Neural Collapse in Deep Linear Networks: From Balanced to Imbalanced Data

Anonymous Authors[1]

Abstract

Modern deep neural networks have achieved impressive performance on tasks from image classification to natural language processing. Surprisingly, these complex systems with massive amounts of parameters exhibit the same structural properties in their last-layer features and classifiers across canonical datasets when training until convergence. In particular, it has been observed that the last-layer features collapse to their class-means, and those class-means are the vertices of a simplex Equiangular Tight Frame (ETF). This phenomenon is known as Neural Collapse ($\mathcal{NC}$). Recent papers have theoretically shown that $\mathcal{NC}$ emerges in the global minimizers of training problems with the simplified "unconstrained feature model". In this context, we take a step further and prove the $\mathcal{NC}$ occurrences in deep linear networks for the popular mean squared error (MSE) and cross entropy (CE) losses, showing that global solutions exhibit $\mathcal{NC}$ properties across the linear layers. Furthermore, we extend our study to imbalanced data for MSE loss and present the first geometric analysis of $\mathcal{NC}$ under bias-free setting. Our results demonstrate the convergence of the last-layer features and classifiers to a geometry consisting of orthogonal vectors, whose lengths depend on the amount of data in their corresponding classes. Finally, we empirically validate our theoretical analyses on synthetic and practical network architectures with both balanced and imbalanced scenarios.

1. Introduction

Despite the impressive performance of deep neural networks (DNNs) across areas of machine learning and artificial intelligence (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; Goodfellow et al., 2016; He et al., 2015; Huang et al., 2017; Brown et al., 2020), the highly non-convex nature of these systems, as well as their massive number of parameters, ranging from hundreds of millions to hundreds of billions, impose a significant barrier to having a concrete theoretical understanding of how they work. Additionally, a variety of optimization algorithms have been developed for training DNNs, which makes it more challenging to analyze the resulting trained networks and learned features (Ruder, 2016). In particular, the modern practice of training DNNs includes training the models far beyond *zero error* to achieve *zero loss* in the terminal phase of training (TPT) (Ma et al., 2017; Belkin et al., 2018; 2019). A mathematical understanding of this training paradigm is important for studying the generalization and expressivity properties of DNNs (Papyan et al., 2020; Han et al., 2021).

Recently, (Papyan et al., 2020) has empirically discovered an intriguing phenomenon, named Neural Collapse ($\mathcal{NC}$), which reveals a common pattern of the learned deep representations across canonical datasets and architectures in image classification tasks. (Papyan et al., 2020) defined Neural Collapse as the existence of the following four properties:

($\mathcal{NC}1$) **Variability collapse:** features of the same class converge to a unique vector, as training progresses.

($\mathcal{NC}2$) **Convergence to simplex ETF:** the optimal class-means have the same length and are equally and maximally pairwise seperated, 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.

Theoretically, it has been proven that $\mathcal{NC}$ emerges in the last layer of DNNs during TPT when the models belong to the class of "unconstrained features model" (UFM) (Mixon et al., 2020) and trained with cross-entropy (CE) loss or mean squared error (MSE) loss. With regard to classification tasks, CE is undoubtedly the most popular loss function to train neural networks. However, MSE has recently been shown to be effective for classification tasks, with comparable or even better generalization performance than CE loss (Hui & Belkin, 2020; Demirkaya et al., 2020; Zhou et al., 2022b).

Contributions: We provide a thorough analysis of the global solutions to the training deep linear network problem with MSE and CE losses under the unconstrained features model defined in Section 2.1. Moreover, we study the geometric structure of the learned features and classifiers under a more practical setting where the dataset is imbalanced among classes. Our contributions are three-fold:

1. UFM + MSE + balanced + deep linear network: We provide the *first mathematical analysis of the global solutions for deep linear networks with arbitrary depths and widths under UFM setting*, showing that the global solutions exhibit $\mathcal{NC}$ properties and how adding the bias term can affect the collapsed structure, when training the model with the MSE loss and balanced data.

2. UFM + MSE + imbalanced + plain/deep linear network: We provide the *first geometric analysis for the plain UFM*, which includes only one layer of weight after the unconstrained features, when training the model with the MSE loss and imbalanced data. Additionally, we also generalize this setting to the deep linear network one.

