the basic setup for training deep neural networks and later specialize to the problem of M-lab with K number of classes. Given a labelled training instance $(\boldsymbol{x}, \boldsymbol{y})$, the goal is to learn the network parameter $\boldsymbol{\Theta}$ to fit the input $\boldsymbol{x}$ to the corresponding training label $\boldsymbol{y}$ such that

$$\boldsymbol{y} \approx \psi_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \underbrace{\boldsymbol{W}_L}_{\text{linear classifier } \boldsymbol{W}} \cdot \underbrace{\sigma\left(\boldsymbol{W}_{L-1} \cdots \sigma\left(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1\right) + \boldsymbol{b}_{L-1}\right)}_{\text{feature } \boldsymbol{h} = \phi_{\boldsymbol{\theta}}(\boldsymbol{x})} + \boldsymbol{b}_L, \quad (1)$$

where $\boldsymbol{W} = \boldsymbol{W}_L$ represents the last-layer linear classifier and $\boldsymbol{h}(\boldsymbol{x}) = \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,i})$ is a deep hierarchical representation (or feature) of the input $\boldsymbol{x}$. For a L-layer deep network $\psi_{\boldsymbol{\Theta}}(\boldsymbol{x})$, each layer is composed of an affine transformation, followed by a nonlinear activation $\sigma(\cdot)$ (e.g., ReLU) and normalization (e.g., BatchNorm (Ioffe & Szegedy, 2015)).

Notations for multi-label dataset. Let $[K] := \{1, 2, \ldots, K\}$ denote the set of labels. For each $m \in [K]$, let $\binom{[K]}{m} := \{S \subseteq [K] : |S| = m\}$ denote the set of all subsets of $[K]$ with size m. Throughout this work, we consider a fixed multi-label training dataset of the form $\{\boldsymbol{x}_i, \boldsymbol{y}_{S_i}\}_{i=1}^N$, where N is the size of the training set and S_i is a nonempty proper subset of the labels. For instance, $S_i = \{\texttt{cat}\}$ and $S_{i'} = \{\texttt{dog}, \texttt{bird}\}$. Each label $\boldsymbol{y}_{S_i} \in \mathbb{R}^K$ is a *multi-hot-encoding* vector:

$$j\text{-th entry of } \boldsymbol{y}_{S_i} = \begin{cases} 1 & : j \in S_i \\ 0 & : \text{otherwise.} \end{cases} \quad (2)$$

The *Multiplicity* of a training sample $(\boldsymbol{x}_i, \boldsymbol{y}_{S_i})$ is defined as the cardinality of S_i, i.e., the number of labels or tags that is related to $\boldsymbol{x}_i$. We refer to a feature learned for the sample $(\boldsymbol{x}_i, \boldsymbol{y}_{S_i})$ as the Multiplicity-m feature if $|S_i| = m$. As such, with the abuse of notation, we also refer to such $\boldsymbol{x}_i$ as a Multiplicity-m sample and such $\boldsymbol{y}_{S_i}$ as a Multiplicity-m label, respectively. Note that a multiplicity-m label is a multi-hot label that can be decomposed as a summation of one-hot multiplicity-1 labels. For example $S_{i'} = \{\texttt{dog}, \texttt{bird}\}$ has two tags $S_{j'} = \{\texttt{dog}\}$ and $S_{k'} = \{\texttt{bird}\}$, and then the corresponding 2-hot label can be decomposed into the associated 1-hot labels of $S_{j'} = \{\texttt{dog}\}$ and $S_{k'} = \{\texttt{bird}\}$. In this work, we show the relationship of labels can be generalized to study the relationship of the associated features trained via deep networks through M-lab NC.

The Multiplicity-m feature matrix $\boldsymbol{H}_m$ is column-wise comprised of a collection of Multiplicity-m feature vectors. Moreover, we use $M := \max_{i \in [N]} |S_i|$ to denote the largest multiplicity in the training set. To distinguish imbalanced class samples between Multiplicities, for each $m \in [M]$, we use $n_m := |\{i \in [N] : |S_i| = m\}|$ to denote the number of samples in each class of a multiplicity order m (or Multiplicity m). Note that $M \in \{1, \ldots, K-1\}$ in general, and a M-lab problem reduces to M-clf when $M = 1$.

The "pick-all-labels" loss. Since M-lab is a generalization of M-clf, recent work (Menon et al., 2019) studied various ways of converting a M-clf loss into a M-lab loss, a process referred to as *reduction*.[2] In this work, we analyze the *pick-all-labels* (PAL) method of reducing the cross-entropy (CE) loss to a M-lab loss, which is the *default* option implemented by `torch.nn.CrossEntropyLoss` from the deep learning library PyTorch (Paszke et al., 2019). The benefit of PAL approach is that the difficult M-lab problem can be approached using insights from M-clf learning using well-understood losses such as the CE loss, which is one of the most commonly used loss functions in classification:

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}_k) := -\log\left(\exp(z_k) / \textstyle\sum_{\ell=1}^K \exp(z_\ell)\right).$$

where $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{h}$ is called the logits, and $\boldsymbol{y}_k$ is the one-hot encoding for the k-th class. To convert the CE loss into a M-clf loss via the PAL method, for any given label set S, consider decomposing a multi-hot label $\boldsymbol{y}_S$ as a summation of one-hot labels: $\boldsymbol{y}_S = \sum_{k \in S} \boldsymbol{y}_k$. Thus, we can define the *pick-all-labels cross-entropy* (PAL-CE) loss as

$$\underline{\mathcal{L}}_{\texttt{PAL}-\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}_S) := \textstyle\sum_{k \in S} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}_k).$$

In this work, we focus exclusively on the CE loss under the PAL framework, we simply write $\underline{\mathcal{L}}_{\texttt{PAL}}$ to denote $\underline{\mathcal{L}}_{\texttt{PAL}-\mathrm{CE}}$. However, by drawing inspiration from recent research (Zhou et al., 2022b), it should be noted that under the PAL framework, the phenomenon of M-lab NC can be generalized beyond cross-entropy to encompass a variety of other loss functions used for M-clf learning, such as mean squared error (MSE), label smoothing (LS),[3] focal loss (FL),[4] and potentially a class of Fenchel-Young Losses that unifies many well-known losses (Blondel et al., 2020).

Putting it all together, training deep neural networks for M-lab learning can be stated as follows:

$$\min_{\boldsymbol{\Theta}} \tfrac{1}{N} \textstyle\sum_{i=1}^N \underline{\mathcal{L}}_{\texttt{PAL}}(\boldsymbol{W}\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + \boldsymbol{b}, \boldsymbol{y}_{S_i}) + \lambda \|\boldsymbol{\Theta}\|_F^2, \quad (3)$$

where $\boldsymbol{\Theta} = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta}\}$ denote all parameters and $\lambda > 0$ controls the strength of weight decay. Here, weight decay prevents the norm of the linear classifier and the feature matrix goes to infinity or 0.

Optimization under the unconstrained feature model (UFM). Analyzing the nonconvex loss (3) can be notoriously difficult due to the highly non-linear characteristic of the deep network $\phi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$. In this work, we simplify the study by treating the feature $\boldsymbol{h}_i = \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ of each input $\boldsymbol{x}_i$ as a *free* optimization variable. Analysis of NC under UFM

[2] "Reduction" refers to reformulating M-lab problems in the simpler framework of M-clf problems.

[3] The loss replaces hard targets in CE with smoothed ones to achieve better calibration and generalization (Szegedy et al., 2016).

[4] The loss adjusts its focus to less on the well-classified samples, enhances calibration, and establishes a curriculum learning framework (Lin et al., 2017; Mukhoti et al., 2020; Smith, 2022).

has been extensively studied in recent works (Zhu et al., 2021; Fang et al., 2021; Ji et al., 2022; Yaras et al., 2022; Mixon et al., 2022; Zhou et al., 2022a; Tirer & Bruna, 2022), the motivation behind the UFM is the fact that modern networks are highly overparameterized and they are universal approximators (Cybenko, 1989; Zhang et al., 2021). More specifically, we study the following problem.

Definition 2.1 (Nonconvex Training Loss under UFM)**.** Let $\boldsymbol{Y} = [\boldsymbol{y}_{S_1} \cdots \boldsymbol{y}_{S_N}] \in \mathbb{R}^{K\times N}$ be the multi-hot encoding matrix whose i-th column is given by the multi-hot vector $\boldsymbol{y}_{S_i} \in \mathbb{R}^K$. We consider

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}} f(\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}) \;:=\; g(\boldsymbol{W}\boldsymbol{H}+\boldsymbol{b},\boldsymbol{Y}) + \lambda_W \|\boldsymbol{W}\|_F^2 + \lambda_H \|\boldsymbol{H}\|_F^2 + \lambda_b \|\boldsymbol{b}\|_2^2 \tag{4}$$

with the penalty $\lambda_W, \lambda_H, \lambda_b > 0$.

Here, the linear classifier $\boldsymbol{W} \in \mathbb{R}^{K\times d}$, the features $\boldsymbol{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_N] \in \mathbb{R}^{d\times N}$, and the bias $\boldsymbol{b} \in \mathbb{R}^K$ are all unconstrained optimization variables, and we refer to the columns of $\boldsymbol{H}$, denoted $\boldsymbol{h}_i$, as the *unconstrained last layer features* of the input samples $\boldsymbol{x}_i$. Additionally, the function $g(\cdot)$ is the PAL loss, denoted by

$$g(\boldsymbol{W}\boldsymbol{H}+\boldsymbol{b},\boldsymbol{Y}) := \tfrac{1}{N}\mathcal{L}_{\texttt{PAL}}(\boldsymbol{W}\boldsymbol{H}+\boldsymbol{b},\boldsymbol{Y}) := \tfrac{1}{N}\sum_{i=1}^N \mathcal{L}_{\texttt{PAL}}(\boldsymbol{W}\boldsymbol{h}_i+\boldsymbol{b},\boldsymbol{y}_{S_i}).$$

Although the objective function is seemingly a simple extension of M-clf case, our work shows that the global optimizers of Problem (4) for M-lab learning substantially differs from that of the M-clf that we present in the following.

3. Main Results

In this section, we show that the global minimizers of Problem (4) exhibit a more generic structure than the vanilla NC in M-clf (see Figure 1), where higher multiplicity features are formed by a scaled tag-wise average of associated Multiplicity-1 features that we introduce in detail below. Theoretically, we rigorously analyze the global geometry of the optimizer of Problem (4) and its nonconvex optimization landscape, and present our main results in Theorem 3.1.

3.1. Multi-label Neural Collapse (M-lab NC)

We assume that the training data is balanced with respect to Multiplicity-1 while high-order multiplicity is imbalanced or even has missing classes. Through empirical investigation, we discover that when a deep network is trained up to the terminal phase using the objective function (3), it exhibits the following characteristics, which we collectively term as "multi-label neural collapse" (M-lab NC):

1. **Variability collapse:** The within-class variability of last-layer features across different multiplicities and different classes all collapses to zero. In other words, the individual features of each class of each multiplicity concentrate to their respective class means.
2. $(*)$ **Convergence to self-duality of multiplicity-1 features $\boldsymbol{H}_1$:** The rows of the last-layer linear classifier $\boldsymbol{W}$ and the class means of Multiplicity-1 feature $\boldsymbol{H}$ are collinear, i.e., $\boldsymbol{h}_i^\star \propto \boldsymbol{w}^{\star k}$ when the label set $S_i = \{k\}$ is a singleton set.
3. $(*)$ **Convergence to the M-lab ETF:** Multiplicity-1 features $\boldsymbol{H}_1 := \left\{\boldsymbol{h}_i^\star | i : |S_i| = 1\right\}$ form a Simplex Equiangular Tight Frame, similar to the M-clf setting (Papyan et al., 2020; Fang et al., 2021; Zhu et al., 2021). Moreover, for any higher multiplicity $m > 1$, *the average feature means for classes with label count m are a scaled, tag-wise aggregation of the corresponding single-label ($Multiplicity - 1$) feature means across the relevant label set.* In other words, $\boldsymbol{h}_i^\star \propto \sum_{k\in S_i} \boldsymbol{w}^{\star k}$ (see the bottom line of Figure 1). This is true regardless of class imbalance between multiplicities.

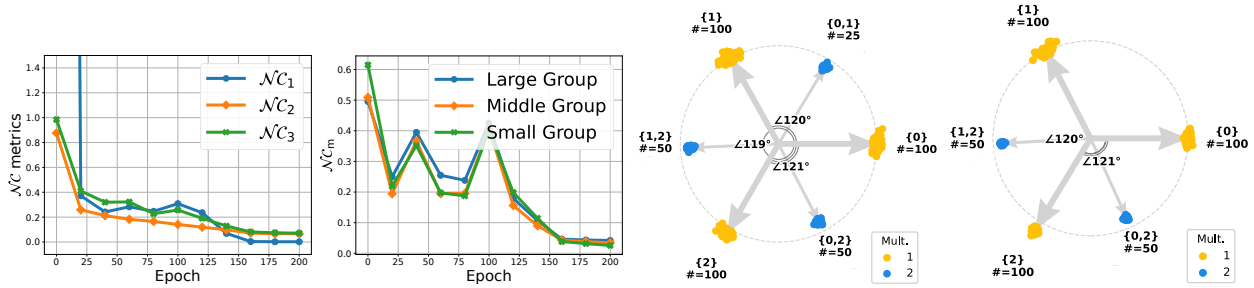
Remarks. The M-lab NC can be viewed as a more general version of the vanilla NC in M-clf (Papyan et al., 2020), where we mark the difference from the vanilla NC above by a "$(*)$". The M-lab ETF implies that, in the pick-all-labels approach to multi-label classification, deep networks learn discriminant and informative features for Multiplicity-1 subset of the training data, and use them to construct higher multiplicity features as the tag-wise average of associated Multiplicity-1 features. To quantify the collapse of high multiplicity NC, we introduce a new measure $\mathcal{NC}_m$ in Section 4 and demonstrate that it collapses for practical neural networks during the terminal phase of training. This result is intuitive: since the multi-hot label vector can be decomposed into the sum of its tag-wise one-hot vectors, the corresponding learned features may exhibit a similar scaled tag-average phenomenon.

Moreover, in the case of data imbalancend-ness, we find that the M-lab NC holds as long as the training samples within the same multiplicity are required to be class balanced, and the number of samples between multiplicities does *not* need to be balanced. This can be later confirmed by our theory in Section 3.2. For example, the M-lab NC still holds if there are more or less training samples for the category $(\texttt{ant}, \texttt{bee})$(Multiplicity-2) than that of $(\texttt{cat}, \texttt{dog}, \texttt{elk})$ (Multiplicity-3).

3.2. Global Optimality of M-lab NC

For M-lab, we show that the M-lab NC is the only global solution to the nonconvex problem in Definition 2.1. We consider the setting that the training data may exhibit imbalanced-ness between different multiplicities while maintaining class-balancedness within each multiplicity.

Theorem 3.1 (**Global Optimality of M-lab NC**)**.** *In the setting of Definition 2.1, assume the feature dimension is no smaller than the number of classes minues one, i.e., $d \geq K - 1$, and assume the training are balanced within each multiplicity as we discussed above. Then any global*



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

Figure 2. **M-lab NC holds with imbalanced data in higher multiplicities.** (a) and (b) plot metrics that measures M-lab NC on M-lab Cifar10; (c) and (d) visualize learned features on M-lab MNIST, where one multiplicity-2 class is missing in the set up which results in the reduced M-lab NC geometry. As we observe, the ETF structure for Multiplicity-1 still holds. More experimental details are deferred to Section 4.

optimizer $\boldsymbol{W}^\star, \boldsymbol{H}^\star, \boldsymbol{b}^\star$ of the optimization Problem (4) satisfies:

$$w^\star := \|\boldsymbol{w}^{\star 1}\|_2 = \cdots = \|\boldsymbol{w}^{\star K}\|_2, \quad \text{and} \quad \boldsymbol{b}^\star = b^\star \mathbf{1}, \quad (5)$$

where either $b^\star = 0$ or $\lambda_{\boldsymbol{b}} = 0$. Moreover, the global minimizer $\boldsymbol{W}^\star, \boldsymbol{H}^\star, \boldsymbol{b}^\star$ satisfies the M-lab NC properties introduced in Section 3.1, in the sense that

- *The linear classifier matrix $\boldsymbol{W}^{\star\top} \in \mathbb{R}^{d\times K}$ forms a K-simplex ETF up to scaling and rotation, i.e., for any $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ s.t. $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_d$, the rotated and normalized matrix $\boldsymbol{M} := \frac{1}{w^\star}\boldsymbol{U}\boldsymbol{W}^{\star\top}$ satisfies*

$$\boldsymbol{M}^\top\boldsymbol{M} = \frac{K}{K-1}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right). \quad (6)$$

- Tag-wise average property. *For each feature $\boldsymbol{h}_i^\star$ (i.e., the i-th column $\boldsymbol{h}_i^\star$ of $\boldsymbol{H}^\star$) with $i \in [N]$, there exist unique positive real numbers $C_1, C_2, \ldots, C_M > 0$ such that the following holds:*

Multiplicity $= 1$ Case:

$$\boldsymbol{h}_i^\star = C_1\boldsymbol{w}^{\star k} \qquad \text{when } S_i = \{k\},\ k \in [K], \quad (7)$$

Multiplicity > 1 Case:

$$\boldsymbol{h}_i^\star = C_m \sum_{k\in S_i} \boldsymbol{w}^{\star k} \quad \text{when } |S_i| = m,\ 1 < m \le M. \quad (8)$$

Moreover, the function $f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ in Problem (4) is a strict saddle function (Ge et al., 2015; Sun et al., 2015; Zhang et al., 2020b) with no spurious local minimum.

We discuss the high-level ideas of the proof in Section 3.3. The detailed proof of our results is deferred to Appendix C and Appendix D. Next, we delve into the implications of our findings from various perspectives.

The global solutions of Problem (4) satisfy M-lab NC. Under the assumption of UFM, our findings imply that every global solution of the loss function of Problem (4) exhibits the M-lab NC that we presented in Section 3.1. First, feature variability within each class and multiplicity can be deduced from Equations (7) and (8). This occurs because all features of the designated class and multiplicity align with the (tag-wise average of) linear classifiers, meaning they are equal to their feature means with no variability. Second, the convergence of feature means to the M-lab ETF can be observed from Equations (6), (7), and (8). For Multiplicity-1 features $\boldsymbol{H}_1^\star$, Equation (7) implies that the feature mean $\overline{\boldsymbol{H}}_1^\star$ converges to $\boldsymbol{W}^\star$; this, coupled with Equation (6), implies that the feature means $\overline{\boldsymbol{H}}_1^\star$ of Multiplicity-1 form a simplex ETF. Moreover, the structure of tag-wise average in Equation (8) implies the M-lab ETF for feature means of high multiplicity samples. Finally, the convergence of Multiplicity-1 features towards self-duality can be deduced from Equation (7).

Data imbalanced-ness in M-lab learning. Due to the scarcity of higher multiplicity labels in the training set, in practice the imbalance of training data samples could be a more serious issue in M-lab than M-clf. It should be noted that there are two types of data imbalanced-ness: (*i*) the imbalanced-ness between classes *within* each multiplicity, and (*ii*) the imbalanced-ness of classes *among* different multiplicities. Interestingly, as long as Multiplicity-1 training samples remain balanced between classes, our results in Figure 2 and Figure 4 suggest that the M-lab NC holds *regardless* of both within and among multiplicity imbalanced-ness in higher multiplicity. Given that achieving balance in Multiplicity-1 sample data is relatively easy, this implies that our result captures a common phenomenon in M-lab learning. However, if classes of Multiplicity-1 are imbalanced, we suspect a more general minority collapse phenomenon would happen (Fang et al., 2021; Thrampoulidis et al., 2022), which is worth of further investigation.

