Neural Collapse for Cross-entropy Class-Imbalanced Learning with Unconstrained ReLU Features Model

Hien Dang [1 2] **Tho Tran** [2] **Tan Nguyen** [* 3] **Nhat Ho** [* 1]

Abstract

The current paradigm of training deep neural networks for classification tasks includes minimizing the empirical risk, pushing the training loss value towards zero even after the training classification error has vanished. In this terminal phase of training, it has been observed that the last-layer features collapse to their class-means and these class-means converge to the vertices of a simplex Equiangular Tight Frame (ETF). This phenomenon is termed as Neural Collapse ($\mathcal{NC}$). However, this characterization only holds in class-balanced datasets where every class has the same number of training samples. When the training dataset is class-imbalanced, some $\mathcal{NC}$ properties will no longer hold true, for example, the geometry of class-means will skew away from the simplex ETF. In this paper, we generalize $\mathcal{NC}$ to imbalanced regime for cross-entropy loss under the unconstrained ReLU features model. We demonstrate that while the within-class features collapse property still holds in this setting, the class-means will converge to a structure consisting of orthogonal vectors with lengths dependent on the number of training samples. Furthermore, we find that the classifier weights (i.e., the last-layer linear classifier) are aligned to the scaled and centered class-means, with scaling factors dependent on the number of training samples of each class. This generalizes $\mathcal{NC}$ in the class-balanced setting. We empirically validate our results through experiments on practical architectures and dataset.

[*]Equal contribution [1]Department of Statistics and Data Sciences, University of Texas at Austin, USA [2]FPT Software AI Center, Vietnam [3]Department of Mathematics, National University of Singapore, Singapore. Correspondence to: Hien Dang <danghoanghien1123@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Cross-entropy (CE) is undoubtedly one of the most popular loss functions used for training neural networks in the current deep learning paradigm. However, some crucial aspects of training networks using this loss function have not been fully explored yet, for example: i) Is there any unique pattern that the models learn when training deep neural networks until convergence, i.e., to reach zero loss?, ii) How do the learned network parameters vary across data distribution, training instances, and model architecture?, iii) What are the geometries of the representations and the classifier obtained from minimizing CE loss?. Understanding these questions is crucial for studying the training and generalization properties of deep neural networks. For instance, it has been a long-standing problem that training networks using CE loss under a long-tailed distribution dataset causes a significant drop in accuracy, especially for classes with a scarce amount of training samples. An important observation for this phenomenon is that the classifier weight vector of a more frequent class tends to have a larger norm, thus biasing the decision boundary toward the less frequent class. As a consequence, a smaller volume of the feature space is allocated for the minority classes, which leads to a drop in performance (Kim & Kim, 2020; Kang et al., 2019; Cao et al., 2019; Ye et al., 2020; Liu et al., 2023; Kang et al., 2020).

A noticeable progress in answering these questions is the discovery of *Neural Collapse* phenomenon (Papyan et al., 2020). Neural Collapse ($\mathcal{NC}$) reveals a common pattern of the learned last-layer features and the classifier weight of deep neural networks across canonical datasets and architectures. Specifically, $\mathcal{NC}$ consists of four properties emerging in the terminal phase of training of training deep neural networks for balanced datasets (i.e., every class has the same number of training instances):

- ($\mathcal{NC}1$) **Variability collapse:** features of the samples within the same class converge to a unique vector (i.e., the *class-mean*), as training progresses.
- ($\mathcal{NC}2$) **Convergence to simplex ETF:** the optimal class-means have the same length and are equally and maximally pairwise separated, i.e., they form a simplex

Equiangular Tight Frame (ETF).

- ($\mathcal{NC}3$) **Convergence to self-duality:** up to rescaling, the class-means and classifiers converge on each other.
- ($\mathcal{NC}4$) **Simplification to nearest class-center:** given a feature, the classifier converges to choosing whichever class has the nearest class-mean to it.

The intriguing empirical observation of Neural Collapse has attracted many theoretical investigations, mostly under a simplified *unconstrained features model (UFM)* (see Section 3 for more details) and for *class-balanced dataset*, demonstrating that $\mathcal{NC}$ properties occur at any global solution of the loss function. However, under *class-imbalanced dataset*, it has been observed that deep neural networks exhibit different geometric structures and some $\mathcal{NC}$ properties are not satisfied anymore (Dang et al., 2023; Thrampoulidis et al., 2022; Hong & Ling, 2023). The last-layer features of samples within the same class still converge to their class-means ($\mathcal{NC}1$), but the class-means and the classifier weights will no longer form a simplex ETF ($\mathcal{NC}2$) (Fang et al., 2021). In a more extreme case where the imbalance level exceeds a certain threshold, the learned classifiers of the minority classes collapse onto each other, becoming indistinguishable from those of other classes (Fang et al., 2021). This phenomenon, known as *Minority Collapse*, explains why the accuracy for these minority classes drops significantly compared to the class-balanced setting.

While the Neural Collapse emergence at the optimal solution in deep neural networks training using CE loss for *balanced dataset* has been extensively studied (Lu & Steinerberger, 2020; Zhu et al., 2021), the corresponding characterization for this loss function in *imbalanced scenario* has remained limited. Under imbalanced regime, several theoretical works have characterized Neural Collapse phenomenon for other loss functions. In particular, (Dang et al., 2023) has demonstrated the convergence geometry of the learned features and learned classifier for the mean squared error (MSE) loss. (Thrampoulidis et al., 2022) studies the support vector machine (SVM) problem, whose global minima follows a different geometry known as Simplex-Encoded-Labels Interpolation (SELI) and later (Behnia et al., 2023) extends it to some other SVM parameterizations.

Comparison to concurrent work (Hong & Ling, 2023): For CE loss, we acknowledge that the concurrent work (Hong & Ling, 2023) is closely related to our work. They investigate Neural Collapse for CE loss with UFM under imbalanced setting. They prove the within-class features collapse property ($\mathcal{NC}1$) and demonstrate the *network output prediction vectors converge to a block structure* where each block corresponds to classes that have the same amount of training samples. However, their analysis does not cover the magnitude of prediction vectors and the magnitude of each block within the structure. Consequently, it is not yet possible to describe the geometry explicitly and quantify how the structure changes under different imbalance levels. Additionally, the geometry of the learned last-layer feature, the classifier weight ($\mathcal{NC}2$) and the relationship between them ($\mathcal{NC}3$) have not been examined in (Hong & Ling, 2023). As a result, the corresponding ($\mathcal{NC}2$) and ($\mathcal{NC}3$) properties has not been characterized for this setting. Moreover, other considerations such as the norm, the norm ratio, and angle between the classifier weights and features are also not addressed.

On the other hand, in this paper, we study CE loss training problem using UFM, but with a different setting, in which the features have to be *element-wise non-negative*. This setting is motivated by the current paradigm that features are typically the outcome of some non-negative activation function, like ReLU or sigmoid. In this setting, we study the global solutions of CE training problem under UFM and present a thorough analysis of the convergence geometry of the last-layer features and classifier. We summarize our contributions as follows:

- We explicitly characterize Neural Collapse for the last-layer features and classifier weights in CE loss training with non-negative features in class-imbalanced settings. We prove that at optimality, $\mathcal{NC}1$ still occurs, and the optimal class-means form an orthogonal structure. We derive the closed-form lengths of the features, in terms of the number of training samples and other hyperparameters.
- We find that the classifier weight is aligned to a scaled and centered version of the class-means, which generalizes the properties $\mathcal{NC}2$ and $\mathcal{NC}3$ from the original definition of Neural Collapse. Additionally, we derive the norms, norm ratios and angles between these classifier weights explicitly.
- We derive the exact threshold of the amount of training samples for a class to collapse and become indistinguishable from other classes. Hence, the threshold for Minority Collapse is also obtained in our analysis.

Notation: For a weight matrix $\mathbf{W}$, we use $\mathbf{w}_j$ to denote its j-th row vector. $\|.\|_F$ denotes the Frobenius norm of a matrix and $\|.\|_2$ denotes L_2-norm of a vector. $\otimes$ denotes the Kronecker product. The symbol "$\propto$" denotes proportional, i.e, equal up to a positive scalar. We also use some common matrix notations: $\mathbf{1}_n$ is the all-ones vector, $\text{diag}\{a_1, \ldots, a_K\}$ is a square diagonal matrix size $K \times K$ with diagonal entries $a_1, \ldots, a_K$. We use $[K]$ to denote the index set $\{1, 2, \ldots, K\}$.

2. Related Works

Neural Collapse on balanced dataset: A surge of theoretical results for $\mathcal{NC}$ under balanced scenario has emerged after the discovery of this phenomenon. Due to the highly non-convexity of the problem of training deep networks, theoretical works have proven the occurrence of $\mathcal{NC}$ for different loss functions and architectures with a simplified unconstrained features model (see Section 3 for more details) (Lu & Steinerberger, 2020; Zhu et al., 2021; Graf et al., 2023; Zhou et al., 2022a;b; Tirer & Bruna, 2022; Dang et al., 2023; Thrampoulidis et al., 2022; Behnia et al., 2023; Kini et al., 2023). In particular, $\mathcal{NC}$ properties are proven to occur at the optimal last-layer features and classifier across different loss functions: cross-entropy (Lu & Steinerberger, 2020; Zhu et al., 2021), mean squared error (Zhou et al., 2022a; Tirer & Bruna, 2022; Dang et al., 2023), supervised contrastive loss (Graf et al., 2023) and also for focal loss and label smoothing (Zhou et al., 2022b). Recent works have spent efforts to extend the UFM to deeper architectures to study the behavior of more layers after the "unconstrained features". Specifically, (Tirer & Bruna, 2022) extends UFM to account for one additional layer, from one-layer linear classifier to two-layer linear classifier after the "unconstrained" features for MSE loss, and later the work (Dang et al., 2023) extends the setting to a general deep linear network for both MSE and CE losses. (Tirer & Bruna, 2022) also extends UFM for MSE loss to a two-layer case with ReLU activation. This setting is later extended by (Súkeník et al., 2023) to the general deep UFM with ReLU activation for the binary classification problem. For multiclass classification problem with MSE loss, recent extensions to account for additional layers in the analysis with non-linearity are studied in (Tirer & Bruna, 2022; Rangamani & Banburski-Fahey, 2022), or with batch normalization (Ergen & Pilanci, 2020). However, these works require strong assumptions on the global optimal solution or the network architecture and capability for their theoretical results to be hold. There are also efforts to mitigate the restriction of UFM, such as (Tirer et al., 2023) analyzes UFM with an additional regularization term to force the features to stay in the vicinity of a predefined feature matrix (e.g., intermediate features). Additionally, (Zhu et al., 2021; Zhou et al., 2022a;b) prove the benign optimization landscape for several loss functions under UFM, demonstrating that critical points can only be global minima or strict saddle points.

Neural Collapse on imbalanced dataset: The work (Fang et al., 2021) is likely the first to observe that for imbalanced setting, the collapse of features within the same class $\mathcal{NC}1$ is preserved, but the geometry skews away from the ETF. They also present the "Minority Collapse" phenomenon, in which the minority classifiers collapse to the same vector if the imbalance level is greater than some threshold. For MSE loss, (Dang et al., 2023) has explicitly characterized the geometry of the learned features and classifiers for imbalanced setting. (Dang et al., 2023) showed that the $\mathcal{NC}1$ still holds and the class-means converge to a General Orthonormal Frame (GOF), which consists of orthonormal vectors but with different lengths. By applying non-negative constraints for the normalized features to incorporate the effect of ReLU activation, (Kini et al., 2023) finds the global minimizers of supervised contrastive loss and proves that the optimal features form an Orthogonal Frame (OF) with equal length and orthogonal vectors, regardless of the imbalance level. (Thrampoulidis et al., 2022) theoretically studies the UFM-SVM problem, whose global minima follow a more general geometry than the ETF, called "SELI". However, this work also makes clear that the unregularized version of CE loss only converges to KKT points of the SVM problem, which are not necessarily global minima. The result for UFM-SVM is later extended by the work (Behnia et al., 2023) to consider several cross-entropy parameterizations.

Regarding CE loss, (Yang et al., 2022) studies the imbalanced setting but with fixed, unlearnable last-layer linear classifiers as a simplex ETF. They prove that no matter whether the data distribution is balanced or not among classes, the features will converge to a simplex ETF in the same direction as the fixed classifier. As mentioned in Section 1, the work (Hong & Ling, 2023) is closely related to our work and they study CE loss with UFM and the features can have negative entries. They prove that at optimality, the within-class features collapse ($\mathcal{NC}1$) and the *network output prediction vectors converge to a block structure*. However, their analysis does not cover the magnitude of prediction vectors and the ratio between each block within the structure are not yet covered. Thus, it is not possible to describe the geometry explicitly and quantify how the structure changes under different imbalance levels. Additionally, the structure of the learned last-layer feature, the classifier weight ($\mathcal{NC}2$) and the relation between them ($\mathcal{NC}3$) have not been derived in (Hong & Ling, 2023). As a result, the corresponding ($\mathcal{NC}2$) and ($\mathcal{NC}3$) properties has not been characterized for this setting.