3. UFM + CE + balanced + deep linear network: We study deep linear networks trained with CE loss and demonstrate the existence of $\mathcal{NC}$ for any global minimizes in this setting.

Related works: In recent years, there has been a rapid increase in interest in $\mathcal{NC}$, resulting in a decent amount of works in a short period of time. Under UFM, these works studied different training problems, proving ETF and $\mathcal{NC}$ properties are exhibited by any global solutions of the loss functions. In particular, a line of works use UFM with CE training to analyze theoretical abstractions of $\mathcal{NC}$ (Zhu et al., 2021; Fang et al., 2021; Lu & Steinerberger, 2020). Other works study UFM with MSE loss (Tirer & Bruna, 2022; Zhou et al., 2022a; Ergen & Pilanci, 2020; Rangamani & Banburski-Fahey, 2022). For MSE loss, recent extensions to account for additional layers with non-linearity are studied in (Tirer & Bruna, 2022; Rangamani & Banburski-Fahey, 2022), or with batch normalization (Ergen & Pilanci, 2020). Furthermore, (Zhu et al., 2021; Zhou et al., 2022a;b) have shown the benign optimization landscape for several loss functions under the plain UFM setting, demonstrating that critical points can only be global minima or strict saddle points. Another line of work exploits the ETF structure to improve the network design by initially fixing the last-layer linear classifier as a simplex ETF and not performing any subsequent learning (Zhu et al., 2021; Yang et al., 2022).

Most recent papers study $\mathcal{NC}$ in a balanced setting, i.e., the number of training samples in every class is identical. This setting is vital for the existence of the ETF structure. To the best of our knowledge, $\mathcal{NC}$ with imbalanced data is studied in (Fang et al., 2021; Thrampoulidis et al., 2022; Yang et al., 2022; Xie et al., 2022). In particular, (Fang et al., 2021) is the first to observe that for imbalanced setting, the collapse of features within the same class is preserved, but the geometry skew away from the ETF. (Thrampoulidis et al., 2022) theoretically studies the SVM problem, whose global minima follows a more general geometry than the simplex 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. Due to space considerations, we defer a full discussion of related works to Appendix B. A comparison of our results with some existing works regarding the study of global optimality conditions is shown in Table 1 in Appendix B.

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. Moreover, we denote the best rank-k approximation of a matrix $\mathbf{A}$ as $\mathcal{P}_k(\mathbf{A})$. We also use some common matrix notations: $\mathbf{1}_n$ is the all-ones vector, $\operatorname{diag}\{a_1, \ldots, a_K\}$ is a square diagonal matrix size $K \times K$ with diagonal entries $a_1, \ldots, a_K$.

2. Problem Setup

We consider the classification task with K classes. Let n_k denote the number of training samples of class k, $\forall\, k \in [K]$ and $N := \sum_{k=1}^{K} n_k$. A typical deep neural network $\psi(\cdot) : \mathbb{R}^D \to \mathbb{R}^K$ can be expressed as follows:

$$\psi(\mathbf{x}) = \mathbf{W}\phi(\mathbf{x}) + \mathbf{b},$$

where $\phi(\cdot) : \mathbb{R}^D \to \mathbb{R}^d$ is the feature mapping, and $\mathbf{W} \in \mathbb{R}^{K \times d}$ and $\mathbf{b} \in \mathbb{R}^K$ are the last-layer linear classifiers and bias, respectively. Formally, the feature mapping $\phi(.)$ consists of a multilayer nonlinear compositional mapping, which can be written as:

$$\phi_\theta(\mathbf{x}) = \sigma(\mathbf{W}_L \ldots \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_L),$$

where $\mathbf{W}_l$ and $\mathbf{b}_l$, $l = 1, \ldots, L$, are the weight matrix and bias at layer l, respectively. Here, $\sigma(\cdot)$ is a nonlinear activation function. Let $\theta := \{\mathbf{W}_l, \mathbf{b}_l\}_{l=1}^{L}$ be the set of parameters in the feature mapping and $\Theta := \{\mathbf{W}, \mathbf{b}, \theta\}$ be the set of all network's parameters. We solve the following optimization problem to find the optimal values for Θ:

$$\min_{\Theta} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathcal{L}(\psi(\mathbf{x}_{k,i}), \mathbf{y}_k) + \frac{\lambda}{2}\|\Theta\|_F^2, \tag{1}$$

where $\mathbf{x}_{k,i} \in \mathbb{R}^D$ is the i-th training sample in the k-th class, and $\mathbf{y}_k \in \mathbb{R}^K$ denotes its corresponding label, which is a one-hot vector whose k-th entry is 1 and other entries are 0. Also, $\lambda > 0$ is the regularization hyperparameter that control the impact of the weight decay penalty, 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 $\mathbf{y}_k$.