Improving M-lab prediction & training via M-lab NC. Guided by the feature collapse phenomenon of M-lab NC, we show that we can improve the prediction accuracy and training efficiency in Section 4.2. For prediction, encoding could use an one-nearest-neighbor (ONN) approach to classify new data based on the nearest feature mean in the feature space. Empirical verification confirms that ONN encoding is more efficient and yields superior testing accuracy

compared to OvA, as illustrated in Table 1. For training, as shown in Table 2, we can achieve parameter efficient training for M-lab by fixing the last layer classifier as simplex ETF and reducing the feature dimension d to K.

Tag-wise average coefficients for M-lab ETF with high multiplicity. The features of high multiplicity are scaled tag-wise average of Multiplicity-1 features, and these coefficients are *simple and structured* as shown in Equation (8). As illustrated in Figure 1 (i.e., $K = 3$, $M = 2$), the feature $\boldsymbol{h}_i^\star$ of Multiplicity-m associated with class-index S_i can be viewed as a *tag-wise average* of Multiplicity-1 features in the index set S_i. Specifically, the high multiplicity coefficients $\{C_m\}_{m=1}^M$ in Equation (8), which are shared across all features of the same multiplicity, could be expressed as

$$C_m = \frac{K-1}{\|\boldsymbol{W}\|_F^2} \log(\frac{K-m}{m} c_{1,m}), \quad \forall m$$

where $\{c_{1,m}\}_{m=1}^M$ exist.[5]

3.3. Proof Ideas of Theorem 3.1 and Comparison

Finally, we briefly outline our proofs for the global optimality in Theorem 3.1 as follows: essentially, our proof method first breaks down the $g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}, \boldsymbol{Y})$ component of the objective function of Problem (4) into numerous subproblems $g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}, \boldsymbol{Y}_m)$, categorized by different multiplicity. We determine lower bounds for each g_m and establish the conditions for equality attainment for each multiplicity level. Subsequently, we confirm that equality for these sets of lower bounds of different m values can be attained simultaneously, thus constructing a global optimizer where the overall global objective of (4) is reached. We demonstrate that all optimizers can be recovered using this approach. As a result, our generalized proof implies M-clf NC with only single-multiplicity data.

Although our work is inspired by the recent work (Zhu et al., 2021), it should be noted that our main results as well as the proving techniques used to establish them significantly differ from that of (Zhu et al., 2021). For M-lab, the derivation of optimality conditions is particularly challenging due to the combinatorial complexity of imbalanced features with higher label counts and how they interact with a single linear classifier. We elaborate on this in the following.

- We incorporate all multiplicity samples by calculating the gradient of the PAL-CE loss function to obtain the initial lower bound. The tightness condition of such bound uncovers M-lab learning's unique "in-group and out-group" property hidden behind the combinatorial structure of high multiplicity features. Comparatively, (Zhu et al., 2021) relied on Jensen's inequality and concavity of log function which falls short under the present of high-multiplicity samples. More details can be found in Lemma C.8.
- We decoupled the interplay between linear classifier across various multiplicity features by decomposing the loss into different components based on feature multiplicities. Through the decomposition, we then showed that the equality condition for each components can be achieved simultaneously. More details are provided in Lemma C.2.
- We further unveil that the higher multiplicity features converge to the "scaled tag-wise average" of its associate tag feature means (Lemma C.3). This requires three new supporting lemmas derived from probabilistic (Lemma C.6) and matrix theory perspective (Lemma C.5, C.7), which is unique in M-lab learning.

[5]They satisfy a set of nonlinear equations (Appendix C).

4. Experiments

In this section, we first conduct a series of experiments to demonstrate and analyze the M-lab NC on different practical deep networks with various multi-label datasets. Second, we show that the geometric structure of M-lab NC could efficiently guide M-lab learning in both testing and training stage for better performance.

The datasets used in our experiment are real-world M-lab SVHN (Netzer et al., 2011), along with synthetically generated M-lab MNIST (LeCun et al., 2010) and M-lab Cifar10 (Krizhevsky et al., 2009). The detail dataset description, generation, and visualization along with experimental setups could be found in Appendix B.

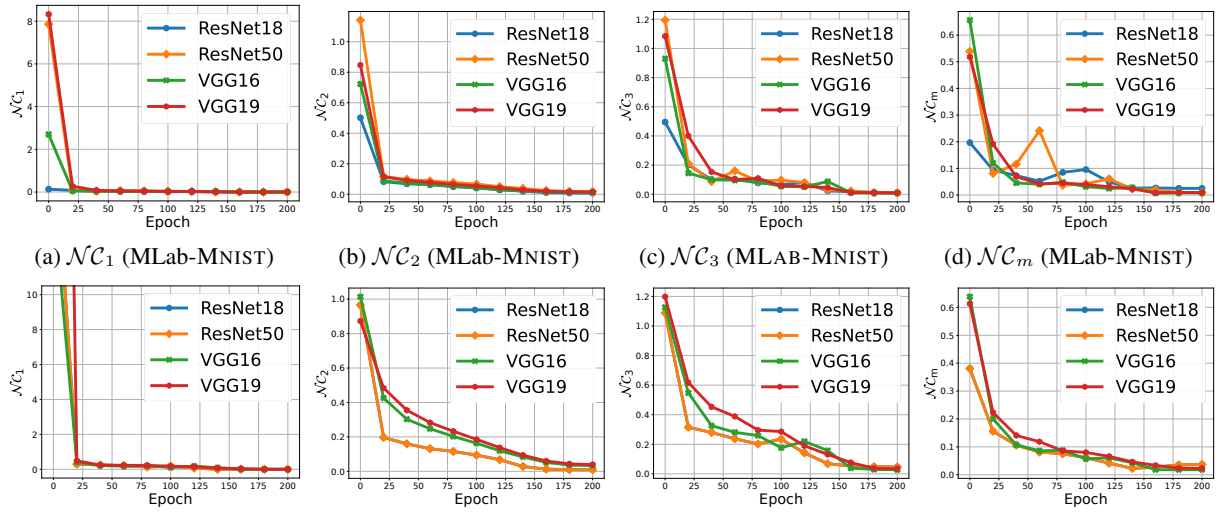
4.1. Verification of M-lab NC

Experimental demonstration of M-lab NC on practical deep networks. When the training data (Appendix B.1) are balanced within multiplicities, Figure 3 shows that all practical deep networks exhibit M-lab NC during the terminal phase of training as implied by our theory. To show this, we introduce new metrics to measure M-lab NC on the last-layer features and classifiers of deep networks.

Based on theoretical results in Section 3.1, we use the original metrics $\mathcal{NC}_1$ (measuring the within-class variability collapse), $\mathcal{NC}_2$ (measuring convergence of learned classifier and feature class means to simplex ETF), and $\mathcal{NC}_3$ (measuring the convergence to self-duality) introduced in (Papyan et al., 2020) to measure M-lab NC on Multiplicity-1 features $\boldsymbol{H}_1$ and classifier $\boldsymbol{W}$. Additionally, we also use the $\mathcal{NC}_1$ metric to measure variability collapse on high multiplicity features $\boldsymbol{H}_m$ $(m > 1)$. Finally, to measure M-lab ETF (the *tag-wise average* property) on Multiplicity-2 features,[6] we propose a new angle metric $\mathcal{NC}_m$, which is defined as:

$$\mathcal{NC}_m = \frac{\text{Avg.}(\{geo_\angle(\overline{\boldsymbol{h}}_i,\ \overline{\boldsymbol{h}}_j + \overline{\boldsymbol{h}}_\ell) : (i,j,\ell) \in \mathcal{F}_1\})}{\text{Avg.}(\{geo_\angle(\overline{\boldsymbol{h}}_{i'},\ \overline{\boldsymbol{h}}_{j'} + \overline{\boldsymbol{h}}_{\ell'}) : (i',j',\ell') \in \mathcal{F}_2\}},$$

[6]This is because our dataset only contains labels up to Multiplicity-2. The $\mathcal{NC}_m$ could be easily extended to capture scaled average for other higher multiplities.



(a) $\mathcal{NC}_1$ (MLab-MNIST) (b) $\mathcal{NC}_2$ (MLab-MNIST) (c) $\mathcal{NC}_3$ (MLAB-MNIST) (d) $\mathcal{NC}_m$ (MLab-MNIST)

(e) $\mathcal{NC}_1$ (MLab-Cifar10) (f) $\mathcal{NC}_2$ (MLab-Cifar10) (g) $\mathcal{NC}_3$ (MLab-Cifar10) (h) $\mathcal{NC}_m$ (MLab-Cifar10)

Figure 3. **Prevalence of M-lab NC across different network architectures** on M-lab MNIST (top) and M-lab Cifar10 (bottom). From the left to the right, the plots show the four metrics, $\mathcal{NC}_1, \mathcal{NC}_2, \mathcal{NC}_3$, and $\mathcal{NC}_m$, for measuring M-lab NC. More details about dataset and training setups could be found in Appendix B.1.

with the sets

$$\mathcal{F}_1 = \{i, j, \ell \mid |S_i| = 2,\ |S_j| = |S_\ell| = 1,\ S_i = S_j \cup S_\ell\},$$
$$\mathcal{F}_2 = \{i, j, \ell \mid |S_i| = 2,\ |S_j| = |S_\ell| = 1\}.$$

Here, $geo\angle$ represents the geometric angle between two vectors, and $\overline{\boldsymbol{h}}_i$ is the mean of all features in the label set S_i. Intuitively, our $\mathcal{NC}_m$ measures the angle between features means of different label sets or classes. The numerator calculates the average angle difference between multiplicity-2 features means and the sum of their multiplicity-1 component features means. While the denominator serves as a normalization factor that is the average of all existing pairs regardless of the relationship.[7] As training progresses, the numerator will converge to 0, while the denominator becomes larger demonstrating the angle collapsing.

As shown in Figure 3 and Figure 4, practical networks do exhibit M-lab NC, and such a phenomenon is prevalent across network architectures and datasets. Specifically, the four metrics, evaluated on four different network architectures and two different datasets, all converge to zero as the training progresses toward the terminal phase.

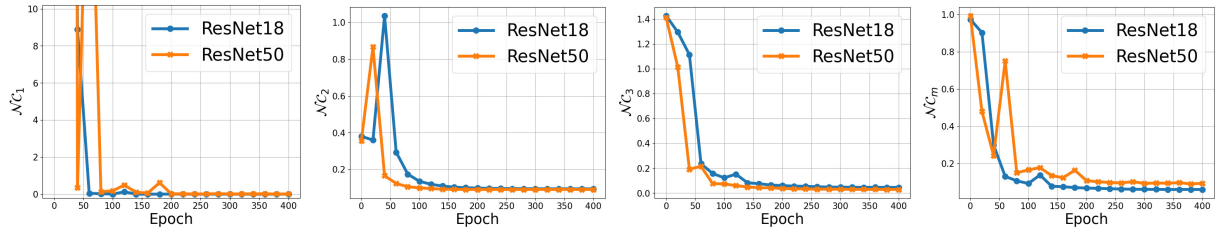
M-lab NC holds despite of class imbalanced-ness in high order multiplicity labels. Moreover, our experiments imply that maintaining balance in single-label training samples ensures the persistence of M-lab NC, even amidst imbalance in higher-order label multiplicities, across both synthetic and real-world data sets. To verify this, we create imbalanced M-lab cifar10 and MNIST datasets, and real-world M-lab SVHN dataset, with more details in Appendix B.

- **Experimental results on imbalanced M-lab Cifar10 dataset.** We run a ResNet18 model with these datasets (Appendix B.2) and report the metrics of measuring M-lab NC in Figure 2 (a) (b). We can observe that not only $\mathcal{NC}_1$ to $\mathcal{NC}_3$ collapse to zero, but the $\mathcal{NC}_m$ metric is also converging zero for all 3 groups of different sizes.
- **Experimental results on imbalanced M-lab MNIST dataset.** For Figure 2 (c) (d) on M-lab MNIST (Appendix B.2), we can see from the visualization of the features vectors that the scaled average property still holds despite a missing class in higher multiplicity. Here, we train a simple convolution plus multi-layer perceptron model. This implies that M-lab NC holds even under data imbalanced-ness in high order multiplicity.
- **Experimental results on imbalanced M-lab SVHN dataset.** In addition, we demonstrate that M-lab NC happens independently of higher multiplicity data imbalanced-ness on real-world M-lab SVHN dataset (Appendix B.3). We evaluated the behavior of NC metrics on this dataset as illustrated in Figure 4, affirming the continued validity of our analysis in real-world settings.

4.2. Practical Implications for M-lab Learning

Finally, we show that our findings lead to improved prediction and training for M-lab learning. For prediction, Our ONN encoding approach attains greater accuracy than the OvA method with improved efficiency, eliminating the need for extra classifier training, as shown in Table 1. For training, our theory supports reducing feature dimension and

[7] For example, if we have 4 total classes for multiplicity-1 samples, they corresponds to 4 features means and hence 6 different sums if we randomly pick 2 features means to sum up. Multiplicity-2 then has $\binom{4}{2} = 6$ features means, there are then 36 possible angles to calculate, we averaged these 36 angles as the denominator.



(a) $\mathcal{NC}_1$ (b) $\mathcal{NC}_2$ (c) $\mathcal{NC}_3$ (d) $\mathcal{NC}_m$

Figure 4. **Prevalence of M-lab NC on the M-lab SVHN dataset**. We train ResNets models on the M-lab SVHN dataset (Netzer et al., 2011) for 400 epochs and report $\mathcal{NC}_1, \mathcal{NC}_2, \mathcal{NC}_3$, and $\mathcal{NC}_m$, for measuring M-lab NC, respectively. See Appendix B.3 for more details.

Dataset	c10-L		c10-M		c10-S		SVHN	
Encoding Mechanism	OvA	**ONN**	OvA	**ONN**	OvA	**ONN**	OvA	**ONN**
Test Metrics (F1-score / SubsetZero-one Accuracy)								
Overall	0.896/0.787	**0.899/0.805**	0.885/0.778	**0.888/0.794**	0.857/0.740	**0.863/0.760**	0.937/0.831	**0.942/0.837**
Mul-1	0.893/0.828	**0.894/0.829**	0.890/0.832	0.890/**0.833**	0.865/0.812	**0.867/0.815**	0.905/0.761	**0.928/0.836**
Mul-2	0.898/0.775	**0.901/0.797**	0.883/0.754	**0.887/0.777**	0.851/0.691	**0.861/0.723**	**0.953/0.871**	0.949/0.842
Mul-3	*						0.926/0.771	**0.935/0.814**
Computational Complexity								
FLOPs(B)	2.46	**0.0307**	2.00	**0.0249**	1.54	**0.0192**	1.068	**0.0135**

Table 1. **Test accuracy and computational complexity comparison between ONN (ours) and OvA approaches across different multiplicity imbalance-ness for both real and synthetic datasets.** Both approaches are trained with fixed ETF classifier. Models are trained using ResNet18 structure and reported accuracy are the average over 3 models with random initialization. We report both F1-score and subset zero-one accuracy that only count for perfect predictions (Dembczyński et al., 2010).

maintaining a fixed classifier structure without sacrificing training accuracy as shown in Table 2. Additionally, the detail descriptions of training datasets with experimental setups can be found in Appendix B.4

Implication I: M-lab NC guided methods for improved test performance. As discussed in Section 1, the classical OvA and PAL methods have several fundamental limitations, specially, when it comes to how to convert outputs of the model into tags (i.e., binary vectors). In this part, we show that we can improve the PAL based method by our findings. Specifically, our proposed method and comparison baseline are the following.

- ***Proposed*** **"one-nearest-neighbor" (ONN) encoding method**: supported by the M-lab NC, features within each class collapse to their means across all multiplicities. Utilizing this, encoding new testing data into binary vectors is simplified: a one-nearest-neighbor calculation is performed between the testing data's features and all class means.
- ***Classical*** **"one-versus-all" (OvA) encoding method**: divides the task into multiple binary classification subtasks, where each needs to train an individual binary classifier for every label based on pre-trained features. During testing, thresholding is used to convert the outputs of the model (i.e., the logits) into tags.

We compared our ONN method with OvA across three synthetic M-lab Cifar-10 datasets and one real-world dataset with different data imbalanced-ness. Two types of test accuracies, F1-score and subset zero-one, are reported in Table 1. The F1-score provides a more balanced performance measure, whereas subset zero-one accuracy only considers perfect matches to the ground truth, disregarding partially correct predictions. The detailed setup of training and dataset could be find in Appendix B.4. As we observe from Table 1, our ONN method uniformly outperforms OvA in overall accuracy, with higher accuracy especially in higher multiplicities. Simultaneously, our ONN eliminates the necessity of training multiple binary classifiers unlike OvA, leading to substantially lower computational complexity for predictions, quantified in billions of FLOPs. Remarkably, even when dealing with the class-imbalanced real M-lab SVHN dataset, our ONN method consistently achieves an overall higher accuracy than OvA.

Implication II: M-lab NC guided parameter-efficient training. With the knowledge of M-lab NC in hand, we can make direct modifications to the model architecture to achieve parameter savings without compromising performance for M-lab classification. Specifically, parameter saving could come from two folds: (i) given the existence of $\mathcal{NC}$ in the multi-label case with $d \geq K$, we can reduce the dimensionality of the penultimate features to match the number of labels (i.e., we set $d = K$); (ii) recognizing that the final linear classifier will converge to a simplex ETF as the training converges, we can initialize the weight matrix of the classifier as a simplex ETF from the start and refrain

Dataset / Arch.	ResNet18		ResNet50		VGG16		VGG19	
	Learned	ETF	Learned	ETF	Learned	ETF	Learned	ETF
Test IoU (%)								
MLab-MNIST	99.5	99.4	99.4	99.4	99.5	99.5	99.5	99.5
MLab-Cifar10	87.7	87.7	88.9	88.6	86.9	87.4	88.7	87.0
Percentage of parameter saved (%)								
MLab-MNIST	0	**20.7**	0	**4.5**	0	**15.8**	0	**11.6**
MLab-Cifar10								

Table 2. **Comparison of the performances and parameter efficiency between learned and fixed ETF classifier.** When counting parameters, we consider all parameters that require gradient calculation during back-propagation.

from updating it during training. By doing so, our experimental results in Table 2 demonstrate that we can achieve parameter reductions of up to 20% without sacrificing the performance of the model.[8]

5. Conclusion

In this study, we extensively analyzed the NC phenomenon in M-lab learning. In theory, our results establish that M-lab ETFs are the only global minimizers of the PAL loss function, incorporating weight decay and bias. In practice, these findings hold significant implications for improving the performance and efficiency of M-lab tasks in both testing and training stages. Our work also fosters future directions in multi-label learning such as dealing with data imbalanced-ness, investigating other losses (Han et al., 2022; Zhou et al., 2022a) or designing a better training loss are all worthy directions of pursuit.

- **Dealing with data imbalanced-ness.** As many multi-label datasets are imbalanced, another important direction is to investigate the more challenging cases where the Multiplicity-1 training data are imbalanced. We suspect a more general minority collapse phenomenon would happen (Fang et al., 2021; Thrampoulidis et al., 2022; Zhai et al., 2023). A promising approach might be employing the *Simplex-Encoded-Labels Interpolation (SELI) framework*, which is related to the singular value decomposition of the simplex-encoded label matrix (Thrampoulidis et al., 2022). Nonetheless, we conjecture that the scaled average property will still hold between higher multiplicity features and their multiplicity-1 features. On the other hand, when Multiplicity-1 training data are imbalanced, it is also worth studying creating a balanced dataset through data augmentation by leveraging recent advances in diffusion models (Ho et al., 2020; Trabucco et al., 2023; Zhang et al., 2023).

- **Designing better training loss.** Prior research has underscored the significance of mitigating within-class variability collapse to improve the transferability of learned models (Li et al., 2022). In the context of M-lab problems, the principle of Maximal Coding Rate Reduction (MCR^2) has been designed and effectively employed to foster feature diversity and discrimination, thereby preventing collapse (Yu et al., 2020; Chan et al., 2022). We believe that our M-lab NC could offer insights into the development of analogous loss functions for M-lab learning, with the goal of promoting diversity in features.