3. Problem Setup

Training Neural Network with Cross-Entropy Loss: In this work, we focus on neural network trained using the cross-entropy (CE) loss function on an imbalanced dataset. We consider the classification task with K classes. Let n_k denote the number of training samples in class $k \in [K]$, and $N := \sum_{k=1}^{K} n_k$, the total number of training samples. A typical deep neural network classifier consists of a feature mapping function $\mathbf{h}(\boldsymbol{x})$ and a linear classifier parameterized as $\mathbf{W}$. Specifically, a typical L-layer deep neural network

can be expressed as follows:

$$\psi_{\mathbf{W},\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbf{W}_L \underbrace{\sigma(\mathbf{W}_{L-1}\ldots\sigma(\mathbf{W}_1\boldsymbol{x}+\mathbf{b}_1)+\mathbf{b}_{L-1})}_{\text{Feature } \mathbf{h}=\mathbf{h}(x)},$$

where each layer composes of an affine transformation parameterized by a weight matrix $\mathbf{W}_l$ and bias $\mathbf{b}_l$, followed by a non-linear activation σ (e.g., $\text{ReLU}(x) = \max(x, 0)$). Here, $\boldsymbol{\theta} := \{\mathbf{W}_l, \mathbf{b}_l\}_{l=1}^{L-1}$ is the set of all learnable parameters in the feature mapping. We denote the last layer linear classifier as $\mathbf{W} := \mathbf{W}_L$ for convenience. Current paradigm trains the network by minimizing the empirical risk over all training samples $\{(\boldsymbol{x}_{k,i}, \boldsymbol{y}_k)\}_{k,i}$ where $\boldsymbol{x}_{k,i}$ denoted the i-th sample of class k and $\boldsymbol{y}_k$ is the one-hot label vector for class k:

$$\min_{\mathbf{W},\boldsymbol{\theta}} \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathcal{L}(\psi_{\mathbf{W},\theta}(\boldsymbol{x}_{k,i}), \boldsymbol{y}_k) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_\theta}{2}\|\boldsymbol{\theta}\|^2,$$

where $\lambda_W, \lambda_\theta > 0$ are the weight decay parameters and $\mathcal{L}(\psi(\mathbf{x}_{k,i}), \mathbf{y}_k)$ is the loss function that measures the difference between the output $\psi(\mathbf{x}_{k,i})$ and the target $\mathbf{y}_k$. For a vector $\mathbf{z} = [z_1, z_2, \ldots, z_K] \in \mathbb{R}^K$ and a target one-hot vector $\boldsymbol{y}_k$, CE loss is defined as:

$$\mathcal{L}_{CE}(\mathbf{z}, \boldsymbol{y}_k) = -\log\left(\frac{\exp(z_k)}{\sum_{m=1}^{K}\exp(z_j)}\right) \tag{1}$$

Unconstrained Features Model (UFM) with non-negative features: Due to the significant challenges of analyzing the highly non-convex neural network training problem, recent theoretical works study $\mathcal{NC}$ phenomenon using a simplified model called unconstrained features model (UFM), or, layer-peeled model (Fang et al., 2021). In particular, UFM peels down the last-layer of the network and treats the last-layer features $\mathbf{h}_{k,i} = \mathbf{h}(\mathbf{x}_{k,i}) \in \mathbb{R}^d$ as free optimization variables in order to capture the main characteristics of the last layers related to $\mathcal{NC}$ during training. This relaxation can be justified by the well-known result that an overparameterized deep neural network can approximate any continuous function (Hornik et al., 1989; Hornik, 1991; Zhou, 2018; Yarotsky, 2018).

In this work, we consider a slight variant of UFM, in which the *features are constrained to be non-negative*, motivated by the fact that features are usually the output of ReLU activations in many common architectures. Formally, we consider the following modified version of UFM trained with CE loss with non-negative features:

$$\min_{\mathbf{W},\mathbf{H}} \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathcal{L}_{CE}(\mathbf{W}\mathbf{h}_{k,i}, \mathbf{y}_k) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 \tag{2}$$
$$+ \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2, \quad \text{s.t.} \quad \mathbf{H} \geq 0, \lambda_W > 0, \lambda_H > 0,$$

where $\mathbf{H} := [\mathbf{h}_{1,1}, \ldots, \mathbf{h}_{1,n_1}, \mathbf{h}_{2,1}, \ldots, \mathbf{h}_{K,n_K}] \in \mathbb{R}^{d\times N}$ and $\mathbf{H} \geq 0$ denotes entry-wise non-negativity. We note that similar settings with ReLU features were previously considered in (Nguyen et al., 2022), where $\mathcal{NC}$ configuration was derived for the label smoothing loss under balanced setting, and in (Kini et al., 2023), which studied the convergence geometry for supervised contrastive loss under imbalanced setting. We denote this setting as UFM_+, as (Kini et al., 2023), to differentiate it from the original UFM.

By denoting $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K]^\top \in \mathbb{R}^{K\times d}$ be the last-layer weight matrix, with $\mathbf{w}_k \in \mathbb{R}^d$ is the k-th row of $\mathbf{W}$, the CE loss can be written as:

$$\mathcal{L}_{CE}(\mathbf{W}\mathbf{h}_{k,i}, \mathbf{y}_k) = -\log\left(\frac{\exp(\mathbf{w}_k^\top\mathbf{h}_{k,i})}{\sum_{m=1}^{K}\exp(\mathbf{w}_m^\top\mathbf{h}_{k,i})}\right).$$

We also denote the *class-mean* of a class $k \in [K]$ as $\mathbf{h}_k := n_k^{-1}\sum_{i=1}^{n_k}\mathbf{h}_{k,i}$ and the *global-mean* $\mathbf{h}_G := N^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathbf{h}_{k,i}$. The *class-mean matrix* is denoted as $\overline{\mathbf{H}} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K] \in \mathbb{R}^{d\times K}$.

Neural Collapse for balanced dataset: With the notations defined above, we recall the $\mathcal{NC}$ properties in the balanced setting as follows:

- ($\mathcal{NC}1$) **Variability collapse**:

$$\mathbf{h}_{k,i} = \mathbf{h}_k, \quad \forall k \in [K], i \in [n_k].$$

- ($\mathcal{NC}2$) **Convergence to simplex ETF**:

$$(\overline{\mathbf{H}} - \mathbf{h}_G\mathbf{1}_K^\top)^\top(\overline{\mathbf{H}} - \mathbf{h}_G\mathbf{1}_K^\top) \propto \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top.$$

- ($\mathcal{NC}3$) **Convergence to self-duality**:

$$\mathbf{W} \propto (\overline{\mathbf{H}} - \mathbf{h}_G\mathbf{1}_K^\top)^\top.$$

Orthogonal Frame and General Orthogonal Frame: Some previous results, such as (Tirer & Bruna, 2022; Nguyen et al., 2022), derive that under the balanced setting, the optimal class-means $\{\mathbf{h}_k\}$ form an *orthogonal frame (OF)*, i.e., $\overline{\mathbf{H}}^\top\overline{\mathbf{H}} \propto \mathbf{I}_K$. By centering the OF structure with its mean vector, we will receive a simplex ETF. Thus, this structure still follows ($\mathcal{NC}2$) property. For MSE loss under class-imbalanced scenario, (Dang et al., 2023) proves that the class-means $\{\mathbf{h}_k\}$ form an orthogonal structure consisting of pairwise orthogonal vectors but having different lengths. They termed this structure as *general orthogonal frame* (GOF). We will use this notation for our results in Section 4.

SELI geometry: Simplex-Encoded-Labels Interpolation (SELI) is the geometric structure of the optimal classifier,

feature and the prediction matrix of the imbalanced SVM training problem under UFM (Thrampoulidis et al., 2022). In particular, the prediction matrix $\mathbf{Z} = \mathbf{W}\mathbf{H} \in \mathbb{R}^{K\times N}$ is proven to have its i-th column to be $\boldsymbol{y}_i - \frac{1}{K}\mathbf{1}_K$ for all $i \in [N]$. Then, if we denote the SVD of the matrix $\mathbf{Z} = \mathbf{V}\Lambda\mathbf{U}^\top$, the classifier and feature matrices satisfy that $\mathbf{W}\mathbf{W}^\top = \mathbf{V}\Lambda\mathbf{V}^\top$ and $\mathbf{H}^\top\mathbf{H} = \mathbf{U}\Lambda\mathbf{U}^\top$. For UFM-CE training problem but without regularization ($\lambda = 0$), (Ji et al., 2021) showed that its gradient flow converges in direction to a Karush-Kuhn-Tucker (KKT) point of the UFM-SVM problem. Thus, SELI geometry is not necessarily the global minima of the unregularized UFM-CE problem.

4. Main Result: Global Structure of UFM$_+$ Cross-Entropy Imbalanced

In this section, we characterize the global solution $(\mathbf{W}, \mathbf{H})$ of the non-convex problem (2) and analyze its geometries. We prove that irrespective of the label distribution, the optimal features form an orthogonal structure in the non-negative orthant while the classifiers align with the scaled-and-centered features and spread across the entire feature space with $\sum_{k=1}^K \mathbf{w}_k = \mathbf{0}$. For convenience, we define the following constants for every class $k \in [K]$:

$$\overline{M}_k := \log\left((K-1)\left(\frac{\sqrt{n_k}}{N\sqrt{\frac{K-1}{K}\lambda_W\lambda_H}} - 1\right)\right),$$
$$M_k := \begin{cases} \overline{M}_k & \text{if } \overline{M}_k > 0 \\ 0 & \text{if } \overline{M}_k \le 0 \text{ or } \overline{M}_k \text{ is undefined} \end{cases}. \quad (3)$$

Note that the inequality $M_k = \overline{M}_k > 0$ is equivalent to $\frac{N}{\sqrt{n_k}}\sqrt{\lambda_W\lambda_H} < \sqrt{\frac{K-1}{K}}$ and $M_k = 0$ when and only when $\frac{N}{\sqrt{n_k}}\sqrt{\lambda_W\lambda_H} \ge \sqrt{\frac{K-1}{K}}$. We state the our main result in the following theorem.

Theorem 4.1 (Geometry of UFM$_+$ Cross-Entropy Imbalanced minimizers)**.** *Suppose $d \ge K$ and $\frac{N}{\sqrt{n_k}}\sqrt{\lambda_W\lambda_H} < \sqrt{\frac{K-1}{K}}\ \forall\, k \in [K]$, then any global minimizer $(\mathbf{W}, \mathbf{H})$ of the problem (2) obeys*

(a) Within-class feature collapse:

$$\forall\, k \in [K],\ \mathbf{h}_{k,i} = \mathbf{h}_{k,j}, \quad \forall\, i \ne j. \quad (4)$$

(b) Class-mean orthogonality:

$$\mathbf{h}_k^\top \mathbf{h}_l = 0, \quad \forall\, k \ne l. \quad (5)$$

(c) Class-mean norm:

$$\|\mathbf{h}_k\|^2 = \sqrt{\frac{K-1}{K}\frac{\lambda_W}{\lambda_H}\frac{1}{n_k}}M_k. \quad (6)$$

(d) Relation between the classifier and class-means:

$$\mathbf{w}_k = \sqrt{\frac{\lambda_H}{\lambda_W K(K-1)}}\left(K\sqrt{n_k}\mathbf{h}_k - \sum_{m=1}^K \sqrt{n_m}\mathbf{h}_m\right) \quad (7)$$
$$\text{and } \sum_{k=1}^K \mathbf{w}_k = \mathbf{0}.$$

(e) Prediction vector of class k-th sample:

$$\mathbf{z}_k^{(k)} = (\mathbf{W}\mathbf{h}_k)^{(k)} = \frac{K-1}{K}M_k, \quad (8)$$
$$\mathbf{z}_k^{(m)} = (\mathbf{W}\mathbf{h}_k)^{(m)} = -\frac{1}{K}M_k, \quad \forall m \ne k, \quad (9)$$
$$\text{and } \sum_{m=1}^K \mathbf{z}_k^{(m)} = 0.$$

If there is any $k \in [K]$ such that $\frac{N}{\sqrt{n_k}}\sqrt{\lambda_W\lambda_H} \ge \sqrt{\frac{K-1}{K}}$, then the k-th class-mean $\mathbf{h}_k = \mathbf{0}$ and all properties above still hold.

We postpone the detailed proof until Section B in the Appendix. At a high level, our proof finds the lower bound of the loss function and studies the conditions to achieve the bound. We start by bounding the cross-entropy term to move the logit $\mathbf{z}_k = \mathbf{W}\mathbf{h}_k$ out of the logarithm and exponent, using arguments based on Cauchy-Schwartz and Jensen inequalities. Next, the technical challenging part is the realization of the alignment between $\mathbf{w}_k$ and $K\sqrt{n_k}\mathbf{h}_k - \sum_{m=1}^K \sqrt{n_m}\mathbf{h}_m$ to separate the weights and the features from the logits $\mathbf{z} = \mathbf{W}\mathbf{h}$ in the CE loss. The constant coefficients go with each logit vector are also chosen carefully to be able to sum all logit vectors altogether, with a different amount from each class due to class-imbalance, without violating the subsequent equal conditions. After the separation of the weights and the features in logit terms, we leverage zero gradient condition of critical points to further simplify the loss function to have features as the remaining optimization variables. Then we finish the bounding and study the equal conditions.

We discuss the implications of Theorem 4.1 as following.

Optimal features form a General Orthogonal Frame: As we observe from Equation (4) in Theorem 4.1, every global solution exhibits the $\mathcal{NC}1$, i.e., within-class features collapse to their class-mean. Under the nonnegativity constraint, the optimal features form a general orthogonal frame (GOF), which consists of pairwise orthogonal vectors but with different lengths. The geometry of the optimal features for UFM class-imbalanced training problem with MSE loss is also a GOF (Dang et al., 2023), but the lengths are clearly

different between two losses. For UFM$_+$ imbalanced with supervised contrastive loss and normalized features (i.e., $\|\mathbf{h}\| = 1$), it is observed that optimal $\mathbf{H}$ exhibits OF structure with equal length class-mean vectors, irrespective to the imbalanced level (Kini et al., 2023).

Classifier converges to scaled-and-centered class-means: The optimal classifier $\mathbf{w}_k$ of problem (2) with CE loss does not form an orthogonal structure as in the case of MSE loss class-imbalance (Dang et al., 2023). Our results indicate that class k's classifier, $\mathbf{w}_k$, is aligned with the scaled and centered class-mean $\mathbf{h}_k$, with scaling factor $\sqrt{n_k}$, i.e., $\mathbf{w}_k \propto K\sqrt{n_k}\mathbf{h}_k - \sum_{m=1}^K \sqrt{n_m}\mathbf{h}_m$. We note that the proportional ratio between $\mathbf{w}_k$ and $K\sqrt{n_k}\mathbf{h}_k - \sum_{m=1}^K \sqrt{n_m}\mathbf{h}_m$ is identical across k's. This property generalizes the original $\mathcal{NC}3$ - Convergence to self-duality property, $\mathbf{w}_k \propto K\mathbf{h}_K - \sum_{m=1}^K \mathbf{h}_m$ in class-balanced setting. From Eqn. (5), (6) and (7), we can readily derive the $\mathcal{NC}_2$ - Geometry of the class-means and the classifiers in this setting. We prove that the original $\mathcal{NC}_2$ property in class-balanced setting is a special case of our result in Corrolary 4.2 below.