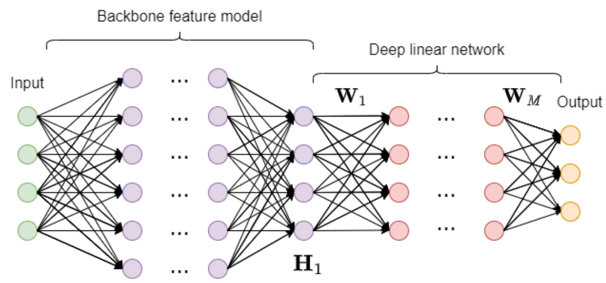
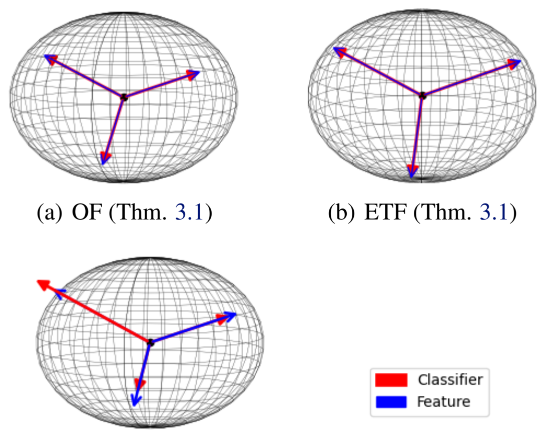


Figure 1. Illustration of UFM, followed by linear layers.



(a) OF (Thm. 3.1)

(b) ETF (Thm. 3.1)

(c) GOF (Thm. 4.1)

Figure 2. Visualization of geometries of Frobenius-normalized classifiers and features with $K = 3$ classes. For imbalanced example, the number of samples for each class is 30, 10, and 5.

2.1. Formulation under Unconstrained Features Model

Following recent studies of the $\mathcal{NC}$ phenomenon, we adopt the *unconstrained features model (UFM)* in our setting. UFM treats the last-layer features $\mathbf{h} = \phi(\mathbf{x}) \in \mathbb{R}^d$ as free optimization variables. 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). Using the UFM, we consider the following slight variant of (1):

$$\min_{\mathbf{W},\mathbf{H},\mathbf{b}} f(\mathbf{W},\mathbf{H},\mathbf{b}) := \frac{1}{2N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathcal{L}(\mathbf{W}\mathbf{h}_{k,i}+\mathbf{b},\mathbf{y}_k) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2 + \frac{\lambda_b}{2}\|\mathbf{b}\|_2^2, \quad (2)$$

where $\mathbf{h}_{k,i}$ is the feature of the i-th training sample in the k-th class. We let $\mathbf{H} := [\mathbf{h}_{1,1},\ldots,\mathbf{h}_{1,n_1},\mathbf{h}_{2,1},\ldots,\mathbf{h}_{K,n_K}] \in \mathbb{R}^{d\times N}$ be the matrix of unconstrained features. The feature class-means and global-mean are computed as $\mathbf{h}_k := n_k^{-1}\sum_{i=1}^{n_k}\mathbf{h}_{k,i}$ for $k = 1,\ldots,K$ and $\mathbf{h}_\mathbf{G} := N^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathbf{h}_{k,i}$, respectively. In this paper, we also denote $\mathbf{H}$ by $\mathbf{H_1}$ and use these notations interchangeably.