More discussions on other future directions with preliminary experimental results can be found in Appendix A.

Code Availability

Our code is publicly available at `https://github.com/Heimine/NC_MLab/`.

Acknowledgements

The authors acknowledge support from NSF CAREER CCF-2143904, NSF CCF-2212066, NSF CCF-2212326, NSF IIS 2312842, ONR N00014-22-1-2529, an Amazon AWS AI Award, and a gift grant from KLA. YW also acknowledges the support from the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program. Results presented in this paper were obtained using CloudBank, which is supported by the NSF under Award #1925001.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Behnia, T., Kini, G. R., Vakilian, V., and Thrampoulidis, C. On the implicit geometry of cross-entropy parameterizations for label-imbalanced data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10815–10838. PMLR, 2023.

Bengio, Y., Courville, A., and Vincent, P. Representation

[8] We use intersection over union (IoU) to measure model performances, in M-lab, we define $\text{IoU}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = ||\boldsymbol{y}||_0 \cdot (\hat{\boldsymbol{y}}^\top \boldsymbol{y}) \in [0, 1]$. Here, the ground truth $\boldsymbol{y}$ represents a probability vector.

learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Blondel, M., Martins, A. F., and Niculae, V. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21:1–69, 2020.

Brinker, K., Fürnkranz, J., and Hüllermeier, E. A unified model for multilabel classification and ranking. In *Proceedings of the 2006 conference on ECAI 2006: 17th European Conference on Artificial Intelligence August 29–September 1, 2006, Riva del Garda, Italy*, pp. 489–493, 2006.

Chan, K. H. R., Yu, Y., You, C., Qi, H., Wright, J., and Ma, Y. Redunet: A white-box deep network from the principle of maximizing rate reduction. *The Journal of Machine Learning Research*, 23(1):4907–5009, 2022.

Chang, W.-C., Yu, H.-F., Zhong, K., Yang, Y., and Dhillon, I. S. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3163–3171, 2020.

Chen, M., Fu, D. Y., Narayan, A., Zhang, M., Song, Z., Fatahalian, K., and Ré, C. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pp. 3090–3122. PMLR, 2022.

Cheng, W., Hüllermeier, E., and Dembczynski, K. J. Bayes optimal multilabel classification via probabilistic classifier chains. In *International Conference on Machine Learning*, pp. 279–286, 2010.

Cybenko, G. Approximation by superposition of sigmoidal functions. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Dang, H., Tran, T., Osher, S., Tran-The, H., Ho, N., and Nguyen, T. Neural collapse in deep linear networks: From balanced to imbalanced data. In *International Conference on Machine Learning*, 2023.

Dembczyński, K., Waegeman, W., Cheng, W., and Hüllermeier, E. Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21*, pp. 280–295. Springer, 2010.

Dembczynski, K., Kotlowski, W., and Hüllermeier, E. Consistent multilabel ranking through univariate losses. In *International Conference on Machine Learning*. PMLR, 2012.

Dembczyński, K., Waegeman, W., Cheng, W., and Hüllermeier, E. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88: 5–45, 2012.

Fang, C., He, H., Long, Q., and Su, W. J. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

Galanti, T. A note on the implicit bias towards minimal depth of deep neural networks. *arXiv preprint arXiv:2202.09028*, 2022.

Galanti, T., György, A., and Hutter, M. Generalization bounds for transfer learning with pretrained classifiers. *arXiv preprint arXiv:2212.12532*, 2022a.

Galanti, T., György, A., and Hutter, M. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022b.

Gao, P., Xu, Q., Wen, P., Shao, H., Yang, Z., and Huang, Q. A study of neural collapse phenomenon: Grassmannian frame, symmetry, generalization. *arXiv preprint arXiv:2304.08914*, 2023.

Gao, W. and Zhou, Z.-H. On the consistency of multi-label learning. In *Proceedings of the 24th annual conference on learning theory*, pp. 341–358. JMLR Workshop and Conference Proceedings, 2011.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pp. 797–842, 2015.

Graf, F., Hofer, C., Niethammer, M., and Kwitt, R. Dissecting supervised contrastive learning. In *International Conference on Machine Learning*, pp. 3821–3830. PMLR, 2021.

Han, X., Papyan, V., and Donoho, D. L. Neural collapse under mse loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022.

He, H. and Su, W. J. A law of data separation in deep learning. *Proceedings of the National Academy of Sciences*, 120(36):e2221704120, 2023. doi: 10.1073/pnas.2221704120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2221704120.

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

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hui, L., Belkin, M., and Nakkiran, P. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

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

Ji, W., Lu, Y., Zhang, Y., Deng, Z., and Su, W. J. An unconstrained layer-peeled perspective on neural collapse. In *International Conference on Learning Representations*, 2022.

Jiang, J., Zhou, J., Wang, P., Qu, Q., Mixon, D., You, C., and Zhu, Z. Generalized neural collapse for a large number of classes. *arXiv preprint arXiv:2310.05351*, 2023.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *International conference on machine learning*, pp. 1724–1732. PMLR, 2017.

Kothapalli, V. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=QTXocpAP9p.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Lanchantin, J., Wang, T., Ordonez, V., and Qi, Y. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16478–16488, 2021.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. at&t labs, 2010.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176:311–337, 2019.

Li, X., Liu, S., Zhou, J., Lu, X., Fernandez-Granda, C., Zhu, Z., and Qu, Q. Principled and efficient transfer learning of deep models via neural collapse. *arXiv preprint arXiv:2212.12206*, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

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

Liu, W., Wang, H., Shen, X., and Tsang, I. W. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7955–7974, 2021.

Liu, W., Yu, L., Weller, A., and Schölkopf, B. Generalizing and decoupling neural collapse via hyperspherical uniformity gap. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=inU2quhGdNU.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Lu, J. and Steinerberger, S. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.011. URL https://www.sciencedirect.com/science/article/pii/S1063520321001123. Special Issue on Harmonic Analysis and Machine Learning.

Lu, Y., Ji, W., Izzo, Z., and Ying, L. Importance tempering: Group robustness for overparameterized models. *arXiv preprint arXiv:2209.08745*, 2022.

Menon, A. K., Rawat, A. S., Reddi, S., and Kumar, S. Multilabel reductions: what is my loss optimising? *Advances in Neural Information Processing Systems*, 32, 2019.

Mixon, D. G., Parshall, H., and Pi, J. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 2022.

Moyano, J. M., Gibaja, E. L., Cios, K. J., and Ventura, S. Review of ensembles of multi-label classifiers: models, experimental study and prospects. *Information Fusion*, 44:33–45, 2018.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., and Dokania, P. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Papyan, V. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020.

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Rangamani, A., Lindegaard, M., Galanti, T., and Poggio, T. A. Feature learning in deep classifiers through intermediate neural collapse. In *International Conference on Machine Learning*, pp. 28729–28745. PMLR, 2023.

Reddi, S. J., Kale, S., Yu, F., Holtmann-Rice, D., Chen, J., and Kumar, S. Stochastic negative mining for learning with large output spaces. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1940–1949. PMLR, 2019.

Reeve, H. and Kaban, A. Optimistic bounds for multi-output learning. In *International Conference on Machine Learning*, pp. 8030–8040. PMLR, 2020.

Ridnik, T., Sharir, G., Ben-Cohen, A., Ben-Baruch, E., and Noy, A. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 32–41, 2023.

Samei, R., Semukhin, P., Yang, B., and Zilles, S. Sample compression for multi-label concept classes. In *Conference on Learning Theory*, pp. 371–393. PMLR, 2014a.

Samei, R., Yang, B., and Zilles, S. Generalizing labeled and unlabeled sample compression to multi-label concept classes. In *Algorithmic Learning Theory: 25th International Conference, ALT 2014, Bled, Slovenia, October 8-10, 2014. Proceedings 25*, pp. 275–290. Springer, 2014b.

Sharma, S., Xian, Y., Yu, N., and Singh, A. Learning prototype classifiers for long-tailed recognition. *arXiv preprint arXiv:2302.00491*, 2023.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Smith, L. N. Cyclical focal loss. *arXiv preprint arXiv:2202.08978*, 2022.

Sun, J., Qu, Q., and Wright, J. When are nonconvex problems not scary? In *NIPS Workshop on Nonconvex Optimization for Machine Learning*, 2015.

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63 (2):885–914, 2016.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Thrampoulidis, C., Kini, G. R., Vakilian, V., and Behnia, T. Imbalance trouble: Revisiting neural-collapse geometry. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27225–27238, 2022.

Tirer, T. and Bruna, J. Extended unconstrained features model for exploring deep neural collapse. In *International Conference on Machine Learning*, 2022.

Trabucco, B., Doherty, K., Gurinas, M., and Salakhutdinov, R. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023.

Wang, P., Liu, H., Yaras, C., Balzano, L., and Qu, Q. Linear convergence analysis of neural collapse with unconstrained features. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

Wang, P., Li, X., Yaras, C., Zhu, Z., Balzano, L., Hu, W., and Qu, Q. Understanding deep representation learning via layerwise feature compression and discrimination. *arXiv preprint arXiv:2311.02960*, 2023a.

Wang, Z., Luo, Y., Zheng, L., Huang, Z., and Baktashmotlagh, M. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5549–5558, 2023b.

Xie, L., Yang, Y., Cai, D., and He, X. Neural collapse inspired attraction-repulsion-balanced loss for imbalanced learning. *Neurocomputing*, 2023.

Xie, S., Qiu, J., Pasad, A., Du, L., Qu, Q., and Mei, H. Hidden state variability of pretrained language models can guide computation reduction for transfer learning. In *Empirical Methods in Natural Language Processing*, 2022.

Xu, C., Liu, T., Tao, D., and Xu, C. Local rademacher complexity for multi-label learning. *IEEE Transactions on Image Processing*, 25(3):1495–1507, 2016.

Yang, Y., Chen, S., Li, X., Xie, L., Lin, Z., and Tao, D. 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*, 2022.

Yang, Y., Yuan, H., Li, X., Lin, Z., Torr, P., and Tao, D. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=y5W8tpojhtJ.

Yaras, C., Wang, P., Zhu, Z., Balzano, L., and Qu, Q. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. In *Advances in Neural Information Processing Systems*, 2022.

Yaras, C., Wang, P., Hu, W., Zhu, Z., Balzano, L., and Qu, Q. The law of parsimony in gradient descent for learning deep linear networks. *arXiv preprint arXiv:2306.01154*, 2023.

Yu, L., Hu, T., HONG, L., Liu, Z., Weller, A., and Liu, W. Continual learning by modeling intra-class variation. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iDxfGaMYVr.

Yu, Y., Chan, K. H. R., You, C., Song, C., and Ma, Y. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems*, 33:9422–9434, 2020.

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J., and Ma, Y. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, H., Zhou, J., Lu, Y., Guo, M., Shen, L., and Qu, Q. The emergence of reproducibility and consistency in diffusion models. *arXiv preprint arXiv:2310.05264*, 2023.

Zhang, M., Ramaswamy, H. G., and Agarwal, S. Convex calibrated surrogates for the multi-label f-measure. In *International Conference on Machine Learning*, pp. 11246–11255. PMLR, 2020a.

Zhang, Y., Qu, Q., and Wright, J. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020b.

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

Zhou, J., Li, X., Ding, T., You, C., Qu, Q., and Zhu, Z. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *International Conference on Machine Learning*, pp. 27179–27202. PMLR, 2022a.

Zhou, J., You, C., Li, X., Liu, K., Liu, S., Qu, Q., and Zhu, Z. Are all losses created equal: A neural collapse perspective. *Advances in Neural Information Processing Systems*, 2022b.

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

Appendix

Organization of Appendices

Appendix A	**Related Works and Futue Directions**
Appendix B	**Dataset Illustrations, Visualizations and Experimental Setups**
Appendix C	**Proofs for Optimality Condition**
Appendix D	**Nonconvex Landscape Analysis**

Table 3. **Table of Contents for Appendices**

In Appendix A, we provided more related work on multi-label learning and Neural Collapse with more detailed future direction of our paper. In Appendix B, We present all the datasets utilized for validating multi-label NC, guiding testing and training, along with the experimental setups introduced in this study. In Appendix C and Appendix D, we present the proofs for results from the main paper Theorem 3.1 (global optimality) and Theorem D.1 (benign landscapes), respectively.

A. Related Works and Future Direction

A.1. Discussion on Related Works

Related works on multi-label learning. In contrast to M-clf, where each sample has a single label, in M-lab the samples are tagged with multiple labels. As such, the final model output must be set-valued. This presents theoretical and practical challenges unique to the regime of M-lab learning, especially when the class number is large (Liu et al., 2021). Besides BR and PAL, there are also non-decomposable approaches to M-lab learning (Dembczyński et al., 2012). However, to the best of our knowledge, the contribution of the work (Dembczyński et al., 2012) is more on the theoretical side, while it has limited impacts on training practical neural networks for M-lab. Therefore, in this work we only focus on reducible formulations such as PAL and BR. On the practical side, many modern deep neural network architectures have been successfully adapted to the multi-label task (Chang et al., 2020; Lanchantin et al., 2021; Ridnik et al., 2023). However, the methods often suffer from the challenges of imbalanced training data, given that high Multiplicity labels are scarce.

On the theory side, consistency of surrogate methods for M-lab has been initiated by (Gao & Zhou, 2011) and followed up by several works in (Menon et al., 2019; Dembczynski et al., 2012; Zhang et al., 2020a; Blondel et al., 2020). Many other concepts from classical learning theory have also been extended successfully to the M-lab regime, e.g., Vapnik-Chervonenkis theory and sample-compression schemes (Samei et al., 2014a;b), (local) Rademacher complexity (Xu et al., 2016; Reeve & Kaban, 2020), and Bayes-optimal prediction (Cheng et al., 2010).

Related works on neural collapse. The phenomenon known as NC was initially identified in recent groundbreaking research (Papyan et al., 2020; Han et al., 2022) conducted on M-clf. These studies provided empirical evidence demonstrating the prevalence of NC across various network architectures and datasets. The significance of NC lies in its elegant mathematical characterization of learned representations or features in deep learning models for M-clf. Notably, this characterization is independent of network architectures, dataset properties, and optimization algorithms, as also highlighted in a recent review paper (Kothapalli, 2023). Subsequent investigations, building upon the "unconstrained feature model" (Mixon et al., 2022) or the "layer-peeled model" (Fang et al., 2021), have contributed theoretical evidence supporting the existence of NC. This evidence pertains to the utilization of a range of loss functions, including cross-entropy (CE) loss (Lu & Steinerberger, 2022; Zhu et al., 2021; Fang et al., 2021; Yaras et al., 2022), mean-square-error (MSE) loss (Mixon et al., 2022; Zhou et al., 2022a; Tirer & Bruna, 2022; Rangamani & Banburski-Fahey, 2022; Wang et al., 2022; Dang et al., 2023), and CE variants (Graf et al., 2021; Zhou et al., 2022b). More recent studies have explored other theoretical aspects of NC, such as its relationship with generalization (Hui et al., 2022; Galanti et al., 2022b;a; Galanti, 2022; Chen et al., 2022), its applicability to large classes (Liu et al., 2023; Gao et al., 2023; Jiang et al., 2023), and the progressive collapse of feature variability across intermediate network layers (Hui et al., 2022; Papyan, 2020; He & Su, 2023; Yaras et al., 2023; Rangamani et al., 2023). Theoretical findings related to NC have also inspired the development of new techniques to improve practical performance in various scenarios, including the design of loss functions and architectures (Yu et al., 2020; Zhu et al., 2021;

Chan et al., 2022), transfer learning (Li et al., 2022; Xie et al., 2022; Wang et al., 2023b) where a model trained on one task or dataset is adapted or fine-tuned to perform a different but related task, imbalanced learning (Fang et al., 2021; Xie et al., 2023; Lu et al., 2022; Yang et al., 2022; Thrampoulidis et al., 2022; Behnia et al., 2023; Zhong et al., 2023; Sharma et al., 2023) which is a characteristic of dataset where one or more classes have significantly fewer instances compared to other classes, and continual learning (Yu et al., 2023; Yang et al., 2023; Zhai et al., 2023) in which a model is designed to learn and adapt to new data continuously over time, rather than being trained on a fixed dataset.

A.2. Other Future Directions

In this study, we extensively analyzed the NC phenomenon in M-lab (Zhu et al., 2021; Fang et al., 2021; Ji et al., 2022). Based upon the UFM, our results establish that M-lab ETFs are the only global minimizers of the PAL loss function, incorporating weight decay and bias. These findings hold significant implications for improve the performance and training efficiency of M-lab tasks. We believe that our results open several interesting directions that is worth of further exploration that we discuss below.

Extreme Multi-label classification The goal of extreme multi-label classification is to tag a data point with the most relevant subset of labels from an extremely large label set. We have explored our M-lab NC phenomenon on the more challenging Microsoft COCO dataset (Lin et al., 2014) with minimal preprocessing to maintain balance among multiplicity-1 data. The visualization of M-lab NC measures is shown in Figure 5. Specifically, a subset of the COCO dataset comprising 32,083 samples was extracted. We included 50 classes with all higher multiplicity samples possible. Most samples belong to higher multiplicity (29,583 samples) compared to multiplicity-1 samples (only 2,500). The ResNet18 architecture was trained for 200 epochs, achieving a training IoU of around 98%. For training efficiency, we fixed the classifier W as an ETF and downsampled images to 64 by 64 pixels. Under these experimental setups, we observed that $\mathcal{NC}_1$, $\mathcal{NC}_3$, and $\mathcal{NC}_m$ all converged to small values, demonstrating the Mlab-NC phenomenon.

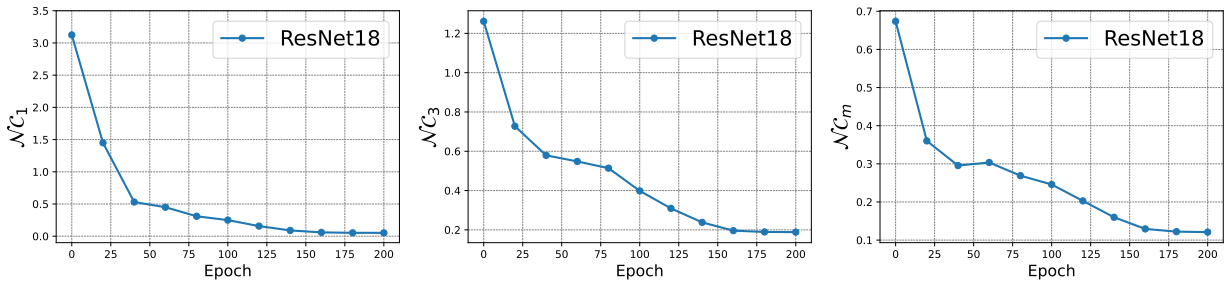


Figure 5. M-lab NC phenomenon in extreme MS-COCO dataset. As we can see that all M-lab NC measures converges to small values.

While the Microsoft COCO dataset indeed contains a large number of labels, there is still a gap between its setup and the most extreme multi-label datasets. The work by (Jiang et al., 2023) studied the phenomenon of neural collapse under extreme multi-class classification where the number of labels exceeds the dimension of learned features. To our knowledge, no one has studied neural collapse in the context of extreme multi-label classification. Our tentative experiments on the Microsoft COCO dataset demonstrate the potential to develop a generalized neural collapse for multi-label classification with a large number of labels.

Features for data with all possible tags While our theoretical analysis of the scaled tag-wise average property assumes that the multiplicity of a data point is less than the total number of classes (i.e., $M < K$), experimental results on the toy Microsoft COCO dataset indicate that the tag-wise average property still holds for data samples with multiplicity $M = K$. The visualization of learned data features is shown in Figure 6. Specifically, we selected three classes: Car, Train, and Truck, where each sample can have 1 to 3 labels. This means that the dataset comprises samples with multiplicities of 1, 2, and 3, based on the number of labels they carry. We then trained a 5-layer convolutional network using the selected data and we chose a feature dimension 2 for visualization purposes. As shown in the visualization, both the global feature mean and the multiplicity-3 features are indeed located at the origin.