Logit matrix and Margin: Each column $\mathbf{z}_k$ of the logit matrix $\mathbf{Z} = \mathbf{W}\mathbf{H}$ is of a factor of the vector $\mathbf{y}_k - \frac{1}{K}\mathbf{1}_k$, but the factors are different among classes. Thus, the optimal matrix $\mathbf{Z} = \mathbf{W}\mathbf{H}$ of the problem (2) is different from the SELI geometry, i.e., the global structure of the UFM-SVM imbalanced problem. This observation further confirms Proposition 1 in (Thrampoulidis et al., 2022), which asserts that SELI is not the optimal structure for the CE imbalanced problem for any finite regularization parameter $\lambda > 0$. Furthermore, we find that the optimal classifier weight and features of the problem (2) are also different from those of SELI, for both finite ($\lambda > 0$) and vanishing regularization levels ($\lambda \to 0$). See Appendix C for the details.

The margin for any data point $\boldsymbol{x}_{k,i}$ from class k is:

$$q_{k,i}(\mathbf{W}, \mathbf{H}) = \mathbf{w}_k^\top \mathbf{h}_{k,i} - \max_{j \neq k} \mathbf{w}_j^\top \mathbf{h}_{k,i} = M_k. \quad (10)$$

We derive the results of Theorem 4.1 in the special case of balanced dataset as follows.

Corollary 4.2 (Balanced dataset as a special case)**.** *Under balanced setting where $n_1 = n_2 = \ldots = n_K$, we have from Eqn. (7) that $\mathbf{W} \propto (\overline{\mathbf{H}} - \mathbf{h}_G\mathbf{1}_K^\top)^\top$, and thus,*

$$\overline{\mathbf{H}}^\top\overline{\mathbf{H}} \propto \mathbf{I}_K,$$

$$\mathbf{W}\mathbf{W}^\top \propto (\overline{\mathbf{H}} - \mathbf{h}_G\mathbf{1}_K^\top)^\top(\overline{\mathbf{H}} - \mathbf{h}_G\mathbf{1}_K^\top) \propto \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top,$$

$$\mathbf{Z} = \mathbf{W}\overline{\mathbf{H}} \propto (\overline{\mathbf{H}} - \mathbf{h}_G\mathbf{1}_K^\top)^\top\overline{\mathbf{H}} \propto \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top.$$

Proof. The results are directly obtained from Theorem 4.1 and by noting that $M_1 = M_2 = \ldots = M_K$. When $\mathbf{H}$ forms an OF, the center class-mean matrix $\mathbf{H} - \mathbf{h}_G\mathbf{1}_K^\top$ is a simplex ETF. This follows from

$$\begin{aligned}
&(\mathbf{H} - \mathbf{h}_G\mathbf{1}_K^\top)^\top(\mathbf{H} - \mathbf{h}_G\mathbf{1}_K^\top) \\
&= (\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)^\top\mathbf{H}^\top\mathbf{H}(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top) \\
&\propto (\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)^\top\mathbf{I}_K(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top) \\
&= \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top.
\end{aligned}$$

The logit matrix $\mathbf{Z}$ also forms an ETF structure because

$$\begin{aligned}
\mathbf{Z} = \mathbf{W}\mathbf{H} &\propto (\mathbf{H} - \mathbf{h}_G\mathbf{1}_K^\top)^\top\mathbf{H} \\
&= (\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)^\top\mathbf{H}^\top\mathbf{H} = \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top.
\end{aligned}$$

We obtain the conclusion of Corollary 4.2. □

For the special case where dataset is balanced, Theorem 4.1 recovers the ETF structure for classifier matrix $\mathbf{W}$ and logit matrix $\mathbf{Z}$. The optimal class-mean matrix forms an orthogonal frame since it is constrained to be on non-negative orthant.

4.1. Classifier Norm and Angle

By expressing the geometry of the optimal solutions explicitly, Theorem 4.1 allows us to derive closed-form expressions for the norms and angles between any individual classifiers and features. Under imbalanced regime, the k-th class classifier $\mathbf{w}_k \propto K\sqrt{n_k}\mathbf{h}_k - \sum_{m=1}^K \sqrt{n_m}\mathbf{h}_m$, indicating that its norm is positively correlated with the number of sample n_k. We study the norm and angle of the classifier in the following proposition.

Proposition 4.3 (Classifers norm and angle)**.** *Let $\alpha = \frac{1}{K\sqrt{K(K-1)}}\sqrt{\frac{\lambda_H}{\lambda_W}}$. The optimal classifier $\{\mathbf{w}_k\}_{k=1}^K$ of problem (2) obeys:*

$$\|\mathbf{w}_k\|^2 = \alpha\left((K-1)^2\sqrt{n_k}M_k + \sum_{m\neq k}\sqrt{n_m}M_m\right),$$

$$\begin{aligned}
\mathbf{w}_k^\top\mathbf{w}_j = \alpha\Bigg[&-(K-1)\sqrt{n_k}M_k - (K-1)\sqrt{n_j}M_j \\
&+ \sum_{m\neq k,j}\sqrt{n_m}M_m\Bigg], \forall\, k \neq j,
\end{aligned}$$

$$\cos(\mathbf{w}_k, \mathbf{w}_j) = \frac{\mathbf{w}_k^\top\mathbf{w}_j}{\|\mathbf{w}_k\|\|\mathbf{w}_j\|}.$$

The fact that the weight norm is positively correlated with the number of training instances has been studied in the literature (Kang et al., 2019; Huang et al., 2016; Kim & Kim,

2020). Our result in Proposition 4.3 supports this observation. We proceed to derive the norm ratio of the classifier weight and class-means and the angles of the classifiers when we will have only two group of classes with equal number of samples in each group.

Corollary 4.4 (Norm ratios)**.** *Suppose $d \geq K$ and $(\mathbf{W}, \mathbf{H})$ is a global minimizer of problem (2). Then, for any $i, j \in [K]$, we have*

$$\frac{\|\mathbf{w}_i\|^2}{\|\mathbf{w}_j\|^2} = \frac{(K-1)^2\sqrt{n_i}M_i + \sum_{m\neq i}\sqrt{n_m}M_m}{(K-1)^2\sqrt{n_j}M_j + \sum_{m\neq j}\sqrt{n_m}M_m}$$
$$\frac{\|\mathbf{h}_i\|^2}{\|\mathbf{h}_j\|^2} = \sqrt{\frac{n_j}{n_i}}\frac{M_i}{M_j} \tag{11}$$

As a consequence, if $n_i \geq n_j$, $\|\mathbf{w}_i\| \geq \|\mathbf{w}_j\|$.

Proof. The results are direct consequences of Proposition 4.3. □

For the norm of optimal features, one might expect it is negatively correlated with the number of training samples and the results for UFM-MSE and UFM-SVM training problem agree with this expectation (Dang et al., 2023; Thrampoulidis et al., 2022). However, for CE loss, we find that this statement is not always true because the function $M_i/\sqrt{n_i}$ is not always a decreasing function with respect to n_i.

Corollary 4.5 (Classifer angles)**.** *Assume the dataset has K_A majority classes with n_A samples per class and K_B minority classes with n_B samples per class, then we have*

$$\cos(\mathbf{w}_{major}, \mathbf{w}'_{major}) \tag{12}$$
$$= 1 - \frac{K^2\sqrt{n_A}M_A}{K(K-1)\sqrt{n_A}M_A - K_B\sqrt{n_A}M_A + K_B\sqrt{n_B}M_B},$$
$$\cos(\mathbf{w}_{minor}, \mathbf{w}'_{minor}) \tag{13}$$
$$= 1 - \frac{K^2\sqrt{n_B}M_B}{K(K-1)\sqrt{n_B}M_B - K_A\sqrt{n_B}M_B + K_A\sqrt{n_A}M_A}.$$

Consequently, we have:

$$\cos(\mathbf{w}_{major}, \mathbf{w}'_{major}) < \cos(\mathbf{w}_{minor}, \mathbf{w}'_{minor}).$$

Proof. The results are direct consequences of Proposition 4.3. □

From Corollary 4.5, we deduce that the angle between classifier of major classes will form larger angles than those of minor classes. This observation further explains the smaller volume of feature space allocated for minority classes, which is one of the main reasons for the drop in model performance for these classes. Additionally, since Eqn. (12) and (13) are true for any pair of classes within the same category (i.e., major or minor), this means that classifiers of classes with the same number of training instances have the *same pairwise angle*.

4.2. Heavy Imbalances Cause Minority Collapse and Complete Collapse

Data naturally exhibit imbalance in their class distribution. Models trained on highly-skewed class distribution data tends to be biased towards the majority classes, resulting in poor performance on the minority classes (Huang et al., 2016; Kang et al., 2019; Kim & Kim, 2020). Especially, (Fang et al., 2021) observes that when the imbalance ratio $R := n_{\text{major}}/n_{\text{minor}}$ is larger than some threshold, the angle between minority classifiers becomes zero, and these classifiers have the same length. Consequently, these classifiers become indistinguishable and the network would predict the same probabilities for these minor classes. This phenomenon is termed as Minority Collapse. From Theorem 4.1, we obtain the exact threshold of the Minority Collapse occurrence for every class for training problem (2), in terms of the number of training samples and hyperparameters.

Corollary 4.6 (Minority Collapse and Complete Collapse)**.** *For any class $k \in [K]$, if $n_k \leq C(N, K, \lambda_W, \lambda_H) := N^2\frac{K}{K-1}\lambda_W\lambda_H$, then at the optimal solution of problem (2), $\mathbf{h}_k = 0$ and $\mathbf{w}_k = \mathbf{w}_{k'}$ for any $k' \in [K]$ such that $n_{k'} \leq C(N, K, \lambda_W, \lambda_H)$.*

(a) Minority Collapse: If the dataset has K_A majority classes with n_A samples per class and K_B minority classes with n_B samples per class, then Minority Collapse happens if the imbalance ratio

$$R := \frac{n_A}{n_B} \geq \frac{1}{K_A}\left(\frac{K-1}{NK\lambda_W\lambda_H} - K_B\right), \tag{14}$$

with $K = K_A + K_B$ and $N = n_AK_A + n_BK_B$.

(b) Complete Collapse: If

$$\frac{N^2}{n_A} \geq \frac{K-1}{K\lambda_W\lambda_H}, \tag{15}$$

then all classes collapse and the optimal solution is trivial, i.e., $(\mathbf{W}, \mathbf{H}) = (\mathbf{0}, \mathbf{0})$.

Proof. The results are direct consequences of Theorem 4.1. □

Corollary 4.6 implies that even the head classes will collapse when the ratio N^2/n_A is large enough, i.e., the dataset has a huge amount of samples or has too many classes. The bound (15) suggests that we should lower the regularization level to avoid this *complete collapse* phenomenon.

5. Experimental Results

5.1. Metric definition

We recall the notation $\mathbf{h}_k := \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{k,i}$, i.e., the class-means of class k and $\mathbf{h}_G := \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbf{h}_{k,i}$ is the feature global-mean. We calculate the within-class covariance

matrix $\mathbf{\Sigma}_W \coloneqq \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}(\mathbf{h}_{k,i}-\mathbf{h}_k)(\mathbf{h}_{k,i}-\mathbf{h}_k)^\top$ and the between-class covariance matrix $\mathbf{\Sigma}_B \coloneqq \frac{1}{K}\sum_{k=1}^{K}(\mathbf{h}_k - \mathbf{h}_G)(\mathbf{h}_k-\mathbf{h}_G)^\top$.

Feature collapse. Following previous works (Papyan et al., 2020; Han et al., 2021; Zhu et al., 2021; Tirer & Bruna, 2022), we measure feature collapse using $\mathcal{NC}1$ metric

$$\mathcal{NC}1 \coloneqq \frac{1}{K}\text{trace}(\mathbf{\Sigma}_W\mathbf{\Sigma}_B^\dagger),$$

where $\mathbf{\Sigma}_B^\dagger$ is the Moore-Penrose inverse of $\mathbf{\Sigma}_B$.

Relation between the classifier $\mathbf{W}$ and features $\mathbf{H}$. To verify the relation in Eqn. (7) in Theorem 4.1, we measure the similarity between the learned classifier $\mathbf{W}$ and the UFM_+ structure described as follows

$$\mathcal{NC}2-(\mathbf{W}-\overline{\mathbf{H}}^\top) \coloneqq \left\| \frac{\mathbf{W}}{\|\mathbf{W}\|_F} - \frac{\mathbf{W}_{\text{UFM}_+}(\mathbf{H})}{\|\mathbf{W}_{\text{UFM}_+}(\mathbf{H})\|_F}\right\|_F,$$

$$\text{where } \mathbf{W}_{\text{UFM}_+}(\mathbf{H}) = \begin{bmatrix} K\sqrt{n_1}\mathbf{h}_1^\top - \sum_{m=1}^{K}\sqrt{n_m}\mathbf{h}_m^\top \\ K\sqrt{n_2}\mathbf{h}_2^\top - \sum_{m=1}^{K}\sqrt{n_m}\mathbf{h}_m^\top \\ \cdots \\ K\sqrt{n_K}\mathbf{h}_K^\top - \sum_{m=1}^{K}\sqrt{n_m}\mathbf{h}_m^\top \end{bmatrix}.$$

Classifier and Class-means Gram matrix. We verify the geometry of the classifier and class-means matrix as follows,

$$\mathcal{NC}2-(\mathbf{W}\mathbf{W}^\top) \coloneqq \left\| \frac{\mathbf{W}\mathbf{W}^\top}{\|\mathbf{W}\mathbf{W}^\top\|_F} - \frac{\mathbf{W}\mathbf{W}_{\text{UFM}}^\top}{\|\mathbf{W}\mathbf{W}_{\text{UFM}}^\top\|_F}\right\|,$$

$$\mathcal{NC}2-(\overline{\mathbf{H}}^\top\overline{\mathbf{H}}) \coloneqq \left\| \frac{\overline{\mathbf{H}}^\top\overline{\mathbf{H}}}{\|\overline{\mathbf{H}}^\top\overline{\mathbf{H}}\|_F} - \frac{\overline{\mathbf{H}}^\top\overline{\mathbf{H}}_{\text{UFM}}}{\|\overline{\mathbf{H}}^\top\overline{\mathbf{H}}_{\text{UFM}}\|_F}\right\|,$$

where $\mathbf{W}\mathbf{W}_{\text{UFM}}^\top$ and $\overline{\mathbf{H}}^\top\overline{\mathbf{H}}_{\text{UFM}}$ are derived from Theorem 4.1 (see details in Appendix A).