Extending UFM to the setting with M linear layers: $\mathcal{NC}$ phenomenon has been studied extensively for different loss functions under UFM but with only 1 to 2 layers of weights. In this work, we study $\mathcal{NC}$ under UFM in its significantly more general form with $M \geq 2$ linear layers by generalizing (2) to deep linear networks with arbitrary depths and widths (see Fig. 1 for an illustration). We consider the following generalization of (2) in the M-linear-layer setting:

$$\min_{\substack{\mathbf{W}_M,\ldots,\mathbf{W}_1\\ \mathbf{H}_1,\mathbf{b}}} \frac{1}{2N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathcal{L}(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1\mathbf{h}_{k,i}+\mathbf{b},\mathbf{y}_k) + \frac{\lambda_{W_M}}{2}\|\mathbf{W}_M\|_F^2 + \frac{\lambda_{W_{M-1}}}{2}\|\mathbf{W}_{M-1}\|_F^2 + \ldots + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2 + \frac{\lambda_b}{2}\|\mathbf{b}\|_2^2, \quad (3)$$

where $M \geq 2$, $\lambda_{W_M},\ldots,\lambda_{W_1},\lambda_{H_1},\lambda_b > 0$ are regularization hyperparameters, and $\mathbf{W}_M \in \mathbb{R}^{K\times d_M}$, $\mathbf{W}_{M-1} \in \mathbb{R}^{d_M\times d_{M-1}},\ldots,\mathbf{W}_1 \in \mathbb{R}^{d_2\times d_1}$ with $d_M, d_{M-1},\ldots,d_1$ are arbitrary positive integers. In our setting, we do not consider the biases of intermediate hidden layers.

Imbalanced data: Without loss of generality, we assume $n_1 \geq n_2 \geq \ldots \geq n_K$. This setting is more general than those in previous works, where only two different class sizes are considered, i.e., the majority classes of n_A training samples and the minority classes of n_B samples with the imbalance ratio $R := n_A/n_B > 1$ (Fang et al., 2021; Thrampoulidis et al., 2022).

We now define the "General Orthogonal Frame" (GOF), which is the convergence geometry of the class-means and classifiers in imbalanced MSE training problem with no bias (see Section 4).

Definition 2.1 (General Orthogonal Frame)**.** A standard general orthogonal frame (GOF) is a collection of points in $\mathbb{R}^K$ specified by the columns of:

$$\mathbf{N} = \frac{1}{\sqrt{\sum_{k=1}^{K} a_k^2}}\operatorname{diag}(a_1,a_2,\ldots,a_K),\ a_i > 0\ \forall\, i \in [K].$$

We also consider the general version of GOF as a collection of points in $\mathbb{R}^d$ ($d \geq K$) specified by the columns of $\mathbf{PN}$ where $\mathbf{P} \in \mathbb{R}^{d\times K}$ is an orthonormal matrix, i.e. $\mathbf{P}^\top\mathbf{P} = \mathbf{I}_K$. In the special case where $a_1 = a_2 = \ldots = a_K$, we have $\mathbf{N}$ follows OF structure in (Tirer & Bruna, 2022), i.e., $\mathbf{N}^\top\mathbf{N} \propto \mathbf{I}_K$. Fig. 2 shows a visualization for GOF versus OF and ETF in (Papyan et al., 2020).

3. Neural Collapse in Deep Linear Networks under the UFM Setting with Balanced Data

In this section, we present our study on the global optimality conditions for the M-layer deep linear networks ($M \geq 2$), trained with the MSE loss under the balanced setting, i.e., $n_1 = n_2 = ... = n_K := n$, extending the prior results that consider only one or two hidden layers. We consider the following optimization problem for training the model:

$$\min_{\substack{\mathbf{W}_M,\dots,\mathbf{W}_1\\ \mathbf{H}_1,\mathbf{b}}} \frac{1}{2N}\|\mathbf{W}_M\mathbf{W}_{M-1}\dots\mathbf{W}_1\mathbf{H}_1 + \mathbf{b}\mathbf{1}_n^\top - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_M}}{2}\|\mathbf{W}_M\|_F^2 + \dots + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2, \tag{4}$$

where $\mathbf{Y} = \mathbf{I}_K \otimes \mathbf{1}_n^\top \in \mathbb{R}^{K\times N}$ is the one-hot vectors matrix. Note that (4) is a special case of (3) when $\lambda_{b_M} = 0$.

We further consider two different settings from (4): (i) bias-free, i.e., excluding $\mathbf{b}$, and (ii) last-layer unregularized bias, i.e., including $\mathbf{b}$. We now state the characteristics of the global solutions to these problems.