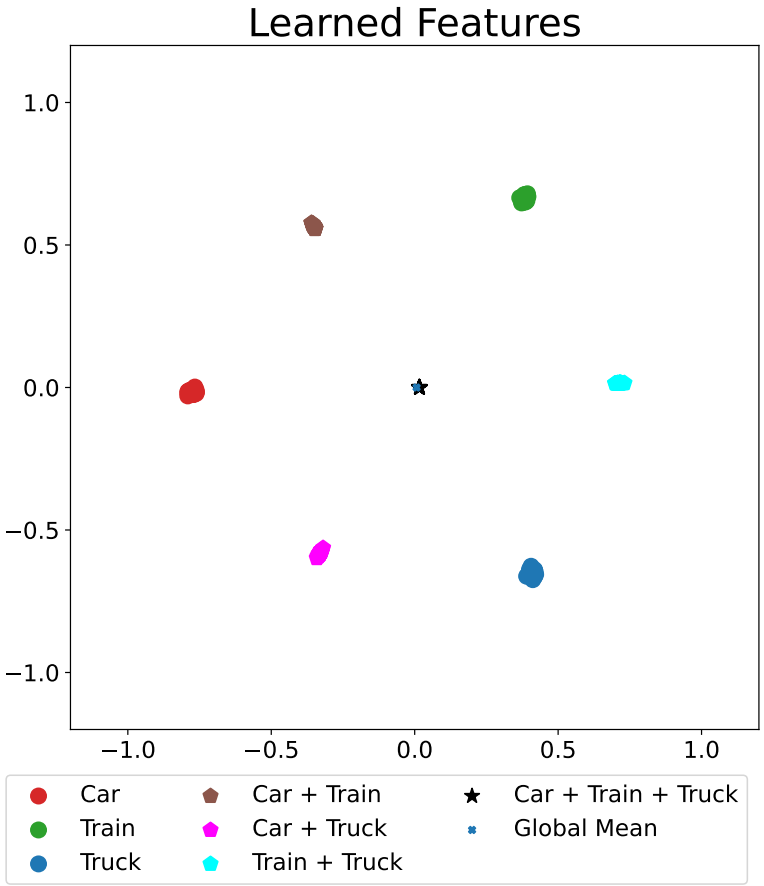
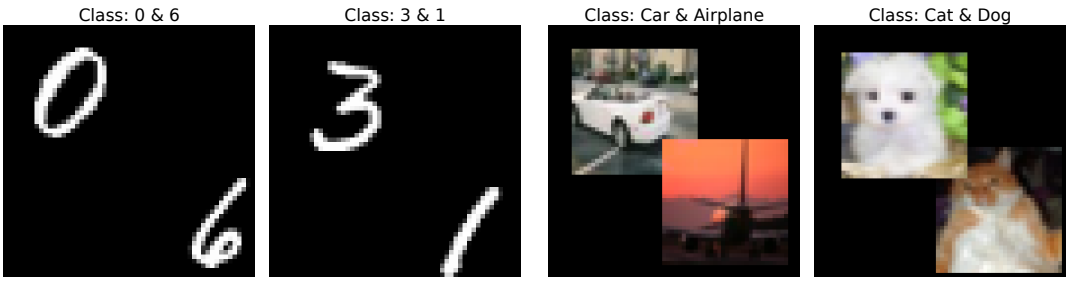


Figure 6. M-lab NC: learned features for data samples with all possible labels is at the origin and follows tag-wise average property.



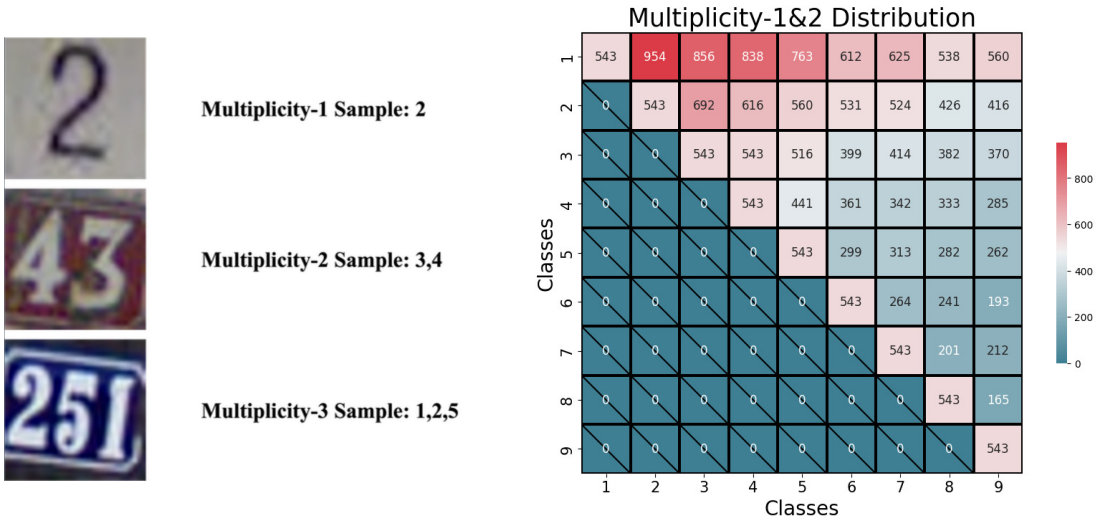
(a) MLab-MNIST (b) MLab-Cifar10

Figure 7. **Illustration of synthetic multi-label MNIST (left) and Cifar10 (right) datasets.**

B. Dataset Illustration and Visualization

B.1. M-lab MNIST and M-lab Cifar10 dataset

We created synthetic Multi-label MNIST (LeCun et al., 2010) and Cifar10 (Krizhevsky et al., 2009) datasets by applying zero-padding to each image, increasing its width and height to twice the original size, and then combining it with another padded image from a different class. An illustration of generated multi-label samples can be found in Figure 7. To create the training dataset, for $m = 1$ scenario, we randomly picked 3100 images in each class, and for $m = 2$, we generated 200 images for each combination of classes using the pad-stack method described earlier. Therefore, the total number of images in the training dataset is calculated as $10 \times 3100 + \binom{10}{2} \times 200 = 40000$. Those dataset are used to generate results



(a) The SVHN Dataset

(b) Class Distribution

Figure 8. **Illustration of M-lab SVHN dataset.** As illustrated in (a), the Street View House Numbers (SVHN) Dataset (Netzer et al., 2011) comprises labeled numerical characters and inherently serves as a M-lab learning dataset. We applied minimal preprocessing to achieve balance specifically within the Multiplicity-1 scenario, as evidenced by the diagonal entries in (b). Furthermore, we omitted samples with Multiplicity-4 and above, as these images posed considerable recognition challenges. Notably, the Multiplicity-2 case remained largely imbalanced, as observed in the off-diagonal entries in (b). Nonetheless, our findings remained robust and consistent in this scenario, as evidenced in Figure 4.

in Figure 3 and Table 2.

In terms of training deep networks for M-lab, we use standard ResNet (He et al., 2016) and VGG (Simonyan & Zisserman, 2014) network architectures. Throughout all the experiments, we use an SGD optimizer with fixed batch size 128, weight decay $(\lambda_{\boldsymbol{W}}, \lambda_{\boldsymbol{H}}) = (5 \times 10^{-4}, 5 \times 10^{-4})$ and momentum 0.9. The learning rate is initially set to 1×10^{-1} and dynamically decays to 1×10^{-3} following a CosineAnnealing learning rate scheduler as described in (Loshchilov & Hutter, 2017). The total number of epochs is set to 200 for all experiments.

B.2. Multiplicity 2 imbalanced M-lab MNIST and M-lab Cifar10 dataset

Following the same padding rule described in Appendix B.1, the multiplicity-2 imbalanced data used to generate Figure 2 are created as follows. The cifar10 dataset has balanced Multiplicity-1 samples (5000 for each class). For the classes of Multiplicity-2, we divide them into 3 groups: the large group (500 samples), the middle group (50 samples), and the small group (5 samples).

B.3. Multiplicity 2 imbalanced M-lab SVHN dataset

To further explore our findings, we conducted additional experiments on the practical SVHN dataset (Netzer et al., 2011) alongside the synthetic datasets. In order to preserve the natural characteristics of the SVHN dataset, we applied minimal pre-processing only to ensure a balanced scenario for multiplicity-1, while leaving other aspects of the dataset untouched. The dataset are visualized in Figure 8.

B.4. Dataset used to compare test accuracy and efficiency between ONN and OvA

For training dataset, following the same generation method described in Appendix B.1 and simply varying data balancedness, the synthetic multiplicity imbalanced data used in Table 1 are generated from Cifar10 datasets. All 3 datasets has 1500 sample in every class in multiplicity-1, we reduce the sample in every class in multiplicity-2 to 1000, 750, and 500, resulting in the “c10-Large”, “c10-Medium”, and “c10-Small” datasets. The testing datasets are independently generated, each with a sample size equivalent to 20% of the training datasets.

Standard ResNet (He et al., 2016) network architecture are used for training with only fixing the last layer classifier of as ETF. The SGD optimizer with fixed batch size 128 are used. Specifically, for the three Cifar-10 datasets, models are trained with weight decay $(\lambda_{\boldsymbol{W}}, \lambda_{\boldsymbol{H}}) = (5 \times 10^{-5}, 10^{-5})$ with 200 epochs with learning rate of 0.1. For the SVHN dataset, models are trained with weight decay $(\lambda_{\boldsymbol{W}}, \lambda_{\boldsymbol{H}}) = (5 \times 10^{-6}, 1.5 \times 10^{-6})$ with 100 epochs with learning rate of 0.09. For testing with OvA, the additional linear classifiers are trained until the training loss reaches 0, typically after approximately 2 epochs.

C. Optimality Condition

The purpose of this section is to prove Theorem 3.1. As such, throughout this section, we assume that we are in the situation of the statement of said theorem. Due to the additional complexity of the M-lab setting compared to the M-clf setting, analysis of the M-lab NC requires substantially more notations. These notations, which are defined in Appendix C.1, while not necessary for *stating* Theorem 3.1, are crucial for the proofs in Appendix C.2 .

C.1. Additional notations

For the reader's convenience, we recall the following:

$$N := \text{number of samples} \tag{9}$$

$$N_m := \text{number of samples } i \in [N] \text{ such that } |S_i| = m \tag{10}$$

$$n_m := N_m / \binom{K}{m} \tag{11}$$

$$\binom{[K]}{m} := \{S \subseteq [K] : |S| = m\} \tag{12}$$

$$M := \text{largest } m \text{ such that } n_m \neq 0 \tag{13}$$

$$d := \text{dimension of the last layer features} \tag{14}$$

C.1.1. Lexicographical ordering on subsets

For each $m \leq K$, recall from the above that the set of subsets of $[K]$ of size m is denoted by the commonly used, suggestive notation $\binom{[K]}{m}$. Moreover, $|\binom{[K]}{m}| = \binom{K}{m}$.

▷ **Notation convention**. Assume the *lexicographical ordering* on $\binom{[K]}{m}$. Thus, for each $k \in \binom{K}{m}$, the *k-th subset of* $\binom{[K]}{m}$ is well-defined.

For example, when $K = 5$ and $m = 2$, there are $\binom{5}{2} = 10$ elements in $\binom{[5]}{2}$ which, when listed in the lexicographic ordering, are

$$\underbrace{\{1,2\}}_{\text{1st}}, \underbrace{\{1,3\}}_{\text{2nd}}, \underbrace{\{1,4\}}_{\text{3rd}}, \underbrace{\{1,5\}}_{\text{4th}}, \underbrace{\{2,3\}}_{\cdots}, \{2,4\}, \{2,5\}, \{3,4\}, \{3,5\}, \underbrace{\{4,5\}}_{\text{10th}}.$$

In general, we use the notation $S_{m,k}$ to denote the k-th subset of $\binom{[K]}{m}$. In other words,

$$\binom{[K]}{m} = \{S_{m,1},\ S_{m,2},\ \ldots,\ S_{m,\binom{K}{m}}\}.$$

C.1.2. Block submatrices of the last layer feature matrix

Without the loss of generality, we assume that the sample indices $i \in [N]$ are sorted such that $|S_i|$ is non-decreasing, i.e., $|S_1| \leq \cdots \leq |S_i| \leq \cdots \leq |S_N|$. Clearly, this does not affect the optimization problem itself. Denote the set of indices of Multiplicity-m samples by $\mathcal{I}_m := \{i \in [N] : |S_i| = m\}$. Thus, we have

$$\mathcal{I}_1 = \{1, \ldots, N_1\},\ \mathcal{I}_2 = \{1 + N_1, \ldots, N_2 + N_1\},\ \cdots \mathcal{I}_m = \{1 + \sum_{\ell=1}^{m} N_\ell, \ldots, N_m + \sum_{\ell=1}^{m} N_\ell\}, \cdots$$

Below, it will be helpful to define the notation

$$\mathcal{I}_{m,S} := \{i \in [N] : S_i = S\}$$

for each $m = 1, \dots, M$ and $S \in \binom{[K]}{m}$.

▷ **Notation convention**. Define the block-submatrices $\boldsymbol{H}_1, \dots, \boldsymbol{H}_M$ of $\boldsymbol{H}$ such that

1. $\boldsymbol{H}_m \in \mathbb{R}^{d \times N_m}$
2. $\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_1 & \boldsymbol{H}_2 & \cdots & \boldsymbol{H}_M \end{bmatrix}$

Thus, as in the main paper, the columns of $\boldsymbol{H}_m$ correspond to the features of $\mathcal{I}_m$.

C.1.3. Decomposition of the PAL-CE loss

Define

$$g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}, \boldsymbol{Y}) := \frac{1}{N_m} \sum_{i \in \mathcal{I}_m} \underline{\mathcal{L}}_{\texttt{PAL}}(\boldsymbol{W}\boldsymbol{h}_i + \boldsymbol{b}, \boldsymbol{y}_{S_i}). \tag{15}$$

Intuitively, g_m is the contribution to g from the Multiplicity-m samples. More precisely, the function $g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}, \boldsymbol{Y})$ from Equation (4) can be decomposed as

$$g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}, \boldsymbol{Y}) = \sum_{m=1}^{M} \frac{N_m}{N} g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}, \boldsymbol{Y}). \tag{16}$$

C.1.4. Triple indices notation

Next, we state precisely the data balanced-ness condition from Theorem 3.1. In order to state the condition, we need some additional notations. Fix some $m \in \{1, \dots, M\}$ and let $S \in \binom{[K]}{m}$. Define

$$n_{m,S} := \{i \in [N] : S_i = S\}. \tag{17}$$

Theorem 3.1 made the following **data balanced-ness condition**:

$$n_{m,S} = N_m / \tbinom{K}{m} =: n_m \text{ for all } S \in \tbinom{[K]}{m}. \tag{18}$$

In other words, for a fixed $m \in [M]$, the set $\mathcal{I}_{m,S}$ has the same constant cardinality equal to n_m ranging across all $S \in \binom{[K]}{m}$.

By the data balanced-ness condition, we have for a fixed $m = 1, \dots, M$ that $\mathcal{I}_{m,S}$ have the same number of elements across all $S \in \binom{[K]}{m}$. Moreover, in our notation, we have $|\mathcal{I}_{m,S}| = n_m$. Below, for each $m = 1, \dots, M$ and for each $S \in \binom{[K]}{m}$, choose an arbitrary ordering on $\mathcal{I}_{m,S}$ once and for all. Every sample is *uniquely* specified by the following three indices:

1. $m \in [M]$ the sample's multiplicity, i.e., $m = |S|$
2. $k \in \binom{K}{m}$ the index such that $S_{m,k}$ is the label set of the sample,
3. $i \in [n_m]$ such that the sample is the i-th element of $\mathcal{I}_{m,S_{m,k}}$.

More concisely, we now introduce the

▷ **Notation convention**. Denote each sample by the triplet

$$(m, k, i) \quad \text{where} \quad m \in [M],\ k \in \binom{K}{m},\ i \in [n_m]. \tag{19}$$

Below, (19) will be referred to as the **triple indices notation** and every sample will be referred to by its triple indices (m, k, i) instead of the previous single index $i \in [N]$. Accordingly, throughout the appendix, columns of $\boldsymbol{H}$ are expressed as $\boldsymbol{h}_{m,k,i}$ instead of the previous $\boldsymbol{h}_i$, and thus the block submatrix $\boldsymbol{H}_m$ of $\boldsymbol{H}$ can be, without the loss of generality, be written as $\boldsymbol{H}_m = \left[\boldsymbol{h}_{m,k,i}\right]_{m \in [M],\ k \in \binom{K}{m},\ i \in [n_m]}$.

Moreover, in the triple indices notation, Equation (15) can be rewritten as

$$g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}) = \frac{1}{N_m} \sum_{i=1}^{n_m} \sum_{k=1}^{\binom{K}{m}} \mathcal{L}_{\texttt{PAL}}(\boldsymbol{W}\boldsymbol{h}_{m,k,i}, \boldsymbol{y}_{S_{m,k}}) \tag{20}$$

C.2. Proofs

We will first state the proof of Theorem 3.1 which depends on several lemmas appearing later in the section. Thus, the proof of Theorem 3.1 serves as a roadmap for the rest of this section.

Proof of Theorem 3.1. Recall the definition of a coercive function: a function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is said to be coercive if $\lim_{\|x\|\to\infty} \varphi(x) = +\infty$. It is well-known that a coercive function attains its infimum which is a global minimum.