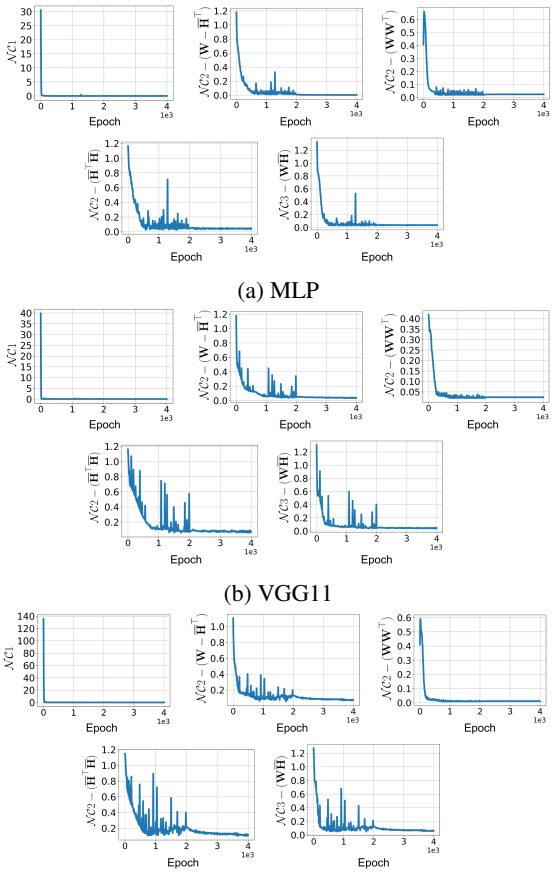
Prediction matrix $\mathbf{W}\overline{\mathbf{H}}$. To measure the similarity of the learned $\mathbf{Z} = \mathbf{W}\overline{\mathbf{H}}$ to the UFM_+ structure described in Theorem 4.1, we define $\mathcal{NC}3$ metric as follows

$$\mathcal{NC}3-(\mathbf{W}\overline{\mathbf{H}}) \coloneqq \left\| \frac{\mathbf{W}\overline{\mathbf{H}}}{\|\mathbf{W}\overline{\mathbf{H}}\|_F} - \frac{\mathbf{W}\overline{\mathbf{H}}_{\text{UFM}_+}}{\|\mathbf{W}\overline{\mathbf{H}}_{\text{UFM}_+}\|_F}\right\|_F,$$

where $\mathbf{W}\overline{\mathbf{H}}_{\text{UFM}_+}$ are described in Appendix A.

5.2. Experiment details

To verify our theoretical results, we train networks that mimic the UFM setting with ReLU features as in Eqn.(2). In particular, we use a 6-layer multilayer perceptron (MLP) model with ReLU activation, VGG11 (Simonyan & Zisserman, 2014), ResNet18 (He et al., 2016) as our three main backbone feature extractors. We train these models on the imbalanced subsets of 4 datasets: MNIST, FashionMNIST, CIFAR10, and CIFAR100. Then we measure the evolution



(a) MLP

(b) VGG11

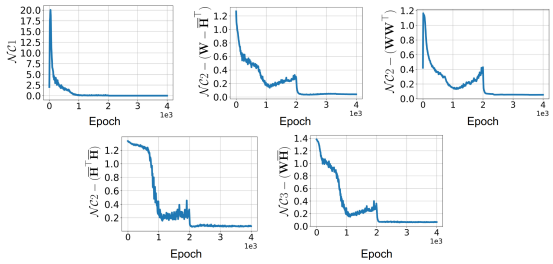
(c) ResNet18

Figure 1. $\mathcal{NC}$ metrics evolution for three models trained on imbalanced subset of CIFAR10 dataset with cross entropy loss.

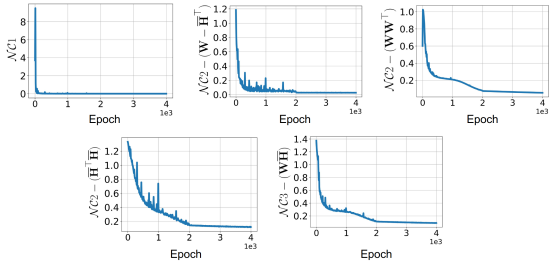
of five $\mathcal{NC}$ metrics in Section 5.1 to study the geometry of the last-layer features and the classifier. Due to space consideration, we show the results for CIFAR10 and CIFAR100 below. The remaining experiments and the training details can be found in Appendix A.

Image classification experiment on CIFAR10: For this experiment, a subset of the CIFAR10 dataset with $\{1000, 1000, 2000, 2000, 3000, 3000, 4000, 4000, 5000, 5000\}$ random samples per class is utilized as training data. We train each backbone model with Adam optimizer with batch size 256, the weight decay is $\lambda_W = 1\times10^{-4}$. Feature decay λ_H is set to 1×10^{-5} for MLP and VGG11, and to 1×10^{-4} for ResNet18. In Figure 1, we observe the convergence of $\mathcal{NC}$ metrics to small values as training progresses, which corroborates our theoretical prediction.

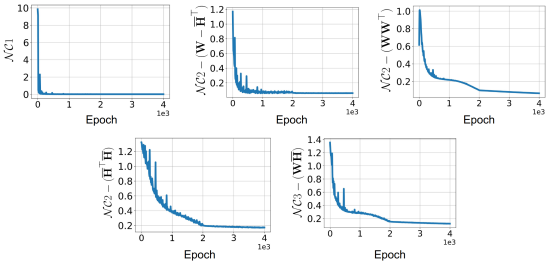
Image classification experiment on CIFAR100: We create a random subset of the CIFAR100 dataset with 100 samples per class for the first 20 classes, 200 samples per class for the next 20 classes,..., 500 samples per class for the remaining 20 classes. Each backbone model is then trained with Adam optimizer with batch size 256, the learning rate is 2×10^{-4}



(a) MLP



(b) VGG11



(c) ResNet18

Figure 2. $\mathcal{NC}$ metrics evolution for three models trained on imbalanced subset of CIFAR100 dataset with cross entropy loss.

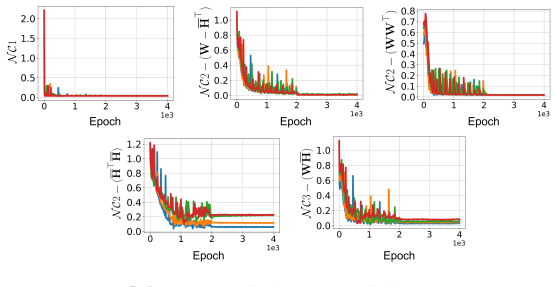
for VGG11, ResNet18 and 1×10^{-4} for MLP. Weight decay λ_W and feature decay λ_H is set to 1×10^{-4} and 1×10^{-5}, respectively. Figure 2 empirically verifies Theorem 4.1 in this setting with a large number of classes ($K = 100$).

Varying imbalance ratio: In this experiment, we validate our theoretical predictions for multiple levels of data imbalance. We train MLP and VGG models on random subsets of the CIFAR10 and MNIST datasets with varying imbalance ratios ($R = 5, 10, 20, 50$). Figures 3 and 7 demonstrate the convergence of $\mathcal{NC}$ metrics for MNIST and CIFAR10 datasets, respectively.

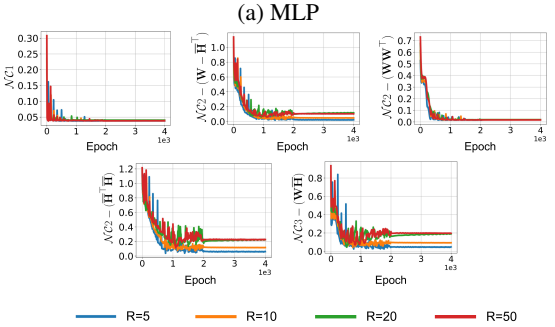
Illustration of $\overline{\mathbf{H}}^\top \overline{\mathbf{H}}$: We normalize the $\overline{\mathbf{H}}^\top \overline{\mathbf{H}}$ matrix obtained from the last epoch of the MLP model trained on the CIFAR10 dataset. The orthogonal structure of the learned features along with the theoretical prediction derived from Theorem 4.1 are demonstrated in Figure 4.

6. Concluding Remarks

In this work, we present a rigorous and explicit study of Neural Collapse phenomenon in the setting of imbalanced dataset using unconstrained non-negative features model



(a) MLP



(b) VGG11

Figure 3. $\mathcal{NC}$ metrics evolution of MLP and VGG11 backbone trained on MNIST imbalanced subsets with cross entropy loss

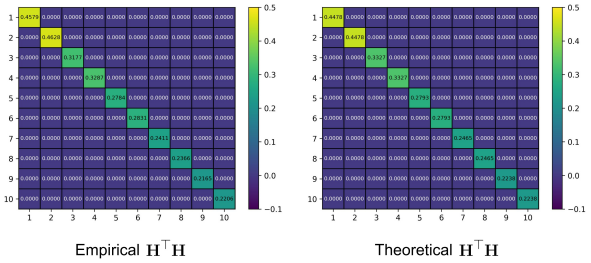


Figure 4. $\overline{\mathbf{H}}^\top \overline{\mathbf{H}}$ matrix extracted from the last epoch of the trained MLP model.

and cross-entropy loss. In particular, we provide a closed-form characterization of the last-layer features and classifier weights learned by the network training. We find that while the variability collapse property still holds, the geometry of the learned features and learned classifier weights are different from the original definition of Neural Collapse, due to the class-imbalance of the training data. Specifically, we prove that at optimality, the features form an orthogonal structure while the classifier weights are aligned to the scaled and centered class-means, which generalizes the original definition of Neural Collapse in class-balanced settings. Furthermore, with closed-form derivations of the solution, we are able to quantify the norms and the angles between the learned features and classifier weights across class distribution. As a limitation, we only study the convergence geometries under the condition that the feature dimension d is at least the number of classes K. The geometric structure of the features and classifier in the bottleneck situation $d < K$ is still unaddressed and we leave it for future work.

Acknowledgements

This research/project is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-012-SGIL). NH acknowledges support from the NSF IFML 2019844 and the NSF AI Institute for Foundations of Machine Learning.

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.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss, 2019. URL https://arxiv.org/abs/1906.07413.

Dang, H., Nguyen, T., Tran, T., Tran, H., and Ho, N. Neural collapse in deep linear network: From balanced to imbalanced data. *arXiv preprint arXiv:2301.00437*, 2023.

Ergen, T. and Pilanci, M. Revealing the structure of deep neural networks via convex duality, 2020. URL https://arxiv.org/abs/2002.09773.

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), oct 2021. doi: 10.1073/pnas.2103091118. URL https://doi.org/10.1073%2Fpnas.2103091118.

Graf, F., Hofer, C. D., Niethammer, M., and Kwitt, R. Dissecting supervised contrastive learning, 2023.

Han, X. Y., Papyan, V., and Donoho, D. L. Neural collapse under mse loss: Proximity to and dynamics on the central path, 2021. URL https://arxiv.org/abs/2106.02073.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Hong, W. and Ling, S. Neural collapse for unconstrained feature model under cross-entropy loss with imbalanced data. *arXiv preprint arXiv:2309.09725*, 2023.

Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(91)90009-T. URL https://www.sciencedirect.com/science/article/pii/089360809190009T.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90020-8. URL https://www.sciencedirect.com/science/article/pii/0893608089900208.

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

Ji, W., Lu, Y., Zhang, Y., Deng, Z., and Su, W. J. An unconstrained layer-peeled perspective on neural collapse, 2021. URL https://arxiv.org/abs/2110.02796.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition, 2019. URL https://arxiv.org/abs/1910.09217.

Kang, B., Li, Y., Xie, S., Yuan, Z., and Feng, J. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020.

Kim, B. and Kim, J. Adjusting decision boundary for class imbalanced learning. *IEEE Access*, 8:81674–81685, 2020.

Kini, G. R., Vakilian, V., Behnia, T., Gill, J., and Thrampoulidis, C. Supervised-contrastive loss learns orthogonal frames and batching matters. *arXiv preprint arXiv:2306.07960*, 2023.

Liu, X., Zhang, J., Hu, T., Cao, H., Yao, Y., and Pan, L. Inducing neural collapse in deep long-tailed learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 11534–11544. PMLR, 2023.

Lu, J. and Steinerberger, S. Neural collapse with cross-entropy loss, 2020. URL https://arxiv.org/abs/2012.08465.

Nguyen, D. A., Levie, R., Lienen, J., Hüllermeier, E., and Kutyniok, G. Memorization-dilation: Modeling neural collapse under noise. In *The Eleventh International Conference on Learning Representations*, 2022.

Papyan, V., Han, X. Y., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *CoRR*, abs/2008.08186, 2020. URL https://arxiv.org/abs/2008.08186.

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, 2022. doi: 10.1109/ICASSP43922.2022.9746778.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition, 2014. URL https://arxiv.org/abs/1409.1556.

Súkeník, P., Mondelli, M., and Lampert, C. Deep neural collapse is provably optimal for the deep unconstrained features model. *arXiv preprint arXiv:2305.13165*, 2023.

Thrampoulidis, C., Kini, G. R., Vakilian, V., and Behnia, T. Imbalance trouble: Revisiting neural-collapse geometry, 2022. URL https://arxiv.org/abs/2208.05512.

Tirer, T. and Bruna, J. Extended unconstrained features model for exploring deep neural collapse, 2022. URL https://arxiv.org/abs/2202.08087.

Tirer, T., Huang, H., and Niles-Weed, J. Perturbation analysis of neural collapse. In *International Conference on Machine Learning*, pp. 34301–34329. PMLR, 2023.

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?, 2022. URL https://arxiv.org/abs/2203.09081.

Yarotsky, D. Universal approximations of invariant maps by neural networks, 2018. URL https://arxiv.org/abs/1804.10306.

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

Zhou, D.-X. Universality of deep convolutional neural networks, 2018. URL https://arxiv.org/abs/1805.10769.

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, 2022a. URL https://arxiv.org/abs/2203.01238.

Zhou, J., You, C., Li, X., Liu, K., Liu, S., Qu, Q., and Zhu, Z. Are all losses created equal: A neural collapse perspective, 2022b. URL https://arxiv.org/abs/2210.02192.

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. A geometric analysis of neural collapse with unconstrained features. *CoRR*, abs/2105.02375, 2021. URL https://arxiv.org/abs/2105.02375.