Theorem 3.1. *Let $R := \min(K, d_M, d_{M-1}, \dots, d_2, d_1)$ and $\left(\mathbf{W}_M^*, \mathbf{W}_{M-1}^*, \dots, \mathbf{W}_1^*, \mathbf{H}_1^*, \mathbf{b}^*\right)$ be any global minimizer of (4). Denoting $a := K\sqrt[M]{Kn\lambda_{W_M}\lambda_{W_{M-1}}\dots\lambda_{W_1}\lambda_{H_1}}$, then the following results hold for both (i) bias-free setting with $\mathbf{b}^*$ excluded and (ii) last-layer unregularized bias setting with $\mathbf{b}^*$ included:*

(a) If $a < \frac{(M-1)^{\frac{M-1}{M}}}{M^2}$, we have:

($\mathcal{NC}1$) $\mathbf{H}_1^* = \overline{\mathbf{H}}^* \otimes \mathbf{1}_n^\top$, where $\overline{\mathbf{H}}^* = [\mathbf{h}_1^*, \dots, \mathbf{h}_K^*] \in \mathbb{R}^{d\times K}$ and $\mathbf{b}^* = \frac{1}{K}\mathbf{1}_K$.

($\mathcal{NC}2$) $\forall j = 1, \dots, M$:

$$\mathbf{W}_M^*\mathbf{W}_M^{*\top} \propto \overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* \propto \mathbf{W}_M^*\mathbf{W}_{M-1}^*\dots\overline{\mathbf{H}}^* \propto (\mathbf{W}_M^*\mathbf{W}_{M-1}^*\dots\mathbf{W}_j^*)(\mathbf{W}_M^*\mathbf{W}_{M-1}^*\dots\mathbf{W}_j^*)^\top$$

and align to:

(i) OF structure if (4) is bias-free:

$$\begin{cases} \mathbf{I}_K & \text{if } R \geq K \\ \mathcal{P}_R(\mathbf{I}_K) & \text{if } R < K \end{cases}.$$

(ii) ETF structure if (4) has last-layer bias $\mathbf{b}$:

$$\begin{cases} \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top & \text{if } R \geq K-1 \\ \mathcal{P}_R\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right) & \text{if } R < K-1 \end{cases}.$$

($\mathcal{NC}3$) $\forall j = 1, \dots, M$:

$$\mathbf{W}_M^*\mathbf{W}_{M-1}^*\dots\mathbf{W}_1^* \propto \overline{\mathbf{H}}^{*\top},$$
$$\mathbf{W}_M^*\mathbf{W}_{M-1}^*\dots\mathbf{W}_j^* \propto (\mathbf{W}_{j-1}^*\dots\mathbf{W}_1^*\overline{\mathbf{H}}^*)^\top.$$

(b) If $a > \frac{(M-1)^{\frac{M-1}{M}}}{M^2}$, (4) only has trivial global minima $(\mathbf{W}_M^*, \mathbf{W}_{M-1}^*, \dots, \mathbf{W}_1^*, \mathbf{H}_1^*, \mathbf{b}^*) = (\mathbf{0}, \mathbf{0}, \dots, \mathbf{0}, \mathbf{0}, \frac{1}{K}\mathbf{1}_K)$.

(c) If $a = \frac{(M-1)^{\frac{M-1}{M}}}{M^2}$, (4) has trivial global solution $(\mathbf{W}_M^*, \dots, \mathbf{W}_1^*, \mathbf{H}_1^*, \mathbf{b}^*) = (\mathbf{0}, .., \mathbf{0}, \mathbf{0}, \frac{1}{K}\mathbf{1}_K)$ and nontrivial global solutions that have the same ($\mathcal{NC}1$) and ($\mathcal{NC}3$) properties as case (a).

For ($\mathcal{NC}2$) property, for $j = 1, \dots, M$, we have:

$$\mathbf{W}_M^*\mathbf{W}_M^{*\top} \propto \overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* \propto \mathbf{W}_M^*\mathbf{W}_{M-1}^*\dots\overline{\mathbf{H}}^* \propto (\mathbf{W}_M^*\mathbf{W}_{M-1}^*\dots\mathbf{W}_j^*)(\mathbf{W}_M^*\mathbf{W}_{M-1}^*\dots\mathbf{W}_j^*)^\top$$

and align to:

$$\begin{cases} \mathcal{P}_r\left(\mathbf{I}_K\right) & \text{if (4) is bias-free} \\ \mathcal{P}_r\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right) & \text{if (4) has last-layer bias} \end{cases},$$

with r is the number of positive singular value of $\overline{\mathbf{H}}^*$.