Now, note that the objective function $f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ in Problem (4) is *coercive* due to the weight decay regularizers (the terms $\|\boldsymbol{W}\|_F^2$, $\|\boldsymbol{H}\|_F^2$ and $\|\boldsymbol{b}\|_F^2$) and that the pick-all-labels cross-entropy loss is non-negative. Thus, a global minimizer, denoted below as $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$, of Problem (4) exists. By Lemma B.2, we know that any critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of Problem (4) satisfies

$$\boldsymbol{W}^\top \boldsymbol{W} = \frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}} \boldsymbol{H}\boldsymbol{H}^\top.$$

Let $\rho := \|\boldsymbol{W}\|_F^2$. Thus, $\|\boldsymbol{H}\|_F^2 = \frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}\rho$

We first provide a lower bound for the PAL cross-entropy term $g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top)$ and then show that the lower bound is tight if and only if the parameters are in the form described in Theorem 3.1. For each $m = 1, \ldots, M$, let $c_{1,m} > 0$ be arbitrary, to be determined below. Now by Lemma C.2 and Lemma C.8, we have

$$g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}) - \Gamma_2 \geq -\frac{1}{N}\sqrt{\sum_{m=1}^{M}\left(\frac{1}{1+c_{1,m}}\frac{m}{K-m}\right)^2 \kappa_m n_m \binom{K}{m}^2}\sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}}\rho$$

where $\Gamma_2 := \sum_{m=1}^{M} c_{2,m}$ and $c_{2,m}$ is as in Lemma C.8. Therefore, we have

$$\begin{aligned} f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b}) &= g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}^\top) + \lambda_{\boldsymbol{W}}\|\boldsymbol{W}\|_F^2 + \lambda_{\boldsymbol{H}}\|\boldsymbol{H}\|_F^2 + \lambda_{\boldsymbol{b}}\|\boldsymbol{b}\|_2^2 \\ &\geq -\frac{1}{N}\sqrt{\sum_{m=1}^{M}\left(\frac{1}{1+c_{1,m}}\frac{m}{K-m}\right)^2 \kappa_m n_m \binom{K}{m}^2}\sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}}\rho + \Gamma_2 + 2\lambda_{\boldsymbol{W}}\rho + \frac{\lambda_{\boldsymbol{b}}}{2}\|\boldsymbol{b}\|_2^2 \\ &\geq -\frac{1}{N}\sqrt{\sum_{m=1}^{M}\left(\frac{1}{1+c_{1,m}}\frac{m}{K-m}\right)^2 \kappa_m n_m \binom{K}{m}^2}\sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}}\rho + \Gamma_2 + 2\lambda_{\boldsymbol{W}}\rho \end{aligned} \tag{21}$$

where the last inequality becomes an equality whenever either $\lambda_{\boldsymbol{b}} = 0$ or $\boldsymbol{b} = \mathbf{0}$. Furthermore, by Lemma C.3, we know that the Inequality (21) becomes an equality *if and only if* $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ satisfy the following:

(I) $\|\boldsymbol{w}^1\|_2 = \|\boldsymbol{w}^2\|_2 = \cdots = \|\boldsymbol{w}^K\|_2$, and $\boldsymbol{b} = b\mathbf{1}$,

(II) $\frac{1}{\binom{K}{m}}\sum_{k=1}^{\binom{K}{m}} \boldsymbol{h}_{m,k,i} = \mathbf{0}$, and $\sqrt{\frac{\binom{K-2}{m-1}}{n_m}}\boldsymbol{w}^k = \sum_{\ell:k\in S_{m,\ell}} \boldsymbol{h}_{m,\ell,i}, \forall m \in [M], k \in [K], i \in [n_m]$,

(III) $\boldsymbol{W}^\top\boldsymbol{W} = \frac{\rho}{K-1}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)$

(IV) There exist unique positive real numbers $C_1, C_2, \ldots, C_M > 0$ such that the following holds:

$\boldsymbol{h}_{1,k,i} = C_1\boldsymbol{w}^\ell$ when $S_{1,k} = \{\ell\}$, $\ell \in [K]$, (Multiplicity = 1 Case)

$\boldsymbol{h}_{m,k,i} = C_m\sum_{\ell\in S_{m,k}}\boldsymbol{w}^\ell$ when $m > 1$. (Multiplicity > 1 Case)

Note that condition (IV) is a restatement of Equation (7) and Equation (8). The choice of the $c_{1,m}$'s is given by (V) from Lemma C.3. □

Lemma C.1. *we have:*

$$\boldsymbol{W}^\top\boldsymbol{W} = \frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}\boldsymbol{H}\boldsymbol{H}^\top \quad \textit{and} \quad \rho = \|\boldsymbol{W}\|_F^2 = \frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}\|\boldsymbol{H}\|_F^2$$

Proof. The proceeds identically as in given by (Zhu et al., 2021) Lemma B.2 and is thus omitted here. □

The following lemma is the generalization of (Zhu et al., 2021) Lemma B.3 to the multilabel case for each multiplicity.

Lemma C.2. *Let $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ be a critical point for the objective f from Problem (4). Let $c_{1,m} > 0$ be arbitrary and let $\gamma_{1,m} := \frac{1}{1+c_{1,m}} \frac{m}{K-m}$. Define $\kappa_m := \left(\frac{K}{m\binom{K}{m}}\right)^2 \binom{K-2}{m-1}$. Then*

$$g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}) - \Gamma_2 \geq -\frac{1}{N}\sqrt{\sum_{m=1}^{M}\left(\frac{1}{1+c_{1,m}}\frac{m}{K-m}\right)^2 \kappa_m n_m \binom{K}{m}^2}\sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}}\rho. \tag{22}$$

where $\rho := \|\boldsymbol{W}\|_F^2$, $\Gamma_2 := \sum_{m=1}^{M} c_{2,m}$ and $c_{2,m}$ is as in Lemma C.8.

Note that Γ_2 depends on $c_{1,1}, c_{1,2}, \ldots, c_{1,M}$ because $c_{2,m}$ depends on $c_{1,m}$ for each $m \in [M]$.

Proof. Throughout this proof, let $\boldsymbol{z}_{m,k,i} := \boldsymbol{W}\boldsymbol{h}_{m,k,i} + \boldsymbol{b}$ and choose the same $\gamma_{1,m}, c_{2,m}$ for all i and k. The first part of this proof aim to find the lower bound for each $g_m(\boldsymbol{W}, \boldsymbol{H}_m, \boldsymbol{b})$ along with conditions when the bound is tight. The rest of the proof focus on sum up g_m to get Equation (22). Thus, using Equation (20) with the $\boldsymbol{z}_{m,k,i}$'s, we have that g_m can be written as

$$g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}) = \frac{1}{N_m}\sum_{i=1}^{n_m}\sum_{k=1}^{\binom{K}{m}} \underline{\mathcal{L}}_{\texttt{PAL}}(\boldsymbol{z}_{m,k,i}, \boldsymbol{y}_{S_{m,k}}) \tag{23}$$

By directly applying Lemma C.8, the following lower bound holds:

$$N_m g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}) \geq \gamma_{1,m}\sum_{i=1}^{n_m}\sum_{k=1}^{\binom{K}{m}} \langle \mathbf{1} - \tfrac{K}{m}\mathbb{I}_S,\, \boldsymbol{W}\boldsymbol{h}_{m,k,i} + \boldsymbol{b}\rangle + N_m c_{2,m}$$

which implies that

$$\begin{aligned}
&\gamma_{1,m}^{-1}(g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}) - c_{2,m}) \\
&\geq \frac{1}{N_m}\sum_{i=1}^{n_m}\sum_{k=1}^{\binom{K}{m}} \langle \mathbf{1} - \tfrac{K}{m}\mathbb{I}_S,\, \boldsymbol{W}\boldsymbol{h}_{m,k,i} + \boldsymbol{b}\rangle \\
&= \underbrace{\frac{1}{N_m}\sum_{i=1}^{n_m}\sum_{k=1}^{\binom{K}{m}} \langle \mathbf{1} - \tfrac{K}{m}\mathbb{I}_S,\, \boldsymbol{W}\boldsymbol{h}_{m,k,i}\rangle}_{(\star)} + \underbrace{\frac{1}{N_m}\sum_{i=1}^{n_m}\sum_{k=1}^{\binom{K}{m}} \langle \mathbf{1} - \tfrac{K}{m}\mathbb{I}_S, \boldsymbol{b}\rangle}_{(\star\star)}
\end{aligned} \tag{24}$$

To further simplify the inequality above, we break it down into two parts, namely, the feature part $(\star)$ and the bias part $(\star\star)$ and analyze each of them separately. We first show that the term $(\star\star)$ is equal to zero. To see this, note that

$$\begin{aligned}
(\star\star) &= \sum_{k=1}^{\binom{K}{m}}\left(\sum_{j=1}^{K} b_j - \frac{K}{m}\sum_{j' \in S_{m,k}} b_{j'}\right) \\
&= \sum_{k=1}^{\binom{K}{m}}\sum_{j=1}^{K} b_j - \frac{K}{m}\sum_{k=1}^{\binom{K}{m}}\sum_{j' \in S_{m,k}} b_{j'} \\
&= K\binom{K}{m}\bar{b} - \frac{K}{m}m\binom{K}{m}\bar{b} \\
&= 0
\end{aligned} \tag{25}$$

where $\bar{b} = \frac{1}{K}\sum_{j=1}^{K} b_j$ and $\sum_{k=1}^{\binom{K}{m}}\sum_{j=1}^{K} b_j = K\binom{K}{m}\bar{b}$. Thus

$$\sum_{k=1}^{\binom{K}{m}}\sum_{j'\in S_{m,k}} b_{j'} \overset{(\diamondsuit)}{=} \sum_{j=1}^{K}\sum_{k:j\in S_{m,k}} b_j = \sum_{j=1}^{K} b_j \#\{k : j \in S_{m,k}\} = \sum_{j=1}^{K}\binom{K}{m}\frac{m}{K} b_j = m\binom{K}{m}\bar{b}.$$

Note that the equality at $(\diamondsuit)$ holds by switching the order of the summation. Now, substituting the result of Equation (25) into the Inequality (24), we have the new lower bound of g_m:

$$\gamma_{1,m}^{-1}(g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}) - c_{2,m}) \geq \frac{1}{N_m}\sum_{i=1}^{n_m}\underbrace{\sum_{k=1}^{\binom{K}{m}}\langle \mathbf{1} - \tfrac{K}{m}\mathbb{I}_S,\, \boldsymbol{W}\boldsymbol{h}_{m,k,i}\rangle}_{(\star)} \tag{26}$$

and the bound is tight when conditions are met in Lemma C.8. To simplify the expression $(\star)$ we first distribute the outer layer summation and further simplify it as:

$$\begin{aligned}
(\star) &= \sum_{k=1}^{\binom{K}{m}}\sum_{j=1}^{K}\boldsymbol{h}_{m,k,i}^{\top}\cdot\boldsymbol{w}^j - \frac{K}{m}\sum_{k=1}^{\binom{K}{m}}\sum_{j'\in S_{m,k}}\boldsymbol{h}_{m,k,i}^{\top}\cdot\boldsymbol{w}^{j'} \\
&= \sum_{k=1}^{\binom{K}{m}}\sum_{j=1}^{K}\boldsymbol{h}_{m,k,i}^{\top}\cdot\boldsymbol{w}^j - \frac{K}{m}\sum_{j=1}^{K}\sum_{k':j\in S_{k'}}\boldsymbol{h}_{m,k',i}^{\top}\boldsymbol{w}^j \qquad (27)\\
&= \sum_{k=1}^{\binom{K}{m}}\sum_{j=1}^{K}\boldsymbol{h}_{m,k,i}^{\top}\cdot\boldsymbol{w}^j - \frac{K}{m}\sum_{j=1}^{K}\boldsymbol{h}_{m,\{j\},i}^{\top}\boldsymbol{w}^j \\
&= \sum_{j=1}^{K}\sum_{k=1}^{\binom{K}{m}}\boldsymbol{h}_{m,k,i}^{\top}\cdot\boldsymbol{w}^j - \frac{K}{m}\sum_{j=1}^{K}\boldsymbol{h}_{m,\{j\},i}^{\top}\boldsymbol{w}^j \qquad (28)\\
&= \sum_{j=1}^{K}\left(\sum_{k=1}^{\binom{K}{m}}\boldsymbol{h}_{m,k,i} - \frac{K}{m}\boldsymbol{h}_{m,\{j\},i}\right)^{\top}\boldsymbol{w}^j \\
&= \sum_{j=1}^{K}\left(\binom{K}{m}\overline{\boldsymbol{h}}_{m,\bullet,i} - \frac{K}{m}\boldsymbol{h}_{m,\{j\},i}\right)^{\top}\boldsymbol{w}^j \qquad (29)
\end{aligned}$$

where we let $\boldsymbol{h}_{m,\{j\},i} = \sum_{k:j\in S_{m,k}}\boldsymbol{h}_{m,k,i}$ and $\overline{\boldsymbol{h}}_{m,\bullet,i}$ be the "average" of $\boldsymbol{h}_{m,k,i}$ over all $k \in \binom{K}{m}$ defined as:

$$\overline{\boldsymbol{h}}_{m,\bullet,i} := \frac{1}{\binom{K}{m}}\sum_{k=1}^{\binom{K}{m}}\boldsymbol{h}_{m,k,i}. \tag{30}$$

Similarly to $(\diamondsuit)$, the Equations (27) and (28) holds since we only switch the order of summation. Continuing simplification, we substitute the result in Equations (29) and (25) into Inequality (24) we have:

$$\begin{aligned}
\gamma_{1,m}^{-1}(g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}) - c_{2,m}) &\geq \frac{1}{N_m}\sum_{i=1}^{n_m}\sum_{j=1}^{K}\left(\binom{K}{m}\overline{\boldsymbol{h}}_{m,\bullet,i} - \frac{K}{m}\boldsymbol{h}_{m,\{j\},i}\right)^{\top}\boldsymbol{w}^j \\
&= \frac{1}{N_m}\sum_{i=1}^{n_m}\sum_{k=1}^{K}\left(\binom{K}{m}\overline{\boldsymbol{h}}_{m,\bullet,i} - \frac{K}{m}\boldsymbol{h}_{m,\{k\},i}\right)^{\top}\boldsymbol{w}^k \\
&= \frac{1}{n_m}\sum_{i=1}^{n_m}\sum_{k=1}^{K}\left(\overline{\boldsymbol{h}}_{m,\bullet,i} - \frac{K}{m\binom{K}{m}}\boldsymbol{h}_{m,\{k\},i}\right)^{\top}\boldsymbol{w}^k
\end{aligned}$$

Furthermore, from the AM-GM inequality (e.g., see Lemma A.2 of (Zhu et al., 2021)), we know that for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^K$ and any $c_{3,m} > 0$,

$$\boldsymbol{u}^\top \boldsymbol{v} \leq \frac{c_{3,m}}{2}\|\boldsymbol{u}\|_2^2 + \frac{1}{2c_{3,m}}\|\boldsymbol{v}\|_2^2 \tag{31}$$

where the above AM-GM inequality becomes an equality when $c_{3,m}\boldsymbol{u} = \boldsymbol{v}$. Thus letting $\boldsymbol{u} = \boldsymbol{w}^k$ and $\boldsymbol{v} = \left(\overline{\boldsymbol{h}}_{m,\bullet,i} - \frac{K}{m\binom{K}{m}}\boldsymbol{h}_{m,\{k\},i}\right)^\top$ and applying the AM-GM inequality, we further have:

$$\begin{aligned}
&\gamma_{1,m}^{-1}(g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}) - c_{2,m}) \\
&\geq \frac{1}{n_m}\sum_{i=1}^{n_m}\sum_{k=1}^{K}\left(\overline{\boldsymbol{h}}_{m,\bullet,i} - \frac{K}{m\binom{K}{m}}\boldsymbol{h}_{m,\{k\},i}\right)^\top \boldsymbol{w}^k \qquad (32)\\
&\geq \frac{1}{n_m}\sum_{i=1}^{n_m}\sum_{k=1}^{K}\left(-\frac{c_{3,m}}{2}\|\boldsymbol{w}^k\|_2^2 - \frac{1}{2c_{3,m}}\|\overline{\boldsymbol{h}}_{m,\bullet,i} - \frac{K}{m\binom{K}{m}}\boldsymbol{h}_{m,\{k\},i}\|_2^2\right) \\
&= \frac{1}{n_m}\sum_{i=1}^{n_m}\sum_{k=1}^{K} -\frac{c_{3,m}}{2}\|\boldsymbol{w}^k\|_2^2 - \frac{1}{n_m}\sum_{i=1}^{n_m}\sum_{k=1}^{K}\frac{1}{2c_{3,m}}\|\overline{\boldsymbol{h}}_{m,\bullet,i} - \frac{K}{m\binom{K}{m}}\boldsymbol{h}_{m,\{k\},i}\|_2^2 \\
&= -\frac{c_{3,m}}{2}\|\boldsymbol{W}\|_F^2 - \frac{1}{2c_{3,m}n_m}\sum_{i=1}^{n_m}\sum_{k=1}^{K}\|\overline{\boldsymbol{h}}_{m,\bullet,i} - \frac{K}{m\binom{K}{m}}\boldsymbol{h}_{m,\{k\},i}\|_2^2 \\
&= -\frac{c_{3,m}}{2}\|\boldsymbol{W}\|_F^2 - \frac{1}{2c_{3,m}n_m}\sum_{i=1}^{n_m}\left(K\|\overline{\boldsymbol{h}}_{m,\bullet,i}\|_2^2 + \left(\frac{K}{m\binom{K}{m}}\right)^2\left(\sum_{k=1}^{K}\|\boldsymbol{h}_{m,\{k\},i}\|_2^2\right)\right. \\
&\qquad\qquad \left. - 2K\langle\overline{\boldsymbol{h}}_{m,\bullet,i}, \overline{\boldsymbol{h}}_{m,\bullet,i}\rangle\right) \\
&= -\frac{c_{3,m}}{2}\|\boldsymbol{W}\|_F^2 - \frac{1}{2c_{3,m}n_m}\sum_{i=1}^{n_m}\left(\left(\frac{K}{m\binom{K}{m}}\right)^2\left(\sum_{k=1}^{K}\|\boldsymbol{h}_{m,\{k\},i}\|_2^2\right) - K\|\overline{\boldsymbol{h}}_{m,\bullet,i}\|_2^2\right) \\
&= -\frac{c_{3,m}}{2}\|\boldsymbol{W}\|_F^2 - \frac{\left(\frac{K}{m\binom{K}{m}}\right)^2}{2c_{3,m}n_m}\sum_{i=1}^{n_m}\left(\sum_{k=1}^{K}\|\boldsymbol{h}_{m,\{k\},i}\|_2^2 - K\|\overline{\boldsymbol{h}}_{m,\bullet,i}\|_2^2\right) \\
&\geq -\frac{c_{3,m}}{2}\|\boldsymbol{W}\|_F^2 - \frac{\left(\frac{K}{m\binom{K}{m}}\right)^2}{2c_{3,m}n_m}\sum_{i=1}^{n_m}\sum_{k=1}^{K}\|\boldsymbol{h}_{m,\{k\},i}\|_2^2 \qquad (33)\\
&= -\frac{c_{3,m}}{2}\|\boldsymbol{W}\|_F^2 - \frac{\left(\frac{K}{m\binom{K}{m}}\right)^2}{2c_{3,m}n_m}\left(\|\boldsymbol{H}_m\boldsymbol{D}_m\|_F^2\right) \qquad (34)\\
&= -\frac{c_{3,m}}{2}\|\boldsymbol{W}\|_F^2 - \frac{\left(\frac{K}{m\binom{K}{m}}\right)^2\binom{K-2}{m-1}}{2c_{3,m}n_m}\left(\|\boldsymbol{H}_m\|_F^2\right) \qquad \text{(by Lemma C.7)} \\
&= -\frac{c_{3,m}}{2}\|\boldsymbol{W}\|_F^2 - \frac{\kappa_m}{2c_{3,m}n_m}\left(\|\boldsymbol{H}_m\|_F^2\right),
\end{aligned}$$

where we let $\boldsymbol{D}_m = \text{diag}(\boldsymbol{Y}_m^\top, \cdots, \boldsymbol{Y}_m^\top) \in \mathbb{R}^{(n_m*\binom{K}{m})\times(n_m*K)}$ and $\boldsymbol{Y}_m \in \mathbb{R}^{K\times\binom{K}{m}}$ is the many-hot label matrix defined as follows[9]:

$$\boldsymbol{Y}_m = \left[\boldsymbol{y}_{S_{m,k}}\right]_{k\in\binom{K}{m}}.$$

The first Inequality (32) is tight whenever conditions mentioned in Lemma C.8 are satisfied and the second inequality is

[9]See Appendix C.1.1 for definition of the $S_{m,k}$ notation

tight if and only if

$$c_{3,m}\boldsymbol{w}^k = \left(\frac{K}{m\binom{K}{m}}\boldsymbol{h}_{m,\{k\},i} - \overline{\boldsymbol{h}}_{m,\bullet,i}\right) \quad \forall k \in [K], \quad i \in [n_m]. \tag{35}$$

Therefore, we have

$$g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}) - c_{2,m} \geq -\gamma_{1,m}\frac{c_{3,m}}{2}\|\boldsymbol{W}\|_F^2 - \gamma_{1,m}\frac{\kappa_m}{2c_{3,m}n_m}\left(\|\boldsymbol{H}_m\|_F^2\right). \tag{36}$$

The last Inequality (33) achieves its equality if and only if

$$\overline{\boldsymbol{h}}_{m,\bullet,i} = \boldsymbol{0}, \quad \forall i \in [n_m]. \tag{37}$$