Appendix for "Neural Collapse for Cross-entropy Class-Imbalanced Learning with Unconstrained ReLU Features Model"

A. Additional Experiments and Network Training details

A.1. Metric definitions

We define the $\mathcal{NC}$ metrics used in our experiments to measure the discrepancy between the learned model and our derived geometry for the last-layer features and classifier. We recall the notation $\mathbf{h}_k \coloneqq \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{k,i}$, i.e., the class-means of class k and $\mathbf{h}_G \coloneqq \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbf{h}_{k,i}$ is the feature global-mean. We calculate the within-class covariance matrix $\mathbf{\Sigma}_W \coloneqq \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}(\mathbf{h}_{k,i}-\mathbf{h}_k)(\mathbf{h}_{k,i}-\mathbf{h}_k)^\top$ and the between-class covariance matrix $\mathbf{\Sigma}_B \coloneqq \frac{1}{K}\sum_{k=1}^{K}(\mathbf{h}_k-\mathbf{h}_G)(\mathbf{h}_k-\mathbf{h}_G)^\top$. The class-mean matrix is denoted as $\overline{\mathbf{H}}$. $\{M_k\}_{k=1}^{K}$ are the constants defined in Eqn. (3) in our main paper.

Feature collapse:

$$\mathcal{NC}1 \coloneqq \frac{1}{K}\text{trace}(\mathbf{\Sigma}_W\mathbf{\Sigma}_B^{\dagger}),$$

where $\mathbf{\Sigma}_B^{\dagger}$ is the Moore-Penrose inverse of $\mathbf{\Sigma}_B$.

Relation between the classifier $\mathbf{W}$ and features $\mathbf{H}$:

$$\mathcal{NC}2-(\mathbf{W}-\overline{\mathbf{H}}^\top) \coloneqq \left\|\frac{\mathbf{W}}{\|\mathbf{W}\|_F}-\frac{\mathbf{W}_{\text{UFM}_+}(\mathbf{H})}{\|\mathbf{W}_{\text{UFM}_+}(\mathbf{H})\|_F}\right\|_F, \text{where } \mathbf{W}_{\text{UFM}_+}(\mathbf{H}) = \begin{bmatrix} K\sqrt{n_1}\mathbf{h}_1^\top - \sum_{m=1}^{K}\sqrt{n_m}\mathbf{h}_m^\top \\ K\sqrt{n_2}\mathbf{h}_2^\top - \sum_{m=1}^{K}\sqrt{n_m}\mathbf{h}_m^\top \\ \cdots \\ K\sqrt{n_K}\mathbf{h}_K^\top - \sum_{m=1}^{K}\sqrt{n_m}\mathbf{h}_m^\top \end{bmatrix}.$$

Classifier Gram matrix $\mathbf{W}\mathbf{W}^\top$:

$$\mathcal{NC}2-(\mathbf{W}\mathbf{W}^\top) \coloneqq \left\|\frac{\mathbf{W}\mathbf{W}^\top}{\|\mathbf{W}\mathbf{W}^\top\|_F}-\frac{\mathbf{W}\mathbf{W}_{\text{UFM}}^\top}{\|\mathbf{W}\mathbf{W}_{\text{UFM}}^\top\|_F}\right\|,$$

$$\text{where } (\mathbf{W}\mathbf{W}_{\text{UFM}}^\top)_{kk} = \|\mathbf{w}_k\|^2 = \alpha\left((K-1)^2\sqrt{n_k}M_k + \sum_{m\neq k}\sqrt{n_m}M_m\right),$$

$$\text{and } (\mathbf{W}\mathbf{W}_{\text{UFM}}^\top)_{kj} = \mathbf{w}_k^\top\mathbf{w}_j = \alpha\left[-(K-1)\sqrt{n_k}M_k-(K-1)\sqrt{n_j}M_j+\sum_{m\neq k,j}\sqrt{n_m}M_m\right], \forall\, k\neq j.$$

The calculation of $\mathbf{W}\mathbf{W}_{\text{UFM}}^\top$ is from the Proposition 4.3.

Class-mean Gram matrix $\overline{\mathbf{H}}^\top\overline{\mathbf{H}}$:

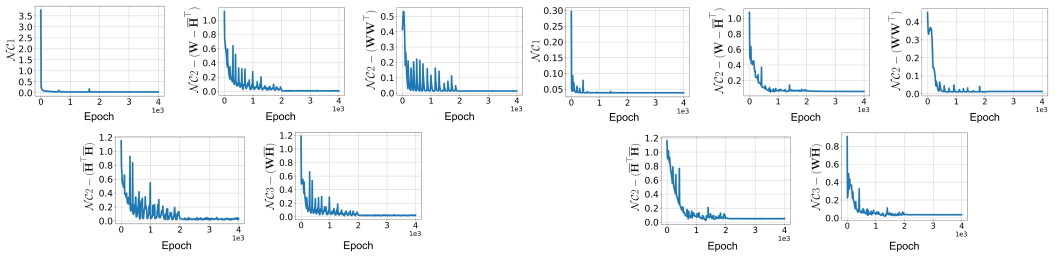
$$\mathcal{NC}2-(\overline{\mathbf{H}}^\top\overline{\mathbf{H}}) \coloneqq \left\|\frac{\overline{\mathbf{H}}^\top\overline{\mathbf{H}}}{\|\overline{\mathbf{H}}^\top\overline{\mathbf{H}}\|_F}-\frac{\overline{\mathbf{H}}^\top\overline{\mathbf{H}}_{\text{UFM}}}{\|\overline{\mathbf{H}}^\top\overline{\mathbf{H}}_{\text{UFM}}\|_F}\right\|,$$

$$\text{where } (\overline{\mathbf{H}}^\top\overline{\mathbf{H}}_{\text{UFM}})_{kk} = \|\mathbf{h}_k\|^2 = \sqrt{\frac{K-1}{K}\frac{\lambda_W}{\lambda_H}\frac{1}{n_k}}M_k \text{ and } (\overline{\mathbf{H}}^\top\overline{\mathbf{H}}_{\text{UFM}})_{kj} = 0.$$

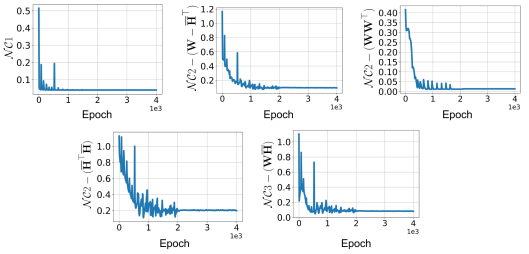
The calculation of $\overline{\mathbf{H}}^\top\overline{\mathbf{H}}_{\text{UFM}}$ is from Theorem 4.1.

Prediction matrix $\mathbf{W}\overline{\mathbf{H}}$:

$$\mathcal{NC}3-\mathbf{W}\overline{\mathbf{H}} \coloneqq \left\|\frac{\mathbf{W}\mathbf{H}}{\|\mathbf{W}\mathbf{H}\|_F}-\frac{\mathbf{W}\mathbf{H}_{\text{UFM}_+}}{\|\mathbf{W}\mathbf{H}_{\text{UFM}_+}\|_F}\right\|_F, \text{where } \mathbf{W}\mathbf{H}_{\text{UFM}_+} = \begin{bmatrix} \frac{K-1}{K}M_1 & \frac{-1}{K}M_2 & \dots & \frac{-1}{K}M_K \\ \frac{-1}{K}M_1 & \frac{K-1}{K}M_2 & \dots & \frac{-1}{K}M_K \\ \dots & \dots & \dots & \\ \frac{-1}{K}M_1 & \frac{-1}{K}M_2 & \dots & \frac{K-1}{K}M_K \end{bmatrix}.$$

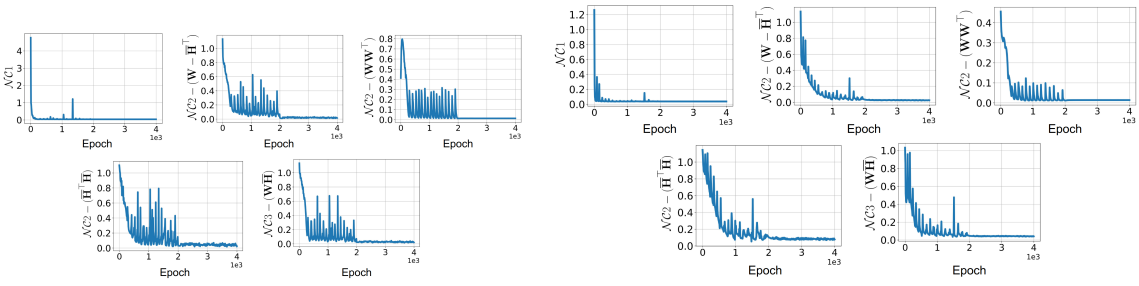


(a) MLP (b) VGG11

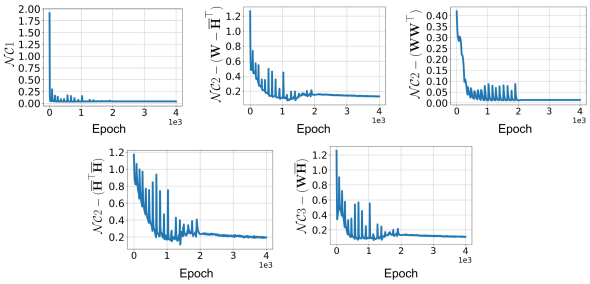


(c) ResNet18

Figure 5. $\mathcal{NC}$ metrics evolution for three models trained on imbalanced subset of MNIST dataset with cross entropy loss.

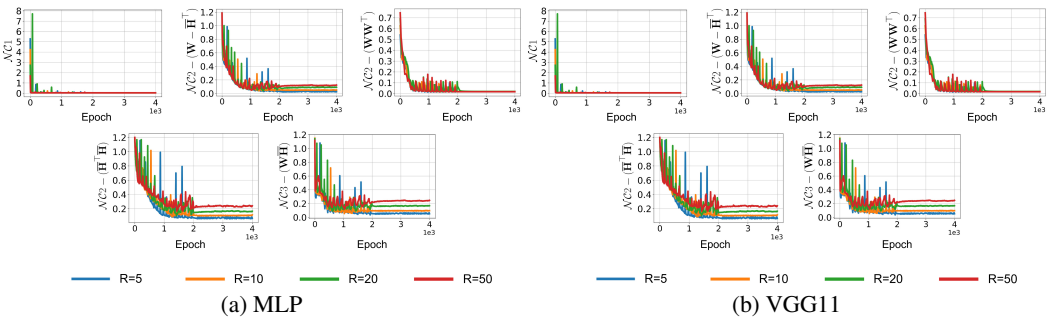


(a) MLP (b) VGG11



(c) ResNet18

Figure 6. $\mathcal{NC}$ metrics evolution for three models trained on imbalanced subset of FashionMNIST dataset with cross entropy loss.



(a) MLP (b) VGG11

Figure 7. $\mathcal{NC}$ metrics evolution of MLP and VGG11 backbone trained on CIFAR10 imbalanced subsets with cross entropy loss

A.2. Network training details

Unless stated otherwise, all models in Section 5.2 are trained for 4000 epochs with Adam optimizer, we set the general learning rate to 1×10^{-3} with decay of 0.1 at 2000-th epoch. All MLP models share the same hidden dimension of 1024.

Image classification experiment on CIFAR10: For this experiment, a subset of the CIFAR10 dataset with $\{1000, 1000, 2000, 2000, 3000, 3000, 4000, 4000, 5000, 5000\}$ random samples per class is utilized as training data. We train each backbone model with Adam optimizer with batch size 256, the weight decay is $\lambda_W = 1 \times 10^{-4}$. Feature decay λ_H is set to 1×10^{-5} for MLP and VGG11, and to 1×10^{-4} for ResNet18.

Image classification experiment on CIFAR100: We create a random subset of the CIFAR100 dataset with 100 samples per class for the first 20 classes, 200 samples per class for the next 20 classes,..., 500 samples per class for the remaining 20 classes. Each backbone model is then trained with Adam optimizer with batch size 256, the learning rate is 2×10^{-4} for VGG11, ResNet18 and 1×10^{-4} for MLP. Weight decay λ_W and feature decay λ_H is set to 1×10^{-4} and 1×10^{-5}, respectively.

Image classification experiment on MNIST dataset: In this experiment, a randomly sampled subset of MNIST dataset with the number of samples per class $\in \{100, 100, 200, 200, 300, 300, 400, 400, 500, 500\}$ is utilized. Each backbone model is trained with batch size 16. Feature decay rate is $\lambda_H = 1 \times 10^{-5}$ and weight decay rate is $\lambda_W = 1 \times 10^{-4}$. The results are shown in Figure 5.

Image classification experiment on FashionMNIST dataset: Similar to MNIST experiment, we randomly sample a subset of FashionMNIST dataset with $\{100, 100, 200, 200, 300, 300, 400, 400, 500, 500\}$ samples per class. Each backbone model is trained with batch size 16. Feature decay rate is $\lambda_H = 1 \times 10^{-5}$ and weight decay rate is $\lambda_W = 1 \times 10^{-4}$. The results are shown in Figure 6.

Varying imbalance ratio: Each model is trained with a batch size of 32, weight decay λ_W of 1×10^{-4}, and feature decay λ_H of 1×10^{-5}. The training data is randomly drawn from CIFAR10 and MNIST dataset with 5 majority classes with 500 samples per class, and the other 5 minority classes with $500/R$ ($R = 5, 10, 20, 50$) samples per class.

B. Proof of Theorem 4.1 and Proposition 4.3

Recall the training problem:

$$\min_{\mathbf{W},\mathbf{H}} \mathcal{L}_0(\mathbf{W}, \mathbf{H}) := \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathcal{L}_{CE}(\mathbf{W}\mathbf{h}_{k,i}, \mathbf{y}_k) + \frac{\lambda_W}{2} \|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2} \|\mathbf{H}\|_F^2, \tag{16}$$

where $\mathbf{h}_{k,i} \geq 0 \, \forall \, k, i$ and:

$$\mathcal{L}_{CE}(\mathbf{z}, \mathbf{y}_k) := -\log \left(\frac{e^{z_k}}{\sum_{m=1}^{K} e^{z_m}} \right).$$

Denoting the class-mean of k-th class as $\mathbf{h}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{h}_{k,i}$, the m-th row vector of $\mathbf{W}$ as $\mathbf{w}_m$. We denote $\mathbf{z}_{k,i} := \mathbf{W}\mathbf{h}_{k,i}$ and $\mathbf{z}^{(m)}$ is the m-th component of vector $\mathbf{z}$.

Step 1: We introduce a lower bound on the loss $\mathcal{L}_0$ by grouping the cross-entropy term and regularization term for features within the same class.

We have:

$$
\begin{aligned}
&\mathcal{L}_0(\mathbf{W},\mathbf{H}) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathcal{L}_{CE}(\mathbf{z}_{k,i},\mathbf{y}_k) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2 \\
&= \frac{1}{N}\sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\log\left(\frac{\sum_{m=1}^{K}\exp\left(\mathbf{z}_{k,i}^{(m)}\right)}{\exp\left(\mathbf{z}_{k,i}^{(k)}\right)}\right)\right) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\|\mathbf{h}_{k,i}\|^2\right) \\
&= \frac{1}{N}\sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\log\left(1+\sum_{m\neq k}\exp\left(\mathbf{z}_{k,i}^{(m)}-\mathbf{z}_{k,i}^{(k)}\right)\right)\right) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\|\mathbf{h}_{k,i}\|^2\right) \\
&\geq \frac{1}{N}\sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\log\left(1+(K-1)\exp\left(\sum_{m\neq k}\frac{\mathbf{z}_{k,i}^{(m)}-\mathbf{z}_{k,i}^{(k)}}{K-1}\right)\right)\right) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\|\mathbf{h}_{k,i}\|^2\right) \\
&= \frac{1}{N}\sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\log\left(1+(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{z}_{k,i}^{(m)}-K\mathbf{z}_{k,i}^{(k)}}{K-1}\right)\right)\right) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\|\mathbf{h}_{k,i}\|^2\right) \\
&\geq \frac{1}{N}\sum_{k=1}^{K}n_k\log\left(1+(K-1)\exp\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\frac{\sum_{m=1}^{K}\mathbf{z}_{k,i}^{(m)}-K\mathbf{z}_{k,i}^{(k)}}{K-1}\right)\right) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K}\left(\frac{1}{n_k}\left\|\sum_{i=1}^{n_k}\mathbf{h}_{k,i}\right\|^2\right) \\
&= \frac{1}{N}\sum_{k=1}^{K}n_k\log\left(1+(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{z}_{k}^{(m)}-K\mathbf{z}_{k}^{(k)}}{K-1}\right)\right) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K}n_k\left\|\mathbf{h}_k\right\|^2 \\
&= \frac{1}{N}\sum_{k=1}^{K}n_k\log\left(1+(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{w}_m\mathbf{h}_k-K\mathbf{w}_k\mathbf{h}_k}{K-1}\right)\right) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K}n_k\left\|\mathbf{h}_k\right\|^2 \\
&:= \mathcal{L}_1(\mathbf{W},\mathbf{H})
\end{aligned}
$$

where $\mathbf{z}_k := \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{z}_{k,i}$. We denote the function:

$$
g(\mathbf{WH}) := \frac{1}{N}\sum_{k=1}^{K}n_k\log\left(1+(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{w}_m\mathbf{h}_k-K\mathbf{w}_k\mathbf{h}_k}{K-1}\right)\right), \tag{17}
$$

thus, $\mathcal{L}_1(\mathbf{W},\mathbf{H}) = g(\mathbf{WH}) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K}n_k\left\|\mathbf{h}_k\right\|^2$.

The first inequality above follows from Jensen inequality that:

$$
\sum_{m\neq k}\exp\left(\mathbf{z}_{k,i}^{(m)}-\mathbf{z}_{k,i}^{(k)}\right) \geq (K-1)\exp\left(\sum_{m\neq k}\frac{\mathbf{z}_{k,i}^{(m)}-\mathbf{z}_{k,i}^{(k)}}{K-1}\right), \tag{18}
$$

which become equality when and only when $\mathbf{z}_{k,i}^{(m)} = \mathbf{z}_{k,i}^{(l)}\ \forall\, m,l\neq k$. The second inequality includes two inequalities as following. The first one is

$$
\sum_{i=1}^{n_k}\log\left(1+(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{z}_{k,i}^{(m)}-K\mathbf{z}_{k,i}^{(k)}}{K-1}\right)\right) \geq n_k\log\left(1+(K-1)\exp\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\frac{\sum_{m=1}^{K}\mathbf{z}_{k,i}^{(m)}-K\mathbf{z}_{k,i}^{(k)}}{K-1}\right)\right), \tag{19}
$$

where we use Jensen inequality since the function $\log(1+(K-1)\exp(x))$ is a convex function. The second sub-inequality is that for any $k\in[K]$, $\sum_{i=1}^{n_k}\|\mathbf{h}_{k,i}\|^2 \geq \frac{1}{n_k}\|\sum_{k=1}^{K}\mathbf{h}_{k,i}\|^2$, which achieves equality if and only if $\mathbf{h}_{k,i} = \mathbf{h}_{k,j}\ \forall\, i\neq j$. This equality condition, $\mathbf{h}_{k,i} = \mathbf{h}_{k,j}\ \forall\, i\neq j$, also satisfies the equality condition of the inequality (19), hence we only need to satisfy this property to achieve equality.

Step 2: We further lower bound the log term of $\mathcal{L}_1$, the idea of this bound is inspired from Lemma D.5 in (Zhu et al., 2021)

For any $k \in [K]$ and any $t_k > 0$, we have:

$$\begin{aligned}
&\log\left(1+(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)}-K\mathbf{z}_k^{(k)}}{K-1}\right)\right)\\
=&\log\left(\frac{t_k}{1+t_k}\frac{1+t_k}{t_k}+\frac{1}{1+t_k}(1+t_k)(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)}-K\mathbf{z}_k^{(k)}}{K-1}\right)\right)\\
\geq&\frac{1}{1+t_k}\log\left((1+t_k)(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)}-K\mathbf{z}_k^{(k)}}{K-1}\right)\right)+\frac{t_k}{1+t_k}\log\left(\frac{1+t_k}{t_k}\right)\\
=&\frac{1}{1+t_k}\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)}-K\mathbf{z}_k^{(k)}}{K-1}+\frac{1}{1+t_k}\log\left((1+t_k)(K-1)\right)+\frac{t_k}{1+t_k}\log\left(\frac{1+t_k}{t_k}\right)\\
=&\underbrace{\frac{\sqrt{n_k}}{1+t_k}}_{c_{1,k}}\sqrt{\frac{1}{n_k}}\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)}-K\mathbf{z}_k^{(k)}}{K-1}+\underbrace{\frac{1}{1+t_k}\log\left((1+t_k)(K-1)\right)+\frac{t_k}{1+t_k}\log\left(\frac{1+t_k}{t_k}\right)}_{c_{2,k}}\\
=&\frac{c_{1,k}}{K-1}\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)}-K\mathbf{z}_k^{(k)}}{\sqrt{n_k}}+c_{2,k},
\end{aligned} \tag{20}$$

where the inequality above is from the concavity of the $\log(x)$ function, i.e., $\log(tx+(1-t)y) \geq t\log(x)+(1-t)\log(y)$ for any x, y and $t \in [0,1]$. The inequality becomes an equality if any only if:

$$\frac{1+t_k}{t_k} = (1+t_k)(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)}-K\mathbf{z}_k^{(k)}}{K-1}\right) \quad \text{or} \quad t_k = 0, \quad \text{or} \quad t_k = +\infty.$$

However, when $t_k = 0$ or $t_k = +\infty$, the equality is trivial. Therefore, we have:

$$t_k = \left[(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)}-K\mathbf{z}_k^{(k)}}{K-1}\right)\right]^{-1}.$$

To summary, at this step, we have that for any $k \in [K]$ and any $t_k > 0$:

$$\log\left(1+(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)}-K\mathbf{z}_k^{(k)}}{K-1}\right)\right) \geq \frac{c_{1,k}}{K-1}\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)}-K\mathbf{z}_k^{(k)}}{\sqrt{n_k}}+c_{2,k}, \tag{21}$$

where $c_{1,k} = \sqrt{n_k}/(1+t_k)$ and $c_{2,k} = \frac{1}{1+t_k}\log\left((1+t_k)(K-1)\right)+\frac{t_k}{1+t_k}\log\left(\frac{1+t_k}{t_k}\right)$. The inequality becomes an equality when:

$$t_k = \left[(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)}-K\mathbf{z}_k^{(k)}}{K-1}\right)\right]^{-1}. \tag{22}$$

Step 3: We apply the result from **Step 2** and choose the same $c_{1,k}$ for all classes to lower bound $g(\mathbf{WH})$ w.r.t. the L2-norm of the class-mean $\|\mathbf{h}_k\|^2 := x_k$.

By using the inequality (21) for $\mathbf{z}_{k,i} = \mathbf{W}\mathbf{h}_{k,i}$ and choosing the same scalar $c_1 := c_{1,1} = \ldots = c_{1,k}$ (recall that $c_{1,k}$ can be

chosen arbitrarily), we have:

$$\begin{aligned}
&\frac{K-1}{c_1}\left[g(\mathbf{WH})-\sum_{k=1}^{K}\frac{n_k}{N}c_{2,k}\right]\\
&=\frac{K-1}{c_1}\left[\frac{1}{N}\sum_{k=1}^{K}n_k\log\left(1+(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{w}_m\mathbf{h}_k-K\mathbf{w}_k\mathbf{h}_k}{K-1}\right)\right)-\sum_{k=1}^{K}\frac{n_k}{N}c_{2,k}\right]\\
&\geq\frac{1}{N}\sum_{k=1}^{K}n_k\sqrt{\frac{1}{n_k}}\left[\sum_{m=1}^{K}\mathbf{w}_m\mathbf{h}_k-K\mathbf{w}_k\mathbf{h}_k\right]\\
&=\frac{1}{N}\sum_{m=1}^{K}\mathbf{w}_m\left(\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k-K\sqrt{n_m}\mathbf{h}_m\right).
\end{aligned}\tag{23}$$

We know that from the Cauchy-Schwarz inequality for inner product that for any $\mathbf{u},\mathbf{v}\in\mathbb{R}^K$ and any $c_3>0$,

$$\mathbf{u}^\top\mathbf{v}\geq-\frac{c_3}{2}\|\mathbf{u}\|_2^2-\frac{1}{2c_3}\|\mathbf{v}\|_2^2.$$

The equality holds when $c_3\mathbf{u}=-\mathbf{v}$. Therefore, by applying this inequality for each term $\mathbf{w}_m\left(\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k-K\sqrt{n_m}\mathbf{h}_m\right)$, we have:

$$\begin{aligned}
&\frac{N(K-1)}{c_1}\left[g(\mathbf{WH})-\sum_{k=1}^{K}\frac{n_k}{N}c_{2,k}\right]\\
&\geq-\frac{c_3}{2}\sum_{m=1}^{K}\|\mathbf{w}_m\|_2^2-\frac{1}{2c_3}\sum_{m=1}^{K}\left\|\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k-K\sqrt{n_m}\mathbf{h}_m\right\|_2^2\\
&=-\frac{c_3}{2}\|\mathbf{W}\|_F^2-\frac{1}{2c_3}\sum_{m=1}^{K}\|\hat{\mathbf{h}}_m\|_2^2,
\end{aligned}\tag{24}$$

where we denote $\hat{\mathbf{h}}_m:=\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k-K\sqrt{n_m}\mathbf{h}_m,\forall\, m\in[K]$, and the above inequality becomes an equality if and only if:

$$c_3\mathbf{w}_m=-\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k+K\sqrt{n_m}\mathbf{h}_m,\ \forall m\in[K]\tag{25}$$

We further have:

$$\begin{aligned}
\sum_{m=1}^{K}\|\hat{\mathbf{h}}_m\|^2&=\sum_{m=1}^{K}\left\|\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k-K\sqrt{n_m}\mathbf{h}_m\right\|^2\\
&=\sum_{m=1}^{K}\left(\left\|\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k\right\|^2+K^2\left\|\sqrt{n_m}\mathbf{h}_m\right\|^2-2K\left\langle\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k,\sqrt{n_m}\mathbf{h}_m\right\rangle\right)\\
&=K^2\sum_{m=1}^{K}\left\|\sqrt{n_m}\mathbf{h}_m\right\|^2+K\left\|\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k\right\|^2-2K\left\langle\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k,\sum_{m=1}^{K}\sqrt{n_m}\mathbf{h}_m\right\rangle\\
&=K^2\sum_{m=1}^{K}\left\|\sqrt{n_m}\mathbf{h}_m\right\|^2-K\left\|\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k\right\|^2
\end{aligned}\tag{26}$$

We lower bound the second term of Eqn. (26) as following:

$$\left\|\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k\right\|^2 = \sum_{k=1}^{K}\|\sqrt{n_k}\mathbf{h}_k\|^2 + \sum_{k,l,k\neq l}\langle\sqrt{n_k}\mathbf{h}_k, \sqrt{n_l}\mathbf{h}_l\rangle \geq \sum_{k=1}^{K}\|\sqrt{n_k}\mathbf{h}_k\|^2, \tag{27}$$

where we use the non-negativity of the features and the equality happens iff $\langle \mathbf{h}_k, \mathbf{h}_l\rangle = 0, \forall\, k \neq l$.

Thus, we have:

$$\sum_{m=1}^{K}\|\hat{\mathbf{h}}_m\|^2 \leq K(K-1)\sum_{k=1}^{K} n_k\|\mathbf{h}_k\|^2, \tag{28}$$

the equality happens iff $\langle \mathbf{h}_k, \mathbf{h}_l\rangle = 0, \forall\, k \neq l$.