Plugging this into (Equation (35)), we have

$$\begin{aligned}
c_{3,m}\boldsymbol{w}^k &= \frac{K}{m\binom{K}{m}}\boldsymbol{h}_{m,\{k\},i} \\
\implies c_{3,m}^2 &= \frac{\left(\frac{K}{m\binom{K}{m}}\right)^2 \sum_{i=1}^n \sum_{k=1}^K \|\boldsymbol{h}_{m,\{k\},i}\|_F^2}{n_m \sum_{k=1}^K \|\boldsymbol{w}^k\|_2^2} \\
&= \frac{\left(\frac{K}{m\binom{K}{m}}\right)^2 \binom{K-2}{m-1}\|\boldsymbol{H}_m\|_F^2}{n_m\|\boldsymbol{W}\|_F^2} \\
&= \frac{\kappa_m}{n_m}\frac{\|\boldsymbol{H}_m\|_F^2}{\|\boldsymbol{W}\|_F^2} \\
\implies c_{3,m} &= \sqrt{\frac{\kappa_m}{n_m}}\frac{\|\boldsymbol{H}_m\|_F}{\|\boldsymbol{W}\|_F} \\
\implies c_{3,m}^2 &= \frac{\kappa_m}{n_m}\frac{\|\boldsymbol{H}_m\|_F^2}{\|\boldsymbol{W}\|_F^2}.
\end{aligned}$$

Now, note that by our definition of ρ and Lemma C.1, we get

$$\|\boldsymbol{H}\|_F^2 = \frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}\rho. \tag{38}$$

Recall from the state of the lemma that we defined $\kappa_m := \left(\frac{K/m}{\binom{K}{m}}\right)^2\binom{K-2}{m-1}$ and that $\gamma_{1,m} := \frac{1}{1+c_{1,m}}\frac{m}{K-m}$. Thus, continuing from Inequality (36), we have

$$\gamma_{1,m}^{-1}(g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}) - c_{2,m}) \geq -\frac{c_{3,m}}{2}\|\boldsymbol{W}\|_F^2 - \frac{\kappa_m}{2c_{3,m}n_m}\|\boldsymbol{H}_m\|_F^2.$$

Next, let $Q > 0$ be an arbitrary constant, to be determined later such that

$$\gamma_{1,m} = \frac{1}{N_m}Qc_{3,m}^{-1}\frac{\|\boldsymbol{H}_m\|_F^2}{\|\boldsymbol{W}\|_F^2}, \qquad \forall m \in \{1,\ldots,M\}. \tag{39}$$

A remark is in order: at this current point in the proof, it is unclear that such a Q exists. However, in Equation (42), we derive an explicit formula for Q such that Equation (39) holds. Now, given Equation (39), we have

$$g_m(\boldsymbol{W}\boldsymbol{H}_m + \boldsymbol{b}) - c_{2,m} \geq \frac{1}{N_m}Q\left(-\frac{1}{2}\|\boldsymbol{H}_m\|_F^2 - \frac{1}{2}\|\boldsymbol{H}_m\|_F^2\right) = -\frac{1}{N_m}Q\|\boldsymbol{H}_m\|_F^2.$$

Let $\Gamma_2 := \sum_{m=1}^M \frac{N_m}{N}c_{2,m}$. Summing the above inequality on both side over $m = 1,\ldots,M$ according to Equation (16), we have

$$g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}) - \Gamma_2 \geq -\frac{1}{N}Q\sum_{m=1}^M \|\boldsymbol{H}_m\|_F^2 = -\frac{1}{N}Q\|\boldsymbol{H}\|_F^2 = -\frac{1}{N}Q\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}\rho. \tag{40}$$

where the last equality is due to Equation (38). Now, we derive the expression for Q, which earlier we set to be arbitrary.

From Equation (39), we have

$$\frac{1}{1+c_{1,m}}\frac{m}{K-m} = \gamma_{1,m} = \frac{1}{N_m} Q \frac{\sqrt{n_m}}{\sqrt{\kappa_m}} \frac{\|\boldsymbol{H}_m\|_F}{\|\boldsymbol{W}\|_F}. \tag{41}$$

Rearranging and using the fact that $N_m = \binom{K}{m} n_m$, we have

$$\binom{K}{m}\frac{1}{1+c_{1,m}}\frac{m}{K-m}\sqrt{\kappa_m n_m} = Q\frac{\|\boldsymbol{H}_m\|_F}{\|\boldsymbol{W}\|_F}.$$

Squaring both side, we have

$$\left(\frac{1}{1+c_{1,m}}\frac{m}{K-m}\right)^2 \kappa_m n_m \binom{K}{m}^2 = Q^2 \frac{\|\boldsymbol{H}_m\|_F^2}{\|\boldsymbol{W}\|_F^2}.$$

Summing over $m = 1, \dots, M$, we have

$$\sum_{m=1}^{M}\left(\frac{1}{1+c_{1,m}}\frac{m}{K-m}\right)^2 \kappa_m n_m \binom{K}{m}^2 = Q^2 \sum_{m=1}^{M}\frac{\|\boldsymbol{H}_m\|_F^2}{\|\boldsymbol{W}\|_F^2} = Q^2 \frac{\|\boldsymbol{H}\|_F^2}{\|\boldsymbol{W}\|_F^2} = Q^2 \frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}$$

Thus, we conclude that

$$Q = \sqrt{\frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}}\sqrt{\sum_{m=1}^{M}\left(\frac{1}{1+c_{1,m}}\frac{m}{K-m}\right)^2 \kappa_m n_m \binom{K}{m}^2}. \tag{42}$$

Substituting Q into Equation (41), we get

$$\frac{1}{1+c_{1,m}}\frac{m}{K-m} = \frac{1}{\binom{K}{m}\sqrt{\kappa_m n_m}}\frac{\|\boldsymbol{H}_m\|_F}{\|\boldsymbol{W}\|_F}\sqrt{\frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}}\sqrt{\sum_{m'=1}^{M}\left(\frac{1}{1+c_{1,m'}}\frac{m'}{K-m'}\right)^2 \kappa_{m'} n_{m'} \binom{K}{m'}^2}. \tag{43}$$

Finally substituting Q into Equation (40),

$$g(\boldsymbol{W}\boldsymbol{H}+\boldsymbol{b}) - \Gamma_2 \geq -\frac{1}{N}\sqrt{\sum_{m=1}^{M}\left(\frac{1}{1+c_{1,m}}\frac{m}{K-m}\right)^2 \kappa_m n_m \binom{K}{m}^2}\sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}}}\rho.$$

which concludes the proof. □

As a sanity check of the validity of Lemma C.2, we briefly revisit the M-clf case where $M = 1$. We show that our Lemma C.2 recovers (Zhu et al., 2021) Lemma B.3 as a special case. Now, from the definition of κ_m, we have that $\kappa_1 = 1$. Thus, the above expression reduces to simply

$$Q = \sqrt{\frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}} n_1}}\frac{1}{1+c_{1,1}}\frac{1}{K-1}.$$

The lower bound from Lemma C.2 reduces to simply

$$g_1(\boldsymbol{W}\boldsymbol{H}_1+\boldsymbol{b}) - \gamma_{2,1} \geq -Q\rho\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}}} = -\frac{1}{1+c_{1,1}}\frac{1}{K-1}\rho\sqrt{\frac{\lambda_{\boldsymbol{W}}}{\lambda_{\boldsymbol{H}} n_1}}$$

which exactly matches that of (Zhu et al., 2021) Lemma B.3.

Next, we show that the lower bound in Inequality (22) is attained if and only if $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ satisfies the following conditions.

Lemma C.3. *Under the same assumptions of Lemma C.2, the lower bound in Inequality (22) is attained for a critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of Problem (4) if and only if all of the following hold:*

(I) $\|\boldsymbol{w}^1\|_2 = \|\boldsymbol{w}^2\|_2 = \cdots = \|\boldsymbol{w}^K\|_2$, *and* $\boldsymbol{b} = b\mathbf{1}$,

(II) $\frac{1}{\binom{K}{m}}\sum_{k=1}^{\binom{K}{m}} \boldsymbol{h}_{m,k,i} = \mathbf{0}$, *and* $\sqrt{\frac{\binom{K-2}{m-1}}{n_m}}\frac{\|\boldsymbol{H}_m\|_F}{\|\boldsymbol{W}\|_F}\boldsymbol{w}^k = \sum_{\ell:k\in S_{m,\ell}} \boldsymbol{h}_{m,\ell,i}, \forall m \in [M], k \in [K], i \in [n_m]$,

(III) $\boldsymbol{W}^\top\boldsymbol{W} = \frac{\rho}{K-1}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)$

(IV) *There exist unique positive real numbers $C_1, C_2, \dots, C_M > 0$ such that the following holds:*

$$\boldsymbol{h}_{1,k,i} = C_1 \boldsymbol{w}^{\ell} \quad \text{when } S_{1,k} = \{\ell\},\ \ell \in [K], \qquad \textit{(Multiplicity} = 1 \textit{ Case)}$$
$$\boldsymbol{h}_{m,k,i} = C_m \textstyle\sum_{\ell \in S_{m,k}} \boldsymbol{w}^{\ell} \quad \text{when } m > 1. \qquad \textit{(Multiplicity} > 1 \textit{ Case)}$$

(See Appendix C.1.1 for the notation $S_{m,k}$.)

(V) *There exists $c_{1,1}, c_{1,2}, \dots, c_{1,M} > 0$ such that*

$$\frac{1}{1+c_{1,m}} \frac{m}{K-m} = \frac{1}{\binom{K}{m}\sqrt{\kappa_m n_m}} \frac{\sqrt{\frac{\binom{K}{m} n_m m(K-m)(K-1)}{K}} * log(\frac{K-m}{m} c_{1,m})}{\rho} \cdot \sqrt{\frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}} \sqrt{\sum_{m'=1}^{M} \left(\frac{1}{1+c_{1,m'}} \frac{m'}{K-m'}\right)^2 \kappa_{m'} n_{m'} \binom{K}{m'}^2} \tag{44}$$

The proof of Lemma C.3 utilizes the conditions in Lemma C.8, and the conditions in Equation (35) and Equation (37) during the proof of Lemma C.2.

Proof. Similar as in the proof of Lemma C.2, define $\boldsymbol{h}_{m,\{k\},i} := \sum_{\ell: k \in S_{m,\ell}} \boldsymbol{h}_{m,\ell,i}$ and $\overline{\boldsymbol{h}}_{m,\bullet,i} := \frac{1}{\binom{K}{m}} \sum_{k=1}^{\binom{K}{m}} \boldsymbol{h}_{m,k,i}$. From the proof of Lemma C.2, the lower bound is attained whenever the conditions in Equation (35) and Equation (37) hold, which respectively is equivalent to the following:

$$\overline{\boldsymbol{h}}_{m,\bullet,i} = \boldsymbol{0} \quad \text{and}$$
$$\sqrt{\frac{\binom{K-2}{m-1}}{n_m}} \frac{\|\boldsymbol{H}_m\|_F}{\|\boldsymbol{W}\|_F} \boldsymbol{w}^k = \boldsymbol{h}_{m,\{k\},i}, \forall m \in [M\ k \in [K], i \in [n_m], \tag{45}$$

In particular, the $m = 1$ case further implies

$$\sum_{k=1}^{K} \boldsymbol{w}^k = \boldsymbol{0}.$$

Next, under the condition described in Equation (45), when $m = 1$, if we want Inequality (22) to become an equality, we only need Inequality (32) to become an equality when $m = 1$, which is true if and only if conditions in Lemma C.8 holds for $\boldsymbol{z}_{1,k,i} = \boldsymbol{W}\boldsymbol{h}_{1,k,i} \forall i \in [n_m]$ and $\forall k \in [K]$. First let $[\boldsymbol{z}_{1,k,i}]_j = \boldsymbol{h}_{1,k,i}^{\top} \boldsymbol{w}^j + b_j$, we would have:

$$\sum_{j=1}^{K} [\boldsymbol{z}_{1,k,i}]_j = K\bar{b} \quad \text{and} \quad K[\boldsymbol{z}_{1,k,i}] = c_{3,1}\left(K\|\boldsymbol{w}^k\|_2^2\right) + Kb_k. \tag{46}$$

We pick $\gamma_{1,1} = 1\beta$, where β is defined in (60), to be the same for all $k \in [K]$ in multiplicity one, which also means to pick $\frac{1}{\beta} - (K-1)$ to be the same for all $k \in [K]$ within one multiplicity. Note under the first (*in-group equality*) and second *out-group equality* condition in Lemma C.8 and utilize the condition (46), we have

$$\begin{aligned}
\frac{1}{\beta} - (K-1) &= \frac{(K-1)\exp(z_{out}) + \exp(z_{in})}{\exp(z_{out})} - (K-1) \\
&= (K-1) + \exp(z_{in} - z_{out}) - (K-1) \\
&= \exp(z_{in} - z_{out}) \\
&= \exp\left(\frac{Kz_{in} - z_{in} - (K-1)z_{out}}{K-1}\right) \\
&= \exp\left(\frac{Kz_{in} - \sum_j z_j}{K-1}\right)
\end{aligned}$$

$$= \left(\exp\left(\frac{\sum_j z_j - Kz_{in}}{K-1}\right)\right)^{-1}$$
$$= \left(\exp\left(\frac{\sum_j z_j - Kz_k}{K-1}\right)\right)^{-1}$$
$$= \exp\left(\frac{K}{K-1}\left(\bar{b} - c_{3,1}\|\boldsymbol{w}^k\|_2^2\right) - b_k\right)^{-1}$$

Since the scalar $\gamma_{1,1}$ is picked the same for one m, but the above equality we have

$$c_{3,1}\|\boldsymbol{w}^k\|_2^2 - b_k = c_{3,1}\|\boldsymbol{w}^\ell\|_2^2 - b_\ell \quad \forall \ell \neq k. \tag{47}$$

this directly follows after Equation (29) from the proof in Lemma B.4 of (Zhu et al., 2021) to conclude all the conditions except the scaled average condition, which we address next. To this end, we use the second condition in (45) which asserts for $m \geq 2$ that:

$$\sqrt{\frac{n_m}{\binom{K-2}{m-1}}}\frac{\|\boldsymbol{W}\|_F}{\|\boldsymbol{H}_m\|_F}\boldsymbol{h}_{m,\{k\},i} = \boldsymbol{w}^k$$
$$\implies \sqrt{\frac{n_1}{\binom{K-2}{1-1}}}\frac{\|\boldsymbol{W}\|_F}{\|\boldsymbol{H}_1\|_F}\boldsymbol{h}_{1,\{k\},i} = \sqrt{n_1}\frac{\|\boldsymbol{W}\|_F}{\|\boldsymbol{H}_1\|_F}\boldsymbol{h}_{1,k,i} = \boldsymbol{w}^k = \sqrt{\frac{n_m}{\binom{K-2}{m-1}}}\frac{\|\boldsymbol{W}\|_F}{\|\boldsymbol{H}_m\|_F}\boldsymbol{h}_{m,\{k\},i}$$
$$\implies \boldsymbol{h}_{m,\{k\},i} = \sqrt{\frac{n_1\binom{K-2}{m-1}}{n_m}}\frac{\|\boldsymbol{H}_m\|_F}{\|\boldsymbol{H}_1\|_F}\boldsymbol{h}_{1,k,i} = c_{h,m}\boldsymbol{h}_{1,k,i} \tag{48}$$

where $c_{h,m} = \sqrt{\frac{n_1\binom{K-2}{m-1}}{n_m}}\frac{\|\boldsymbol{H}_m\|_F}{\|\boldsymbol{H}_1\|_F}$. Let $\widetilde{\boldsymbol{H}}_1$ (resp. $\widetilde{\boldsymbol{H}}_m$) be the block-submatrix corresponding to the first K columns of $\boldsymbol{H}_1$ (resp. first $\binom{K}{m}$ columns of $\boldsymbol{H}_m$). Define $\widetilde{\boldsymbol{Y}}_1$ and $\widetilde{\boldsymbol{Y}}_m$ similarly. Then, Equation (48) can be equivalently stated in the following matrix form:

$$c_{h,m}\widetilde{\boldsymbol{H}}_1 = \widetilde{\boldsymbol{H}}_m\widetilde{\boldsymbol{Y}}_m^\top$$

Let $\boldsymbol{P}_m = \widetilde{\boldsymbol{Y}}_m^\top(\widetilde{\boldsymbol{Y}}_m^\top)^\dagger$ be the projection matrix onto the subspace $\widetilde{\boldsymbol{Y}}_m$, then we have

$$\widetilde{\boldsymbol{H}}_m\boldsymbol{P}_m = \widetilde{\boldsymbol{H}}_m\widetilde{\boldsymbol{Y}}_m^\top(\widetilde{\boldsymbol{Y}}_m^\top)^\dagger = c_{h,m}\widetilde{\boldsymbol{H}}_1(\widetilde{\boldsymbol{Y}}_m^\top)^\dagger,$$

which simplifies as

$$\widetilde{\boldsymbol{H}}_m\boldsymbol{P}_m = c_{h,m}\widetilde{\boldsymbol{H}}_1(\widetilde{\boldsymbol{Y}}_m^\top)^\dagger.$$

Applying Lemma C.4 to the LHS and Lemma C.6 to the RHS we have

$$\widetilde{\boldsymbol{H}}_m = c_{h,m}\widetilde{\boldsymbol{H}}_1(\tau_m\widetilde{\boldsymbol{Y}}_m + \eta_m\boldsymbol{\Theta})$$
$$\widetilde{\boldsymbol{H}}_m = c_{h,m}\cdot\tau_m\widetilde{\boldsymbol{H}}_1\widetilde{\boldsymbol{Y}}_m$$

and substituting $\widetilde{\boldsymbol{H}}_1$ using the relationship between $\widetilde{\boldsymbol{H}}_1$ and $\boldsymbol{W}$, namely, $c_{h,1}\cdot(\boldsymbol{W}^\top) = \widetilde{\boldsymbol{H}}_1$, we now have

$$\widetilde{\boldsymbol{H}}_m = c_{h,m}\cdot\tau_m\cdot c_{1,m}(\boldsymbol{W}^\top\widetilde{\boldsymbol{Y}}_m)$$

where

$$C_m = c_{h,m}\cdot c_{h,1}\cdot\tau_m$$
$$= \sqrt{\frac{n_1}{n_m\binom{K-2}{m-1}}}\frac{\|\boldsymbol{H}_m\|_F}{\|\boldsymbol{H}_1\|_F}\cdot\sqrt{\frac{1}{n_1}}\frac{\|\boldsymbol{H}_1\|_F}{\|\boldsymbol{W}\|_F}$$
$$= \sqrt{\frac{1}{n_m\binom{K-2}{m-1}}}\frac{\|\boldsymbol{H}_m\|_F}{\|\boldsymbol{W}\|_F}$$

This proves (IV). Finally, to proof (V), following from Equation (43) in the proof of Lemma C.2, we only need to further simplify $\|\boldsymbol{H}_m\|_F$.

We first establish a connection the between $\|\boldsymbol{W}\boldsymbol{H}_m\|_F^2$ and $\|\boldsymbol{H}_m\|_F^2$. By definition of Frobenius norm and the last layer

classifier $\boldsymbol{W}$ is an ETF with expression $\boldsymbol{W}^\top\boldsymbol{W} = \frac{\rho}{K-1}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)$, we have

$$\begin{aligned}\|\boldsymbol{W}\boldsymbol{H}_m\|_F^2 &= tr(\boldsymbol{W}\boldsymbol{H}_m\boldsymbol{H}_m^\top\boldsymbol{W}^\top)\\ &= \frac{\rho}{K-1}tr(\boldsymbol{H}_m\boldsymbol{H}_m^\top(\boldsymbol{I}_K - \mathbf{1}_K\mathbf{1}_K^\top))\\ &= \frac{\rho}{K-1}\|\boldsymbol{H}_m\|_F^2\end{aligned}$$

Since variability within feature already collapse at this point, we can express $\|\boldsymbol{W}\boldsymbol{H}_m\|_F^2$ in terms of $z_{m,in}$ and $z_{m,out}$:

$$\|\boldsymbol{W}\boldsymbol{H}_m\|_F^2 = \frac{\rho}{K-1}\|\boldsymbol{H}_m\|_F^2 = \binom{K}{m}n_m(mz_{m,in}^2 + (K-m)z_{m,out}^2).$$

From the second equality we could express $\|\boldsymbol{H}_m\|$ as:

$$\|\boldsymbol{H}_m\|_F = \sqrt{\frac{\binom{K}{m}n_m(K-1)}{\rho}(mz_{m,in}^2 + (K-m)z_{m,out}^2)} \tag{49}$$

Recall from Lemma C.8, we have the following equation to express $z_{m,in}$ and $z_{m,out}$

$$z_{in} - z_{m,out} = log(\frac{K-m}{m}c_{1,m}).$$

As column sum of $\boldsymbol{H}_m$ equals to $\mathbf{0}$, the column sum of $\boldsymbol{W}\boldsymbol{H}_m$ also equals to $\mathbf{0}$ as well. Given the extra constrain of *in-group equality* and *out-group equality* from Lemma C.8, it yields:

$$mz_{m,in} + (K-m)z_{m,out} = 0$$

Now we could solve for $z_{m,in}$ and $z_{m,out}$ in terms of $c_{1,m}$

$$\begin{aligned}z_{m,in} &= \frac{K-m}{K}log\left(\frac{K-m}{m}c_{1,m}\right)\\ z_{m,out} &= -\frac{m}{K}log\left(\frac{K-m}{m}c_{1,m}\right)\end{aligned}$$

Substituting above expression for $z_{m,in}$ and $z_{m,out}$ into Equation (49), we have

$$\|\boldsymbol{H}_m\|_F = \sqrt{\frac{\binom{K}{m}n_m m(K-m)(K-1)}{\rho K}}\log(\frac{K-m}{m}c_{1,m})$$

Finally, we substituting the above expression of $\|\boldsymbol{H}_m\|_F$ in to Equation (43) and conclude:

$$\frac{1}{1+c_{1,m}}\frac{m}{K-m} = \frac{1}{\binom{K}{m}\sqrt{\kappa_m n_m}}\frac{\sqrt{\frac{\binom{K}{m}n_m m(K-m)(K-1)}{K}} * log(\frac{K-m}{m}c_{1,m})}{\rho}\cdot$$
$$\sqrt{\frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}}\sqrt{\sum_{m'=1}^{M}\left(\frac{1}{1+c_{1,m'}}\frac{m'}{K-m'}\right)^2\kappa_{m'}n_{m'}\binom{K}{m'}^2}.$$

Revisiting and combining results from (IV) and (V), we have the scaled-average constant C_m to be

$$\begin{aligned}C_m &= \sqrt{\frac{1}{n_m\binom{K-2}{m-1}}}\frac{\sqrt{\frac{\binom{K}{m}n_m m(K-m)(K-1)}{\rho K}}\log(\frac{K-m}{m}c_{1,m})}{\|\boldsymbol{W}\|_F}\\ &= \frac{K-1}{\rho}\log(\frac{K-m}{m}c_{1,m})\end{aligned}$$

where $c_{1,m}$ is a solution to the system of equation Equation (44). Note that Equation (44) hold for all m. Thus, we could construct a system of equation whose variable are $c_{1,1}, \cdots, c_{1,m}$. Even when missing some multiplicity data, we sill have same number of variable $c_{1,m}$ as equations. We numerically verifies that under various of UFM model setting (i.e. different

number of class and different number of multiplicities), $c_{1,m}$ does solves the above system of equation. □

Lemma C.4. *Let let $\boldsymbol{P}_m = \widetilde{\boldsymbol{Y}}_m^\top (\widetilde{\boldsymbol{Y}}_m^\top)^\dagger$ be the projection matrix then we have, $\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m = \widetilde{\boldsymbol{H}}_m$*

Proof. As $\boldsymbol{P}_m$ is a projection matrix, we have that $\|\widetilde{\boldsymbol{H}}_m\|_F^2 = \|\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m\|_F^2$ if and only if $\widetilde{\boldsymbol{H}}_m = \widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m$. So it is suffice to show that $\|\widetilde{\boldsymbol{H}}_m\|_F^2 = \|\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m\|_F^2$. We denote $\boldsymbol{W}\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m$ as the projection solution and by Lemma C.5 we have that

$$\boldsymbol{W}\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m = \boldsymbol{W}\widetilde{\boldsymbol{H}}_m,$$

which further implies that the projection solution $\boldsymbol{W}\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m$ also solves g

$$g(\boldsymbol{W}\widetilde{\boldsymbol{H}}_m, \widetilde{\boldsymbol{Y}}) = g(\boldsymbol{W}\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m, \widetilde{\boldsymbol{Y}}).$$

When it comes to the regularization term, by minimum norm projection property, we have $\|\widetilde{\boldsymbol{H}}_m\|_F^2 \geq \|\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m\|_F^2$. Note if the projection solution results in a strictly smaller frobenious norm i.e. $\|\widetilde{\boldsymbol{H}}_m\|_F^2 > \|\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m\|_F^2$, then $f(\boldsymbol{W}, \widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m, \boldsymbol{b}) < f(\boldsymbol{W}, \widetilde{\boldsymbol{H}}_m, \boldsymbol{b})$, this contradict the assumption that $\boldsymbol{Z}_m = \boldsymbol{W}\widetilde{\boldsymbol{H}}_m$ is the global solutions of f. Thus, the only possible outcomes is that $\|\widetilde{\boldsymbol{H}}_m\|_F^2 = \|\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m\|_F^2$, which complete the proof. □

Lemma C.5. *We want to show that the optimal global solution of f, $\boldsymbol{W}\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m$, is the same after projected on to the space of $\widetilde{\boldsymbol{Y}}_m$, i.e., $\boldsymbol{W}\widetilde{\boldsymbol{H}}_m \boldsymbol{P}_m = \boldsymbol{W}\widetilde{\boldsymbol{H}}_m$*

Proof. Let $\boldsymbol{Z}_m = \boldsymbol{W}\widetilde{\boldsymbol{H}}_m$ denote the global minimizer of the loss function f for an arbitrary multiplicity m. Since $\boldsymbol{Z}_m$ has both the in-group and out-group equality property, we could express it as

$$\boldsymbol{Z}_m = d_1 \widetilde{\boldsymbol{Y}}_m + d_2 \boldsymbol{\Theta},$$

for some constant d_1, d_2, and all-one matrix $\boldsymbol{\Theta}$ of proper dimension. Note that it is suffice to show that $\boldsymbol{Z}_m$ lives in the subspace of which the projection matrix $\boldsymbol{P}_m$ projects onto. By Lemma C.6, as $(\widetilde{\boldsymbol{Y}}_m^\top)^\dagger$ is the Moore–Penrose pseudo-inverse of $\widetilde{\boldsymbol{Y}}_m^\top$ by, we could rewrite $\boldsymbol{P}_m$ as

$$\begin{aligned}
\boldsymbol{P}_m &= \widetilde{\boldsymbol{Y}}_m^\top (\widetilde{\boldsymbol{Y}}_m^\top)^\dagger \\
&= \widetilde{\boldsymbol{Y}}_m^\top \left(\widetilde{\boldsymbol{Y}}_m \widetilde{\boldsymbol{Y}}_m^\top\right)^\dagger \widetilde{\boldsymbol{Y}}_m \\
&= \widetilde{\boldsymbol{Y}}_m^\top \left(\widetilde{\boldsymbol{Y}}_m \widetilde{\boldsymbol{Y}}_m^\top\right)^{-1} \widetilde{\boldsymbol{Y}}_m.
\end{aligned}$$

Hence we can see that the subspace which $\boldsymbol{P}_m$ projects onto is spanned by columns/rows of $\widetilde{\boldsymbol{Y}}_m$. In order to show that $\boldsymbol{Z}_m = d_1 \widetilde{\boldsymbol{Y}}_m + d_2 \boldsymbol{\Theta}$ is in the subspace spanned by columns of $\widetilde{\boldsymbol{Y}}_m$, it is suffice to see that the columns sum of $\widetilde{\boldsymbol{Y}}_m = \frac{m}{K}\binom{K}{m}\mathbf{1}$. Thus, we finished the proof. □

Lemma C.6. *The Moore-Penrose pseudo-inverse of $\widetilde{\boldsymbol{Y}}_m^\top$ has the form $(\widetilde{\boldsymbol{Y}}_m^\top)^\dagger = \tau_m \widetilde{\boldsymbol{Y}}_m + \eta_m \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the all-one matrix with proper dimension and $\tau_m = \frac{a+c}{bc}$, $\eta_m = -\frac{a}{bc}$, for $a = \frac{m-1}{k-1}\binom{K-1}{m-1}$, $b = \frac{m}{k}\binom{K}{m}$, $c = \frac{m}{k-1}\binom{K-1}{m}$.*

Proof. First, we have the column sum of $\widetilde{\boldsymbol{Y}}_m$ can be written as a constant times an all-one vector

$$\sum_{j}^{\binom{K}{m}} (\widetilde{\boldsymbol{Y}}_m)_{:,j} = \frac{m}{K}\binom{K}{m}\mathbf{1} \tag{50}$$

This property could be seen from a probabilistic perspective. We let $i \in [K]$ be fixed and deterministic, and let $S \subseteq [K]$ be a random subset of size m generating by sampling without replacement. Then

$$Pr\{i \notin S\} = \frac{K-1}{K} \times \frac{K-2}{K-1} \times \cdots \times \frac{K-m}{K-m+1} = \frac{K-m}{K}.$$

This implies that $Pr\{i \in S\} = \frac{m}{K}$ and each entry of the column sum result is exactly $\frac{m}{K}\binom{K}{m}$ as we sum up all $\binom{K}{m}$ columns of $\widetilde{\boldsymbol{Y}}_m$.

Second, the label matrix $\widetilde{\boldsymbol{Y}}_m$ has the property that

$$\widetilde{\boldsymbol{Y}}_m\widetilde{\boldsymbol{Y}}_m^\top = \begin{bmatrix} b & & a \\ & \ddots & \\ a & & b \end{bmatrix}, \quad \widetilde{\boldsymbol{Y}}_m(\boldsymbol{\Theta} - \widetilde{\boldsymbol{Y}}_m^\top) = \begin{bmatrix} 0 & & c \\ & \ddots & \\ c & & 0 \end{bmatrix}, \tag{51}$$

where $a = \frac{m-1}{k-1}\binom{K-1}{m-1}$, $\quad b = \frac{m}{k}\binom{K}{m}$, $\quad c = \frac{m}{k-1}\binom{K-1}{m}$. Again, from a probabilistic perspective, any off-diagonal entry of the product $\widetilde{\boldsymbol{Y}}_m\widetilde{\boldsymbol{Y}}_m^\top$ is equal to $(\widetilde{\boldsymbol{Y}}_m)_{i,:}(\widetilde{\boldsymbol{Y}}_m)_{i',:}^\top$, for $i \neq i'$. Note that $(\widetilde{\boldsymbol{Y}}_m)_{i,:}$ is a row vector of length $\binom{K}{m}$, whose entry are either 0 or 1 and the results of $(\widetilde{\boldsymbol{Y}}_m)_{i,:}(\widetilde{\boldsymbol{Y}}_m)_{i',:}^\top$ would only increase by one if both $(\widetilde{\boldsymbol{Y}}_m)_{i,j} = 1$ and $(\widetilde{\boldsymbol{Y}}_m)_{i',j}^\top = 1$ for $j \in [\binom{K}{m}]$. From the previous property we know that there is $\frac{m}{K}$ probability that $(\widetilde{\boldsymbol{Y}}_m)_{i,j} = 1$. In addition, conditioned on $(\widetilde{\boldsymbol{Y}}_m)_{i,j} = 1$, there are $\frac{m-1}{K-1}$ probability that $(\widetilde{\boldsymbol{Y}}_m)_{i',j} = 1$. Thus, $a = \frac{m}{K}\frac{m-1}{K-1}\binom{K}{m} = \frac{m-1}{K-1}\binom{K-1}{m-1}$. For similar reasoning, we can see that conditioned on $(\widetilde{\boldsymbol{Y}}_m)_{i,j} = 1$, there are $1 - \frac{m-1}{K-1} = \frac{K-m}{K-1}$ probability that $(\boldsymbol{\Theta}_{i',j} - (\widetilde{\boldsymbol{Y}}_m)_{i',j}) = 1$. Thus, $c = \frac{m}{K}\frac{K-m}{K-1}\binom{K}{m} = \frac{m}{K-1}\binom{K-1}{m}$. For the similar probabilistic argument, it is easy to see that diagonal of $\widetilde{\boldsymbol{Y}}_m\widetilde{\boldsymbol{Y}}_m^\top$ are all $b = \frac{m}{K}\binom{K}{m}$ and diagonal of $\widetilde{\boldsymbol{Y}}_m(\boldsymbol{\Theta} - \widetilde{\boldsymbol{Y}}_m^\top)$ are all 0. Then by the second property (Equation (51)), we are about to cook up a left inverse of $\widetilde{\boldsymbol{Y}}^\top$:

$$\begin{aligned}
\frac{1}{b}\left(\widetilde{\boldsymbol{Y}}\widetilde{\boldsymbol{Y}}^\top - \frac{a}{c}(\widetilde{\boldsymbol{Y}}(\boldsymbol{\Theta} - \widetilde{\boldsymbol{Y}}^\top))\right) &= \boldsymbol{I} \\
\widetilde{\boldsymbol{Y}}\left(\frac{1}{b}\widetilde{\boldsymbol{Y}}^\top - \frac{a}{bc}\boldsymbol{\Theta} + \frac{a}{bc}\widetilde{\boldsymbol{Y}}^\top\right) &= \boldsymbol{I} \\
\widetilde{\boldsymbol{Y}}\left(\frac{a+c}{bc}\widetilde{\boldsymbol{Y}}^\top - \frac{a}{bc}\boldsymbol{\Theta}\right) &= \boldsymbol{I} \\
\left(\frac{a+c}{bc}\widetilde{\boldsymbol{Y}} - \frac{a}{bc}\boldsymbol{\Theta}\right)\widetilde{\boldsymbol{Y}}^\top &= \boldsymbol{I}
\end{aligned}$$

Let, $\tau_m = \frac{a+c}{bc}$, $\eta_m = -\frac{a}{bc}$, then the pseudo-inverse of $\widetilde{\boldsymbol{Y}}^\top$, namely $(\widetilde{\boldsymbol{Y}}^\top)^\dagger$ could be written as

$$(\widetilde{\boldsymbol{Y}}^\top)^\dagger = \tau_m\widetilde{\boldsymbol{Y}} + \eta_m\boldsymbol{\Theta}$$

This inverse is also the Moore–Penrose inverse which is unique since it satisfies that:

$$\widetilde{\boldsymbol{Y}}^\top(\widetilde{\boldsymbol{Y}}^\top)^\dagger\widetilde{\boldsymbol{Y}}^\top = \widetilde{\boldsymbol{Y}}^\top\boldsymbol{I} = \widetilde{\boldsymbol{Y}}^\top \tag{52}$$

$$(\widetilde{\boldsymbol{Y}}^\top)^\dagger\widetilde{\boldsymbol{Y}}^\top(\widetilde{\boldsymbol{Y}}^\top)^\dagger = \boldsymbol{I}(\widetilde{\boldsymbol{Y}}^\top)^\dagger = (\widetilde{\boldsymbol{Y}}^\top)^\dagger \tag{53}$$

$$(\widetilde{\boldsymbol{Y}}^\top(\widetilde{\boldsymbol{Y}}^\top)^\dagger)^\top = \widetilde{\boldsymbol{Y}}^\top(\widetilde{\boldsymbol{Y}}^\top)^\dagger \tag{54}$$

$$((\widetilde{\boldsymbol{Y}}^\top)^\dagger\widetilde{\boldsymbol{Y}}^\top)^\top = (\widetilde{\boldsymbol{Y}}^\top)^\dagger\widetilde{\boldsymbol{Y}}^\top \tag{55}$$

□

Lemma C.7. *We would like to show the following equation holds:*

$$\|\boldsymbol{H}_m\boldsymbol{D}_m\|_F^2 = \binom{K-2}{m-1}\|\boldsymbol{H}_m\|_F^2$$

Proof. Note due to how we construct $\boldsymbol{D}_m$, it is suffice to show that $\|\widetilde{\boldsymbol{H}}_m\widetilde{\boldsymbol{Y}}_m^\top\|_F^2 = \binom{K-2}{m-1}\|\widetilde{\boldsymbol{H}}_m\|_F^2$. Recall the definition that $a = \frac{m-1}{k-1}\binom{K-1}{m-1}$ and $b = \frac{m}{k}\binom{K}{m}$. By unwinding the definition of binomial coefficient and simplifying factorial expressions, we can see that $b - a = \binom{K-2}{m-1}$. Along with the assumption that columns sum of $\widetilde{\boldsymbol{H}}_m$ is $\mathbf{0}$ i.e. $\overline{\boldsymbol{h}}_{m,\bullet,i} = \mathbf{0}, \quad \forall i \in [n_m]$ and

the property described in Equation (51), we have

$$
\begin{aligned}
&\|\widetilde{\boldsymbol{H}}_m\widetilde{\boldsymbol{Y}}_m^\top\|_F^2 = \binom{K-2}{m-1}\|\widetilde{\boldsymbol{H}}_m\|_F^2 \\
\iff &\|\tau_m\widetilde{\boldsymbol{H}}_1\widetilde{\boldsymbol{Y}}_m\widetilde{\boldsymbol{Y}}_m^\top\|_F^2 = \binom{K-2}{m-1}\|\tau_m\widetilde{\boldsymbol{H}}_1\widetilde{\boldsymbol{Y}}_m\|_F^2 \\
\iff &\tau_m^2(b-a)^2\|\widetilde{\boldsymbol{H}}_1\|_F^2 = \tau_m^2(b-a)\|\widetilde{\boldsymbol{H}}_1\widetilde{\boldsymbol{Y}}_m\|_F^2 \\
\iff &(b-a)\|\widetilde{\boldsymbol{H}}_1\|_F^2 = \|\widetilde{\boldsymbol{H}}_1\widetilde{\boldsymbol{Y}}_m\|_F^2 \\
\iff &(b-a)\|\widetilde{\boldsymbol{H}}_1\|_F^2 = Tr(\widetilde{\boldsymbol{H}}_1\widetilde{\boldsymbol{Y}}_m\widetilde{\boldsymbol{Y}}_m\widetilde{\boldsymbol{H}}_1^\top) \\
\iff &(b-a)\|\widetilde{\boldsymbol{H}}_1\|_F^2 = Tr((b-a)\widetilde{\boldsymbol{H}}_1\widetilde{\boldsymbol{H}}_1^\top) \\
\iff &(b-a)\|\widetilde{\boldsymbol{H}}_1\|_F^2 = (b-a)\|\widetilde{\boldsymbol{H}}_1\|_F^2
\end{aligned}
$$

Thus, we complete the proof. □

Remarks. We conjecture that the feature learned from data with all possible labels ($M = K$) will collapse to the origin which align with our tag-wise average property. Theoretically, $CE_{PAL}(\mathbf{z}^*, \mathbf{1})$ reaches its minimum as long as $\mathbf{z}^* = \zeta \cdot \mathbf{1}$ for arbitrary constant ζ. Since $\mathbf{z}^* = \mathbf{W}\mathbf{H}_K$ where $\mathbf{W}$ is ETF, it is easy to conclude that $\mathbf{H}_K$ has same index value on a given row. Due to regularization terms on $\mathbf{H}$, we conclude that $\mathbf{H}_K = \mathbf{0}$, i.e, the feature learned from data with all labels collapse to the origin. Extra experiment visualizing this phenomenon on the MSCOCO dataset could be found in Appendix A.