Now, let $x_k := \|\mathbf{h}_k\|^2$, at critical points of $\mathcal{L}_1$, from Lemma B.1 , we have:

$$\|\mathbf{W}\|_F^2 = \frac{\lambda_H}{\lambda_W}\sum_{k=1}^{K} n_k x_k. \tag{29}$$

Hence:

$$\frac{N(K-1)}{c_1}\left[g(\mathbf{WH}) - \sum_{k=1}^{K}\frac{n_k}{N}c_{2,k}\right] \geq -\frac{c_3}{2}\frac{\lambda_H}{\lambda_W}\left(\sum_{k=1}^{K} n_k x_k\right) - \frac{K(K-1)}{2c_3}\left(\sum_{k=1}^{K} n_k x_k\right) \tag{30}$$

We will choose c_3 in advance to let the inequality (30) hold. From the equality conditions (25) and (28), we can choose c_3 as follows:

$$\begin{aligned} c_3\mathbf{w}_m &= -\sum_{k=1}^{K}\sqrt{n_k}\mathbf{h}_k + K\sqrt{n_j}\mathbf{h}_m,\ \forall m \in [K] \\ \Rightarrow c_3^2 &= \frac{\sum_{k=1}^{K}\|\hat{\mathbf{h}}_k\|^2}{\sum_{k=1}^{K}\|\mathbf{w}_k\|^2} = \frac{K(K-1)\left(\sum_{k=1}^{K} n_k x_k\right)}{\frac{\lambda_H}{\lambda_W}\left(\sum_{k=1}^{K} n_k x_k\right)} = \frac{\lambda_W}{\lambda_H}K(K-1) \end{aligned} \tag{31}$$

In summary, from (30), we have the lower bound of $g(\mathbf{WH})$:

$$g(\mathbf{WH}) \geq \frac{-c_1}{N}\sqrt{\frac{\lambda_H}{\lambda_W}\frac{K}{K-1}}\left(\sum_{k=1}^{K} n_k x_k\right) + \sum_{k=1}^{K}\frac{n_k}{N}c_{2,k}, \tag{32}$$

for any $c_1 > 0$. The equality conditions of (32) is derived at Lemma B.2.

Step 4: Now, we use the lower bound of $g(\mathbf{WH})$ above into the bounding of the loss $\mathcal{L}_1(\mathbf{W}, \mathbf{H})$ and use the equality conditions from Lemma B.2 to finish the bounding process.

Recall that $x_k = \|\mathbf{h}_k\|^2$, we have at critical points of $\mathcal{L}_1$:

$$\begin{aligned}\mathcal{L}_1(\mathbf{W},\mathbf{H}) &= g(\mathbf{WH}) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\sum_{k=1}^K n_k \|\mathbf{h}_k\|^2 \\ &\geq \frac{-c_1}{N}\sqrt{\frac{\lambda_H}{\lambda_W}\frac{K}{K-1}}\left(\sum_{k=1}^K n_k x_k\right) + \sum_{k=1}^K \frac{n_k}{N} c_{2,k} + \lambda_H \sum_{k=1}^K n_k x_k \\ &:= \xi(c_1, x_1, \ldots, x_K),\end{aligned}$$

for any $c_1 > 0$ ($c_{2,k}$ can be calculated from c_1). From Lemma B.2, we know that the inequality $\mathcal{L}_1(\mathbf{W},\mathbf{H}) \geq \xi(c_1, x_1, \ldots, x_k)$ becomes an equality if and only if:

$$\begin{aligned}&\mathbf{h}_k^\top \mathbf{h}_l = 0,\ \forall\, k \neq l \\ &\mathbf{w}_m = \sqrt{\frac{1}{K(K-1)}}\sqrt{\frac{\lambda_H}{\lambda_W}}\left(K\sqrt{n_m}\mathbf{h}_m - \sum_{k=1}^K \sqrt{n_k}\mathbf{h}_k\right),\ \forall m \in [K] \\ &t_k = \frac{1}{K-1}\exp\left(\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}\|\mathbf{h}_k\|^2\right), \\ &c_1 = \frac{\sqrt{n_k}}{1+\frac{1}{K-1}\exp\left(\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}\|\mathbf{h}_k\|^2\right)} = \frac{\sqrt{n_l}}{1+\frac{1}{K-1}\exp\left(\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_l}\|\mathbf{h}_l\|^2\right)},\ \forall\, k \neq l\end{aligned}$$

Next, we will lower bound $\xi(c_1, x_1, \ldots, x_K)$ under these equality conditions for arbitrary values of $x_1, \ldots, x_K$, as following:

$$\begin{aligned}&\xi(c_1, x_1, \ldots, x_K) \\ &= \frac{-c_1}{N}\sqrt{\frac{\lambda_H}{\lambda_W}\frac{K}{K-1}}\left(\sum_{k=1}^K n_k x_k\right) + \sum_{k=1}^K \frac{n_k}{N} c_{2,k} + \lambda_H \sum_{k=1}^K n_k x_k \\ &= -\sum_{k=1}^K \frac{1}{N}\sqrt{\frac{\lambda_H}{\lambda_W}\frac{K}{K-1}}\frac{n_k\sqrt{n_k}x_k}{1+t_k} + \sum_{k=1}^K \frac{n_k}{N}\left(\frac{1}{1+t_k}\log((K-1)(1+t_k)) + \frac{t_k}{1+t_k}\log\left(\frac{1+t_k}{t_k}\right)\right) + \lambda_H\sum_{k=1}^K n_k x_k.\end{aligned}$$

Due to the separation of the x_k's, we can minimize them individually. Consider the following function, for any $k \in [K]$:

$$g(x) = -\frac{1}{N}\sqrt{\frac{\lambda_H}{\lambda_W}\frac{K}{K-1}}\frac{n_k\sqrt{n_k}x}{1+t} + \frac{n_k}{N}\left(\frac{1}{1+t}\log((K-1)(1+t)) + \frac{t}{1+t}\log\left(\frac{1+t}{t}\right)\right) + \lambda_H n_k x, \quad x \geq 0 \tag{33}$$

where $t = \frac{1}{K-1}\exp\left(\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}x\right)$.

We note that:

$$\begin{aligned}&\frac{1}{1+t}\log((K-1)(1+t)) + \frac{t}{1+t}\log\left(\frac{1+t}{t}\right) \\ =&\frac{1}{1+t}\log((K-1)(1+t)) - \frac{1}{1+t}\log\left(\frac{1+t}{t}\right) + \log\left(\frac{1+t}{t}\right) \\ =&\frac{1}{1+t}\log((K-1)t) + \log\left(\frac{1+t}{t}\right) \\ =&\frac{\sqrt{\frac{\lambda_H}{\lambda_W}\frac{K}{K-1}n_k}x}{1+t} + \log\left(\frac{1+t}{t}\right).\end{aligned}$$

Hence:

$$g(x) = \frac{n_k}{N}\log\left(1+(K-1)\exp\left(-\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}x\right)\right) + \lambda_H n_k x$$

$$g'(x) = -\frac{n_k}{N}\frac{\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}}{1+\frac{1}{K-1}\exp\left(\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}x\right)} + \lambda_H n_k \tag{34}$$

$$g'(x) = 0 \Rightarrow c_1 = \frac{\sqrt{n_k}}{1+\frac{1}{K-1}\exp\left(\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}x\right)} = N\sqrt{\frac{K-1}{K}\lambda_W\lambda_H}, \tag{35}$$

$$\Rightarrow x^* = \sqrt{\frac{K-1}{K}\frac{\lambda_W}{\lambda_H}\frac{1}{n_k}}\left(\log(K-1)+\log\left(\frac{\sqrt{n_k}}{N\sqrt{\frac{K-1}{K}\lambda_W\lambda_H}}-1\right)\right). \tag{36}$$

Since $x = \|\mathbf{h}\|^2 \geq 0$, we have that $x^* > 0$ if $(K-1)\left(\frac{\sqrt{n_k}}{N\sqrt{\frac{K-1}{K}\lambda_W\lambda_H}}-1\right) > 1$ or equivalently, $\frac{N}{\sqrt{n_k}}\sqrt{\lambda_W\lambda_H} < \sqrt{\frac{K-1}{K}}$. Otherwise, if $\frac{N}{\sqrt{n_k}}\sqrt{\lambda_W\lambda_H} \geq \sqrt{\frac{K-1}{K}}$, we have $g'(x) > 0 \quad \forall x > 0$ and thus, $x^* = 0$.

In conclusion, we have:

$$\mathcal{L}_1(\mathbf{W},\mathbf{H}) = \xi(c_1, x_1, \ldots, x_K) \geq \sum_{k=1}^{K} g(x_k^*) = \text{const}$$

For any $(\mathbf{W},\mathbf{H})$ that the equality conditions at Lemma B.2 do not hold, we have that $\mathcal{L}_1(\mathbf{W},\mathbf{H}) > \xi\left(c_1 = N\sqrt{\frac{K-1}{K}\lambda_W\lambda_H}, x_1, \ldots, x_K\right)$ and:

$$\begin{aligned}
&\xi\left(c_1 = N\sqrt{\frac{K-1}{K}\lambda_W\lambda_H}, x_1, \ldots, x_K\right)\\
=&\frac{-c_1}{N}\sqrt{\frac{\lambda_H}{\lambda_W}\frac{K}{K-1}}\left(\sum_{k=1}^{K} n_k x_k\right) + \sum_{k=1}^{K}\frac{n_k}{N}c_{2,k} + \lambda_H\left(\sum_{k=1}^{K} n_k x_k\right)\\
=&\sum_{k=1}^{K}\frac{n_k}{N}c_{2,k}\\
=&\sum_{k=1}^{K}\frac{n_k}{N}\left(\frac{1}{1+t_k}\log((K-1)t_k) + \log\left(\frac{1+t_k}{t_k}\right)\right) \qquad (\text{with } t_k = \sqrt{n_k}/c_1 - 1)\\
=&\sum_{k=1}^{K} g(x_k^*),
\end{aligned}$$

hence, $(\mathbf{W},\mathbf{H})$ is not optimal.

Step 5: We finish the proof since $\mathcal{L}_0(\mathbf{W},\mathbf{H}) \geq \mathcal{L}_1(\mathbf{W},\mathbf{H}) \geq$ const and we study the equality conditions.

In conclusion, by summarizing all equality conditions, we have that any optimal $(\mathbf{W}^*,\mathbf{H}^*)$ of the original training problem

obey the following:

$$\text{i) } \forall\, k \in [K], \mathbf{h}_{k,i} = \mathbf{h}_{k,j} \quad \forall\, i \neq j$$

$$\text{ii) } \mathbf{h}_k^\top \mathbf{h}_l = 0 \quad \forall\, k \neq l$$

$$\text{iii) } \mathbf{w}_k = \sqrt{\frac{1}{K(K-1)}}\sqrt{\frac{\lambda_H}{\lambda_W}}\left(K\sqrt{n_k}\mathbf{h}_k - \sum_{m=1}^{K}\sqrt{n_m}\mathbf{h}_m\right),\ \forall k \in [K]$$

$$\text{and } \sum_{k=1}^{K}\mathbf{w}_k = \mathbf{0}$$

$$\text{iv) } \|\mathbf{h}_k\|^2 = \sqrt{\frac{K-1}{K}\frac{\lambda_W}{\lambda_H}\frac{1}{n_k}}\log\left((K-1)\left(\frac{\sqrt{n_k}}{N\sqrt{\frac{K-1}{K}\lambda_W\lambda_H}} - 1\right)\right)$$

$$\text{v) For } m \neq k, \mathbf{z}_k^{(m)} = (\mathbf{W}\mathbf{h}_k)^{(m)} = -\frac{1}{K}\log\left((K-1)\left(\frac{\sqrt{n_k}}{N\sqrt{\frac{K-1}{K}\lambda_W\lambda_H}} - 1\right)\right),$$

$$\mathbf{z}_k^{(k)} = (\mathbf{W}\mathbf{h}_k)^{(k)} = \frac{K-1}{K}\log\left((K-1)\left(\frac{\sqrt{n_k}}{N\sqrt{\frac{K-1}{K}\lambda_W\lambda_H}} - 1\right)\right)$$

We proceed to deduce the results of Proposition 4.3:

$$\begin{aligned}
\|\mathbf{w}_k\|^2 &= \frac{1}{K(K-1)}\frac{\lambda_H}{\lambda_W}\left\|(K-1)\sqrt{n_k}\mathbf{h}_k - \sum_{m\neq k}\sqrt{n_m}\mathbf{h}_m\right\|^2 \\
&= \frac{1}{K(K-1)}\frac{\lambda_H}{\lambda_W}\left((K-1)^2 n_k\|\mathbf{h}_k\|^2 + \sum_{m\neq k} n_m\|\mathbf{h}_m\|^2\right) \\
&= \frac{1}{K\sqrt{K(K-1)}}\sqrt{\frac{\lambda_H}{\lambda_W}}\left((K-1)^2\sqrt{n_k}M_k + \sum_{m\neq k}\sqrt{n_m}M_m\right), \\
\mathbf{w}_k^\top\mathbf{w}_j &= \frac{1}{K(K-1)}\frac{\lambda_H}{\lambda_W}\left\langle (K-1)\sqrt{n_k}\mathbf{h}_k - \sum_{m\neq k}\sqrt{n_m}\mathbf{h}_m, (K-1)\sqrt{n_j}\mathbf{h}_j - \sum_{m\neq j}\sqrt{n_m}\mathbf{h}_m\right\rangle \\
&= \frac{1}{K(K-1)}\frac{\lambda_H}{\lambda_W}\left(-(K-1)n_k\|\mathbf{h}_k\|^2 - (K-1)n_j\|\mathbf{h}_j\|^2 + \sum_{m\neq k,j} n_m\|\mathbf{h}_m\|^2\right) \\
&= \frac{1}{K\sqrt{K(K-1)}}\sqrt{\frac{\lambda_H}{\lambda_W}}\left[-(K-1)\sqrt{n_k}M_k - (K-1)\sqrt{n_j}M_j + \sum_{m\neq k,j}\sqrt{n_m}M_m\right], k \neq j
\end{aligned}$$

B.1. Supporting lemmas

Remark: Although our training problem is a constrained optimization problem, the constraints $\mathbf{h}_{k,i} \geq 0$ are affine functions, it is clear that strong duality holds with dual variables equal 0's. Then, the solutions of the primal problem and the optimal dual variables will satisfy KKT conditions and hence, we have $\nabla_{\mathbf{H}}\mathcal{L}_1 = \mathbf{0}$ at optimal.