The following result is a M-lab generalization of Lemma B.5 from (Zhu et al., 2021):

Lemma C.8. *Let $S \subseteq \{1, \dots, K\}$ be a subset of size m where $1 \le m < K$. Then for all $\boldsymbol{z} = (z_1, \dots, z_K)^\top \in \mathbb{R}^K$ and all $c_{1,m} > 0$, there exists a constant $c_{2,m}$ such that*

$$
\mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}, \boldsymbol{y}_S) \ge \frac{1}{1+c_{1,m}}\frac{m}{K-m} \cdot \langle \mathbf{1} - \tfrac{K}{m}\mathbb{I}_S,\ \boldsymbol{z}\rangle + c_{2,m}. \tag{56}
$$

In fact, we have

$$
c_{2,m} := \frac{c_{1,m}m}{c_{1,m}+1}\log(m) + \frac{mc_{1,m}}{1+c_{1,m}}\log\left(\frac{c_{1,m}+1}{c_{1,m}}\right) + \frac{m}{c_{1,m}+1}\log\left((K-m)(c_{1,m}+1)\right).
$$

The Inequality (56) is tight, i.e., achieves equality, if and only if $\boldsymbol{z}$ satisfies all of the following:

1. *For all $i, j \in S$, we have $z_i = z_j$ (in-group equality). Let $z_{\text{in}} \in \mathbb{R}$ denote this constant.*
2. *For all for all $i, j \in S^c$, we have $z_i = z_j$ (out-group equality). Let $z_{\text{out}} \in \mathbb{R}$ denote this constant.*
3. $z_{\text{in}} - z_{\text{out}} = \log\left(\frac{(K-m)}{m}c_{1,m}\right) = \log\left(\gamma_{1,m}^{-1} - \frac{(K-m)}{m}\right)$.

Proof. Let $\boldsymbol{z}$ and $c_{1,m}$ be fixed. For convenience, let $\gamma_{1,m} := \frac{1}{1+c_{1,m}}\frac{m}{K-m}$. Below, let $z_{\text{in}}, z_{\text{out}} \in \mathbb{R}$ be arbitrary to be chosen later. Define $\boldsymbol{z}^* = (z_1^*, \dots, z_K^*) \in \mathbb{R}^K$ such that

$$
z_k^* = \begin{cases} z_{\text{in}} & : k \in S \\ z_{\text{out}} & : k \in S^c. \end{cases} \tag{57}
$$

For any $\boldsymbol{z} \in \mathbb{R}^K$, recall from the definition of pick-all-labels cross-entropy loss that

$$
\mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}, \boldsymbol{y}_S) = \sum_{k \in S}\mathcal{L}_{\text{CE}}(\boldsymbol{z}, \boldsymbol{y}_k)
$$

In particular, the function $\boldsymbol{z} \mapsto \mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}, \boldsymbol{y}_S)$ is a sum of strictly convex functions and is itself also strictly convex. Thus, the first order Taylor approximation of $\mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}, \boldsymbol{y}_S)$ around $\boldsymbol{z}^*$ yields the following lower bound:

$$
\begin{aligned}
\mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}, \boldsymbol{y}_S) &\ge \mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S) + \langle \nabla\mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S),\ \boldsymbol{z} - \boldsymbol{z}^*\rangle \\
&= \langle \nabla\mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S),\ \boldsymbol{z}\rangle + \mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S) - \langle \nabla\mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S),\ \boldsymbol{z}^*\rangle
\end{aligned} \tag{58}
$$

Next, we calculate $\nabla \mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S)$. First, we observe that

$$\nabla \mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S) = \sum_{k \in S} \nabla \mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}^*, \boldsymbol{y}_k).$$

Recall the well-known fact that the gradient of the cross-entropy is given by

$$\nabla \mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}^*, \boldsymbol{y}_k) = \mathrm{softmax}(\boldsymbol{z}^*) - \boldsymbol{y}_k. \tag{59}$$

Below, it is useful to define

$$\alpha := \frac{\exp(z^*_{\mathrm{in}})}{\sum_j \exp(z^*_j)} \quad \text{and} \quad \beta := \frac{\exp(z^*_{\mathrm{out}})}{\sum_j \exp(z^*_j)} \tag{60}$$

where $\sum_j \exp(z^*_j) = m\exp(z^*_{\mathrm{in}}) + (K-m)\exp(z^*_{\mathrm{out}})$. In view of this notation and the definition of $\boldsymbol{z}^*$ in Equation (57), we have

$$\mathrm{softmax}(\boldsymbol{z}^*) = \alpha \mathbb{I}_S + \beta \mathbb{I}_{S^c} \tag{61}$$

where we recall that $\mathbb{I}_S$ and $\mathbb{I}_{S^c} \in \mathbb{R}^K$ are the indicator vectors for the set S and S^c, respectively. Thus, combining Equation (59) and Equation (61), we get

$$\nabla \mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S) = \sum_{k \in S} \nabla \mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}^*, \boldsymbol{y}_k) = \sum_{k \in S} (\alpha \mathbb{I}_S + \beta \mathbb{I}_{S^c} - \boldsymbol{y}_k) = m(\alpha \mathbb{I}_S + \beta \mathbb{I}_{S^c}) - \mathbb{I}_S.$$

The above right-hand-side can be rewritten as

$$\begin{aligned} m(\alpha \mathbb{I}_S + \beta \mathbb{I}_{S^c}) - \mathbb{I}_S &= (m\alpha - 1) \cdot \mathbb{I}_S + m\beta \cdot \mathbb{I}_{S^c} \\ &= (m\alpha - 1 + m\beta - m\beta) \cdot \mathbb{I}_S + m\beta \cdot \mathbb{I}_{S^c} \\ &= m\beta \cdot \mathbf{1} - (m\beta + 1 - m\alpha) \cdot \mathbb{I}_S \\ &= m\beta \cdot \left(\mathbf{1} - \frac{m\beta + 1 - m\alpha}{m\beta} \cdot \mathbb{I}_S \right). \end{aligned}$$

Note that from Equation (61) we have $m\alpha + (K-m)\beta = 1$. Manipulating this expression algebraically, we have

$$\begin{aligned} & m\alpha + (K-m)\beta = 1 \\ \iff & k - m = \frac{1 - m\alpha}{\beta} \\ \iff & \frac{1}{\beta}\left(\frac{1}{m} - \alpha\right) = \frac{K}{m} - 1 \\ \iff & 1 + \frac{1}{m\beta} - \frac{\alpha}{\beta} = \frac{K}{m} \\ \iff & \frac{m\beta + 1 - m\alpha}{m\beta} = \frac{K}{m}. \end{aligned}$$

Putting it all together, we have

$$\nabla \mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S) = m\beta \cdot (\mathbf{1} - \tfrac{K}{m} \cdot \mathbb{I}_S).$$

Thus, combining Equation (58) with the above identity, we have

$$\mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}, \boldsymbol{y}_S) \geq m\beta \cdot \langle \mathbf{1} - \tfrac{K}{m} \cdot \mathbb{I}_S,\ \boldsymbol{z} \rangle + \mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S) - m\beta \cdot \langle \mathbf{1} - \tfrac{K}{m} \cdot \mathbb{I}_S, \boldsymbol{z}^* \rangle. \tag{62}$$

Let

$$c_{2,m} := \mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S) - m\beta \cdot \langle \mathbf{1} - \tfrac{K}{m} \cdot \mathbb{I}_S, \boldsymbol{z}^* \rangle \tag{63}$$

Note that this definition depends on β, which in terms depends in z^*_{in} and z^*_{out} which we have not yet defined. To define these quantities, note that in order to derive Equation (56) from Equation (62), a sufficient condition is to ensure that

$$\frac{1}{1 + c_{1,m}} \frac{m}{K - m} = m\beta = \frac{m \exp(z^*_{\mathrm{out}})}{\sum_j \exp(z^*_j)} = \frac{1}{\exp(z^*_{\mathrm{in}} - z^*_{\mathrm{out}}) + \frac{(K-m)}{m}} \tag{64}$$

Rearranging, the above can be rewritten as

$$(1 + c_{1,m}) \tfrac{K-m}{m} = \exp(z^*_{\mathrm{in}} - z^*_{\mathrm{out}}) + \tfrac{(K-m)}{m} \iff c_{1,m} = \tfrac{m}{K-m} \exp(z^*_{\mathrm{in}} - z^*_{\mathrm{out}})$$

or, equivalently, as

$$z^*_{\text{in}} - z^*_{\text{out}} = \log\left(\tfrac{(K-m)}{m} c_{1,m}\right). \tag{65}$$

Thus, if we choose $z^*_{\text{in}}, z^*_{\text{out}}$ such that the above holds, then Equation (56) holds.

Finally, we compute the closed-form expression for $c_{2,m}$ defined in Equation (63), which we restate below for convenience:

$$c_{2,m} := \mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S) - m\beta \cdot \langle \mathbf{1} - \tfrac{K}{m} \cdot \mathbb{I}_S, \boldsymbol{z}^* \rangle$$

The expression for $m\beta$ is given at Equation (64). Moreover, we have

$$\langle \mathbf{1} - \tfrac{K}{m} \cdot \mathbb{I}_S, \boldsymbol{z}^* \rangle = m z_{\text{in}} + (K-m) z_{\text{out}} - \tfrac{K}{m} m z_{\text{in}} = -(K-m)(z_{\text{in}} - z_{\text{out}}).$$

Thus, we have

$$\begin{aligned}
-m\beta \cdot \langle \mathbf{1} - \tfrac{K}{m} \cdot \mathbb{I}_S, \boldsymbol{z}^* \rangle &= \frac{(K-m)(z_{\text{in}} - z_{\text{out}})}{\exp(z^*_{\text{in}} - z^*_{\text{out}}) + \frac{(K-m)}{m}} \\
&= \frac{(K-m)\log\left(\frac{(K-m)}{m} c_{1,m}\right)}{\frac{(K-m)}{m} c_{1,m} + \frac{(K-m)}{m}} \\
&= \frac{m}{c_{1,m}+1} \log\left(\tfrac{(K-m)}{m} c_{1,m}\right).
\end{aligned}$$

On the other hand,

$$\mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S) = \sum_{k \in S} \mathcal{L}_{\text{CE}}(\boldsymbol{z}^*, \boldsymbol{y}_k)$$

Now,

$$\begin{aligned}
\mathcal{L}_{\text{CE}}(\boldsymbol{z}^*, \boldsymbol{y}_k) &= -\log([\text{softmax}(\boldsymbol{z}^*)]_k) \\
&= -\log(\exp(z^*_{\text{in}})/(m\exp(z^*_{\text{in}}) + (K-m)\exp(z^*_{\text{out}}))) \\
&= \log(m + (K-m)\exp(z^*_{\text{out}} - z^*_{\text{in}})) \\
&= \log\left(m + (K-m)(1/\exp(z^*_{\text{in}} - z^*_{\text{out}}))\right) \\
&= \log\left(m + (K-m)\frac{1}{\frac{(K-m)}{m} c_{1,m}}\right) && \text{by Equation (65)} \\
&= \log\left(m + m\frac{1}{c_{1,m}}\right) \\
&= \log\left(m\left(\frac{c_{1,m}+1}{c_{1,m}}\right)\right)
\end{aligned}$$

Thus

$$\mathcal{L}_{\texttt{PAL}}(\boldsymbol{z}^*, \boldsymbol{y}_S) = m\log\left(m\left(\frac{c_{1,m}+1}{c_{1,m}}\right)\right).$$

Putting it all together, we have

$$c_{2,m} = m\log\left(m\left(\frac{c_{1,m}+1}{c_{1,m}}\right)\right) + \frac{m}{c_{1,m}+1}\log\left(\tfrac{(K-m)}{m} c_{1,m}\right).$$

$$m\log\left(m\left(\frac{c_{1,m}+1}{c_{1,m}}\right)\right) = m\log(m) + m\log\left(\frac{c_{1,m}+1}{c_{1,m}}\right)$$

$$\frac{m}{c_{1,m}+1}\log\left(\tfrac{(K-m)}{m} c_{1,m}\right) = \frac{m}{c_{1,m}+1}\log\left((K-m)c_{1,m}\right) - \frac{m}{c_{1,m}+1}\log(m)$$

Putting it all together, we have

$$c_{2,m} = m\log\left(m\left(\frac{c_{1,m}+1}{c_{1,m}}\right)\right) + \frac{m}{c_{1,m}+1}\log\left(\tfrac{(K-m)}{m}c_{1,m}\right) \tag{66}$$

$$= m\log(m) + m\log\left(\frac{c_{1,m}+1}{c_{1,m}}\right) + \frac{m}{c_{1,m}+1}\log\left((K-m)c_{1,m}\right) - \frac{m}{c_{1,m}+1}\log(m) \tag{67}$$

$$= \frac{c_{1,m}m}{c_{1,m}+1}\log(m) + m\log\left(\frac{c_{1,m}+1}{c_{1,m}}\right) + \frac{m}{c_{1,m}+1}\log\left((K-m)c_{1,m}\right) \tag{68}$$

Next, for simplicity, let us drop the subscript and simply write $c := c_{1,m}$. Then

$$\begin{aligned}
&m\log\left(\frac{c+1}{c}\right) + \frac{m}{c+1}\log\left((K-m)c\right) \\
&= \frac{m}{1+c}\log\left(\frac{c+1}{c}\right) + \frac{mc}{1+c}\log\left(\frac{c+1}{c}\right) + \frac{m}{c+1}\log\left((K-m)c\right) \qquad \because \tfrac{1}{1+c} + \tfrac{c}{1+c} = 1 \\
&= \frac{m}{1+c}\log\left(\frac{c+1}{c}\right) + \frac{mc}{1+c}\log\left(\frac{c+1}{c}\right) + \frac{m}{c+1}\log\left((K-m)c\right) \\
&\quad + \frac{m}{c+1}\log\left((K-m)(c+1)\right) - \frac{m}{c+1}\log\left((K-m)(c+1)\right) \qquad \because \text{add a “zero”} \\
&= \frac{m}{1+c}\log\left(\frac{c+1}{c}\right) + \frac{mc}{1+c}\log\left(\frac{c+1}{c}\right) + \frac{m}{c+1}\log\left(\frac{c}{c+1}\right) \qquad \because \text{property of } \log \\
&\quad + \frac{m}{c+1}\log\left((K-m)(c+1)\right) \\
&= \frac{mc}{1+c}\log\left(\frac{c+1}{c}\right) + \frac{m}{c+1}\log\left((K-m)(c+1)\right) \qquad \because \log(\tfrac{c+1}{c}) = -\log(\tfrac{c}{c+1})
\end{aligned}$$

To conclude, we have

$$c_{2,m} = \frac{c_{1,m}m}{c_{1,m}+1}\log(m) + \frac{mc_{1,m}}{1+c_{1,m}}\log\left(\frac{c_{1,m}+1}{c_{1,m}}\right) + \frac{m}{c_{1,m}+1}\log\left((K-m)(c_{1,m}+1)\right)$$

as desired. □

D. Nonconvex Landscape Analysis

Due to the nonconvex nature of Problem (3), the characterization of global optimality alone in Theorem 3.1 is not sufficient for guaranteeing efficient optimization to those desired global solutions. Thus, we further study the global landscape of Problem (3) by characterizing all of its critical points, we show the following result.

Theorem D.1 (**Benign Optimization Landscape**). *(Generalization of (Zhu et al., 2021) Theorem 3.2) Suppose the same setting of Theorem 3.1, and assume the feature dimension is larger than the number of classes, i.e., $d > K$, and the number of training samples for each class are balanced within each multiplicity. Then the function $f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ in Problem (4) is a strict saddle function with no spurious local minimum in the sense that:*

- *Any local minimizer of f is a global solution of the form described in Theorem 3.1.*
- *Any critical point $(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})$ of f that is not a global minimizer is a strict saddle point with negative curvatures, in the sense that there exists some direction $(\boldsymbol{\Delta_W}, \boldsymbol{\Delta_H}, \boldsymbol{\delta_b})$ such that the directional Hessian $\nabla^2 f(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{b})[\boldsymbol{\Delta_W}, \boldsymbol{\Delta_H}, \boldsymbol{\delta_b}] < 0$.*

The original proof in (Zhu et al., 2021) connects the nonconvex optimization problem to a convex low-rank realization and then characterizes the global optimality conditions based on the convex problem. The proof concludes by analyzing all critical points, guided by the identified optimality conditions. Because the PAL loss for M-lab is reduced from the CE loss in M-clf, the above result can be generalized from the result in (Zhu et al., 2021). Unlike Theorem 3.1, the result of benign landscape in Theorem D.1 does not hold for $d = K - 1$. The reason is that we need to construct a negative curvature direction in the null space of $\boldsymbol{W}$ for showing strict saddle points. Similar to (Zhu et al., 2021), we conjecture the M-lab NC results also hold for $d = K$ and leave it for future work.

In our paper, we establish the theoretical properties of all critical points, demonstrating that the function is a strict saddle

function (Ge et al., 2015) in the context of multi-label learning with respect to $(\boldsymbol{W}, \boldsymbol{H})$. It's worth noting that for strict saddle functions like PAL-CE, there exists a substantial body of prior research in the literature that provides rigorous algorithmic convergence to global minimizers. In our case, this equates to achieving a global multi-label neural collapse solution. These established methods include both first-order gradient descent techniques (Ge et al., 2015; Jin et al., 2017; Lee et al., 2019) and second-order trust-region methods (Sun et al., 2016), all of which ensure efficient algorithmic convergence.

Proof of Theorem D.1. We note that the proof for Theorem 3.2 in (Zhu et al., 2021) could be directly extended in our analysis. More specifically, the proof in (Zhu et al., 2021) relies on a connection for the original loss function to its convex counterpart, in particular, letting $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{H} \in \mathbb{R}^{K\times N}$ with $N = \sum_m n_m$ and $\alpha = \frac{\lambda_{\boldsymbol{H}}}{\lambda_{\boldsymbol{W}}}$, the original proof first shows the following fact:

$$\begin{aligned}\min_{\boldsymbol{H}\boldsymbol{W}=\boldsymbol{Z}} \lambda_{\boldsymbol{W}}||\boldsymbol{W}||_F^2 + \lambda_{\boldsymbol{H}}||\boldsymbol{H}||_F^2 &= \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}} \min_{\boldsymbol{H}\boldsymbol{W}=\boldsymbol{Z}} \frac{1}{\sqrt{\alpha}}(||\boldsymbol{W}||_F^2 + \alpha||\boldsymbol{H}||_F^2) \\ &= \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}||\boldsymbol{Z}||_*.\end{aligned}$$

With the above result, the original proof relates the original loss function

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}} f(\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}) \;:=\; g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top) \;+\; \lambda_{\boldsymbol{W}}||\boldsymbol{W}||_F^2 + \lambda_{\boldsymbol{H}}||\boldsymbol{H}||_F^2 + \lambda_{\boldsymbol{b}}||\boldsymbol{b}||_2^2$$

with

$$g(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top) := \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathcal{L}_{\mathrm{CE}}(\boldsymbol{W}\boldsymbol{h}_{k,i} + \boldsymbol{b}, \boldsymbol{y}_k),$$

to a convex problem:

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{K\times N},\, \boldsymbol{b}\in\mathbb{R}^K} \widetilde{f}(\boldsymbol{Z},\boldsymbol{b}) \;:=\; g(\boldsymbol{Z} + \boldsymbol{b}\mathbf{1}^\top) + \sqrt{\lambda_{\boldsymbol{W}}\lambda_{\boldsymbol{H}}}||\boldsymbol{Z}||_* + \lambda_{\boldsymbol{b}}||\boldsymbol{b}||_2^2.$$

In our analysis, by letting $\widetilde{g}(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}\mathbf{1}^\top) := \frac{1}{Nm}\sum_{m=1}^{m}\sum_{i=1}^{n_m}\sum_{k=1}^{\binom{K}{m}}\mathcal{L}_{\texttt{PAL}}(\boldsymbol{W}\boldsymbol{h}_{m,k,i} + \boldsymbol{b}, \boldsymbol{y}_{S_{m,k}})$, we can directly apply the original proof for our problem. For more details, we refer readers to the proof of Theorem 3.2 in (Zhu et al., 2021).

□