Lemma B.1. *Any critical points* $(\mathbf{W}, \mathbf{H})$ *of* $\mathcal{L}_1(\mathbf{W}, \mathbf{H})$ *satisfy:*

$$\|\mathbf{W}\|_F^2 = \frac{\lambda_H}{\lambda_W}\sum_{k=1}^{K} n_k\|\mathbf{h}_k\|^2 \tag{37}$$

Proof of Lemma B.1. Recall that $\mathcal{L}_1(\mathbf{W}, \mathbf{H}) = g(\mathbf{W}\mathbf{H}) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\sum_{k=1}^{K} n_k\|\mathbf{h}_k\|^2$. We have:

$$\nabla_{\mathbf{W}}\mathcal{L}_1(\mathbf{W}, \mathbf{H}) = \nabla_{\mathbf{Z}=\mathbf{W}\mathbf{H}}\, g(\mathbf{W}\mathbf{H})\mathbf{H}^\top + \lambda_W\mathbf{W} = \mathbf{0},$$

$$\nabla_{\mathbf{H}}\mathcal{L}_1(\mathbf{W},\mathbf{H}) = \mathbf{W}^\top \nabla_{\mathbf{Z}=\mathbf{WH}}\, g(\mathbf{WH}) + \lambda_H \begin{bmatrix} n_1\mathbf{h}_1 & n_2\mathbf{h}_2 & \dots & n_K\mathbf{h}_K \end{bmatrix} = \mathbf{0}.$$

From $\mathbf{0} = \mathbf{W}^\top \nabla_{\mathbf{W}}\mathcal{L}_1(\mathbf{W},\mathbf{H}) - \nabla_{\mathbf{H}}\mathcal{L}_1(\mathbf{W},\mathbf{H})\mathbf{H}^\top$, we have:

$$\lambda_W \mathbf{W}^\top\mathbf{W} = \lambda_H \begin{bmatrix} n_1\mathbf{h}_1 & n_2\mathbf{h}_2 & \dots & n_K\mathbf{h}_K \end{bmatrix} \mathbf{H}^\top$$

Hence, by taking the trace of both sides, we have $\|\mathbf{W}\|_F^2 = \frac{\lambda_H}{\lambda_W}\sum_{k=1}^K n_k\|\mathbf{h}_k\|^2$. □

Lemma B.2. *The lower bound (32) is attained for any critical points $(\mathbf{W},\mathbf{H})$ if and only if the following hold:*

$$\mathbf{h}_k^\top \mathbf{h}_l = 0,\ \forall\, k \neq l \tag{38}$$

$$\mathbf{w}_m = \sqrt{\frac{1}{K(K-1)}}\sqrt{\frac{\lambda_H}{\lambda_W}}\left(K\sqrt{n_m}\mathbf{h}_m - \sum_{k=1}^K \sqrt{n_k}\mathbf{h}_k\right),\ \forall m \in [K] \tag{39}$$

$$t_k = \frac{1}{K-1}\exp\left(\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}\|\mathbf{h}_k\|^2\right) \tag{40}$$

$$c_1 = \frac{\sqrt{n_k}}{1+\frac{1}{K-1}\exp\left(\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}\|\mathbf{h}_k\|^2\right)} = \frac{\sqrt{n_l}}{1+\frac{1}{K-1}\exp\left(\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_l}\|\mathbf{h}_l\|^2\right)},\ \forall\, k \neq l \tag{41}$$

Proof of Lemma B.2. From the proof above, we see that if we want to achieve the lower bound (30), we need that:

$$\mathbf{h}_k^\top \mathbf{h}_l = 0 \quad \forall\, k,l \in [K], k \neq l,$$

to achieve equality for the inequality (28).

We further need the following to obey (25):

$$c_3\mathbf{w}_m = -\left(\sum_{k=1}^K \sqrt{n_k}\mathbf{h}_k - K\sqrt{n_m}\mathbf{h}_m\right),\ \forall k \in [K] \qquad \text{with } c_3 = \sqrt{\frac{\lambda_W}{\lambda_H}K(K-1)}$$

$$\Rightarrow \mathbf{w}_m = \sqrt{\frac{1}{K(K-1)}}\sqrt{\frac{\lambda_H}{\lambda_W}}\left(K\sqrt{n_m}\mathbf{h}_m - \sum_{k=1}^K \sqrt{n_k}\mathbf{h}_k\right),\ \forall k \in [K].$$

Thus:

$$\sum_{k=1}^K \mathbf{w}_k = \mathbf{0} \tag{42}$$

Next, we need the inequality (24) to hold. Equivalently, this means that the equality condition (22) need to hold. Indeed, for a given $k \in [K]$ and any $m \neq k$:

$$\begin{aligned} \mathbf{z}_k^{(m)} &= \mathbf{w}_m\mathbf{h}_k \\ &= \sqrt{\frac{1}{K(K-1)}}\sqrt{\frac{\lambda_H}{\lambda_W}}\left(K\sqrt{n_m}\mathbf{h}_m - \sum_{l=1}^K \sqrt{n_l}\mathbf{h}_l\right)\mathbf{h}_k \\ &= -\sqrt{\frac{1}{K(K-1)}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}\|\mathbf{h}_k\|^2 \\ \Rightarrow\ \mathbf{z}_k^{(m)} &= \mathbf{z}_k^{(l)} \quad \forall m,l \neq k \end{aligned} \tag{43}$$

We further have, for any $k \in [K]$:

$$\sum_{m=1}^K \mathbf{z}_k^{(m)} = \left(\sum_{m=1}^K \mathbf{w}_m\right)\mathbf{h}_k = 0,$$

$$
\begin{aligned}
K\mathbf{z}_k^{(k)} &= K\mathbf{w}_k\mathbf{h}_k = K\sqrt{\frac{1}{K(K-1)}}\sqrt{\frac{\lambda_H}{\lambda_W}}\left(K\sqrt{n_k}\mathbf{h}_k - \sum_{m=1}^{K}\sqrt{n_m}\mathbf{h}_m\right)\mathbf{h}_k \\
&= \sqrt{K(K-1)}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}\|\mathbf{h}_k\|^2 \\
\Rightarrow t_k &= \left[(K-1)\exp\left(\frac{\sum_{m=1}^{K}\mathbf{z}_k^{(m)} - K\mathbf{z}_k^{(k)}}{K-1}\right)\right]^{-1} = \frac{1}{K-1}\exp\left(\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}\|\mathbf{h}_k\|^2\right)
\end{aligned} \tag{44}
$$

Since the scalar c_1 is chosen to be the same for all $k \in [K]$, we have:

$$
c_1 = \frac{\sqrt{n_k}}{1+t_k} = \frac{\sqrt{n_k}}{1+\frac{1}{K-1}\exp\left(\sqrt{\frac{K}{K-1}}\sqrt{\frac{\lambda_H}{\lambda_W}}\sqrt{n_k}\|\mathbf{h}_k\|^2\right)}, \forall\, k \in [K] \tag{45}
$$

□

C. Comparison with SELI geometry

In this section, we make a comparison between our geometry derived in Theorem 4.1, which is the convergence geometry of the UFM Cross-entropy class-imbalance problem with ReLU features, and SELI (Thrampoulidis et al., 2022), the geometry of the UFM SVM class-imbalance problem. Our conclusion is that both the classifier and features of our geometry are different from those of SELI, for both finite ($\lambda > 0$) and vanishing regularization level ($\lambda \to 0$).

First, we have a useful result that used for subsequent analysis in the vanishing regularization scenario:

Lemma C.1. *Let $\{M_i\}_{i=1}^K$ be the constants that we have defined in Eqn. (3) in our main paper. We have:*

$$
\lim_{\lambda_W,\lambda_H\to 0}\frac{M_i}{M_j} = 1
$$

Proof. This property can be easily proved using L'Hôpital's rule. □

Let $\overline{\mathbf{H}}$ is the class-mean matrix. For the prediction matrix, we have from Theorem 4.1, point (e):

$$
\mathbf{Z} = \mathbf{W}\overline{\mathbf{H}} = \begin{bmatrix} (1-1/K)M_1 & -M_2/K & \dots & -M_K/K \\ -M_1/K & (1-1/K)M_2 & \dots & -M_K/K \\ \vdots & \vdots & \ddots & \vdots \\ -M_1/K & -M_2/K & \dots & (1-1/K)M_K \end{bmatrix}
$$

$$
\Rightarrow \lim_{\lambda_W,\lambda_H\to 0}\frac{\mathbf{Z}}{\|\mathbf{Z}\|_F} \propto \begin{bmatrix} 1-1/K & -1/K & \dots & -1/K \\ -1/K & 1-1/K & \dots & -1/K \\ \vdots & \vdots & \ddots & \vdots \\ -1/K & -1/K & \dots & 1-1/K \end{bmatrix}.
$$

Hence, (i) the prediction matrix with finite λ's is different from SELI's prediction matrix due to the multiplication M_k at each column, (ii) in limiting case where the λ's converge to 0, our prediction matrix converges to the ETF matrix. SELI structure, after grouping the identical columns of the SEL matrix (see Definition 2 in (Thrampoulidis et al., 2022)), is also an ETF. It is proven in (Thrampoulidis et al., 2022) that the matrix $\mathbf{WH}$ follows SEL matrix and since the features of the same class converge to their class-mean, we have $\mathbf{Z} = \mathbf{W}\overline{\mathbf{H}}$ follows ETF structure.

To derive the classifier and class-means Gram matrices, i.e., $\mathbf{W}\mathbf{W}^\top$ and $\overline{\mathbf{H}}^\top\overline{\mathbf{H}}$, we consider the same setting $(R, 1/2)$ as (Thrampoulidis et al., 2022) for easier comparison with results derived for SELI at page 24 and 25 in (Thrampoulidis et al., 2022). Specifically, the setting has total K classes, with $K/2$ classes are majority class with n_A samples per class, the other

$K/2$ classes are minority with n_B samples per class. The imbalance ratio $\frac{n_A}{n_B}$ is R.

For the matrix $\mathbf{W}\mathbf{W}^\top$, using the results in the Proposition 4.3, we have:

$$\mathbf{W}\mathbf{W}^\top \propto \begin{bmatrix} \sqrt{R}M_A\mathbf{I}_{K/2} - \frac{1}{K}\left(\frac{3}{2}\sqrt{R}M_A - \frac{1}{2}M_B\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top & -\frac{1}{2K}(\sqrt{R}M_A + M_B)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top \\ -\frac{1}{2K}(\sqrt{R}M_A + M_B)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top & M_B\mathbf{I}_{K/2} - \frac{1}{K}\left(\frac{3}{2}M_B - \frac{1}{2}\sqrt{R}M_A\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top \end{bmatrix}$$

$$\Rightarrow \lim_{\lambda_W,\lambda_H\to 0} \frac{\mathbf{W}\mathbf{W}^\top}{\|\mathbf{W}\mathbf{W}^\top\|_F} \propto \begin{bmatrix} \sqrt{R}\mathbf{I}_{K/2} - \frac{1}{K}\left(\frac{3}{2}\sqrt{R} - \frac{1}{2}\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top & -\frac{1}{2K}(\sqrt{R}+1)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top \\ -\frac{1}{2K}(\sqrt{R}+1)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top & \mathbf{I}_{K/2} - \frac{1}{K}\left(\frac{3}{2} - \frac{1}{2}\sqrt{R}\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top \end{bmatrix},$$

which has the similar block structure as the Gram matrix $\mathbf{W}\mathbf{W}^\top$ in (Thrampoulidis et al., 2022), but the elements are all different with SELI in both finite λ's and vanishing λ's cases (see page 24 in (Thrampoulidis et al., 2022) for the SELI closed-form $\mathbf{W}\mathbf{W}^\top$).

The "centering" considered in (Thrampoulidis et al., 2022) is equivalent to centering the class-mean matrix $\overline{\mathbf{H}}$, which is $\overline{\mathbf{H}} - \overline{\mathbf{h}}_G\mathbf{1}_K^\top$ with $\overline{\mathbf{h}}_G = \frac{1}{K}\sum_{i=1}^K \mathbf{h}_i$. Regarding the "centering" class-mean matrix, we have:

$$(\overline{\mathbf{H}} - \overline{\mathbf{h}}_G\mathbf{1}_K^\top)^\top(\overline{\mathbf{H}} - \overline{\mathbf{h}}_G\mathbf{1}_K^\top) = (\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)^\top\overline{\mathbf{H}}^\top\overline{\mathbf{H}}(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top).$$

From our Theorem 4.1, we have: $\overline{\mathbf{H}}^\top\overline{\mathbf{H}} \propto \mathrm{diag}\left(\frac{M_A}{\sqrt{n_A}}, \ldots, \frac{M_A}{\sqrt{n_A}}, \frac{M_B}{\sqrt{n_B}}, \ldots, \frac{M_B}{\sqrt{n_B}}\right)$.

Thus,

$$(\overline{\mathbf{H}} - \overline{\mathbf{h}}_G\mathbf{1}_K^\top)^\top(\overline{\mathbf{H}} - \overline{\mathbf{h}}_G\mathbf{1}_K^\top) \propto \begin{bmatrix} \frac{M_A}{\sqrt{R}}\mathbf{I}_{K/2} - \frac{1}{K}\left(\frac{3}{2}\frac{M_A}{\sqrt{R}} - \frac{1}{2}M_B\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top & -\frac{1}{2K}\left(\frac{M_A}{\sqrt{R}} + M_B\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top \\ -\frac{1}{2K}\left(\frac{M_A}{\sqrt{R}} + M_B\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top & M_B\mathbf{I}_{K/2} - \frac{1}{K}\left(\frac{3}{2}M_B - \frac{1}{2}\frac{M_A}{\sqrt{R}}\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top \end{bmatrix}$$

$$\Rightarrow \lim_{\lambda_W,\lambda_H\to 0} \frac{(\overline{\mathbf{H}} - \overline{\mathbf{h}}_G\mathbf{1}_K^\top)^\top(\overline{\mathbf{H}} - \overline{\mathbf{h}}_G\mathbf{1}_K^\top)}{\|(\overline{\mathbf{H}} - \overline{\mathbf{h}}_G\mathbf{1}_K^\top)^\top(\overline{\mathbf{H}} - \overline{\mathbf{h}}_G\mathbf{1}_K^\top)\|_F} \propto \begin{bmatrix} \frac{1}{\sqrt{R}}\mathbf{I}_{K/2} - \frac{1}{K}\left(\frac{3}{2\sqrt{R}} - \frac{1}{2}\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top & -\frac{1}{2K}\left(\frac{1}{\sqrt{R}} + 1\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top \\ -\frac{1}{2K}\left(\frac{1}{\sqrt{R}} + 1\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top & \mathbf{I}_{K/2} - \frac{1}{K}\left(\frac{3}{2} - \frac{1}{2\sqrt{R}}\right)\mathbf{1}_{K/2}\mathbf{1}_{K/2}^\top \end{bmatrix},$$

which again has the similar block structure as the matrix $\mathbf{H}^\top\mathbf{H}$ in (Thrampoulidis et al., 2022), but the elements are all different with SELI in both finite λ's and vanishing λ's cases (see page 25 of (Thrampoulidis et al., 2022) for SELI closed-form $\mathbf{H}^\top\mathbf{H}$).