Our proofs (in Appendix D) first characterize critical points of the loss function, showing that the weight matrices of the network have the same set of singular values, up to a factor depending on the weight decay. Then, we use the singular value decomposition on these weight matrices to transform the loss function into a function of singular values of $\mathbf{W}_1$ and singular vectors of $\mathbf{W}_M$. Due to the separation of the singular values/vectors in the expression of the loss function, we can optimize each one individually. This method shares some similarities with the proof for bias-free case in (Tirer & Bruna, 2022) where they transform a lower bound of the loss function into a function of singular values. Furthermore, the threshold $(M-1)^{\frac{M-1}{M}}/M^2$ of the constant a is derived from the minimizer of the function $g(x) = 1/(x^M+1) + bx$ for $x \geq 0$. For instance, if $b > (M-1)^{\frac{M-1}{M}}/M$, $g(x)$ is minimized at $x = 0$ and the optimal singular values will be 0's, leading to the stated solution.

The main difficulties and novelties of our proofs for deep linear networks are: i) we observe that the product of many matrices can be simplified by using SVD with identical orthonormal bases between consecutive weight matrices (see Lemma D.4) and, thus, only the singular values of $\mathbf{W}_1$ and left singular vectors of $\mathbf{W}_M$ remain in the loss function, ii) optimal singular values are related to the minimizer of the function $g(x) = 1/(x^M+1) + bx$ (see Appendix D.2.1), and iii) we study the properties of optimal singular vectors to derive the geometries of the global solutions.

Theorem 3.1 implies the following interesting results:

- **Features collapse**: For each $k \in [K]$, with class-means matrix $\overline{\mathbf{H}}^* = [\mathbf{h}_1^*, \dots, \mathbf{h}_K^*] \in \mathbb{R}^{d\times K}$, we have $\mathbf{H}_1^* = \overline{\mathbf{H}}^* \otimes \mathbf{1}_n^\top$, implying the collapse of features within the same class to their class-mean.

- **Convergence to OF/Simplex ETF:** The class-means matrix, the last-layer linear classifiers, or the product of consecutive weight matrices converge to OF in the case of bias-free and simplex ETF in the case of having last-layer bias. This result is consistent with the two and three-layer cases in (Tirer & Bruna, 2022; Zhou et al., 2022a).
- **Convergence to self-duality:** If we separate the product $\mathbf{W}_M^* \ldots \mathbf{W}_1^* \overline{\mathbf{H}}^*$ (once) into any two components, they will be perfectly aligned to each other up to rescaling. This generalizes from the previous results which demonstrate that the last-layer linear classifiers are perfectly matched with the class-means after rescaling.

Remark 3.2. The convergence of the class-means matrix to OF/Simplex ETF happens when $d_m \geq K$ (or $K-1$) $\forall m \in [M]$, which often holds in practice (Krizhevsky et al., 2012; He et al., 2015). Otherwise, they converge to the best rank-R approximation of $\mathbf{I}_K$ or $\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1_K}^\top$, where the class-means neither have the equinorm nor the maximally pairwise separation properties. This result is consistent with the two-layer case observed in (Zhou et al., 2022a).

Remark 3.3. From the proofs, we can show that under the condition $d_m \geq K, \ \forall m \in [M]$, the optimal value of the loss function is strictly smaller than when this condition does not hold. Our result is aligned with (Zhu et al., 2018), where they empirically observe that a larger network (i.e., larger width) tends to exhibit severe $\mathcal{NC}$ and have smaller training errors.

Remark 3.4. We study deep linear networks under UFM and balanced data for CE loss in Appendix A. The result demonstrates $\mathcal{NC}$ properties of every global solutions, whose the matrices product $\mathbf{W}_M \times \mathbf{W}_{M-1} \times \ldots \times \mathbf{W}_1$ and $\mathbf{H}_1$ converge to the ETF structure when training progresses.

4. Neural Collapse in Deep Linear Networks under the UFM Setting with MSE Loss and Imbalanced Data

The majority of theoretical results for NC only consider the balanced data setting, i.e., the same number of training samples for each class. This assumption plays a vital role in the existence of the well-structured ETF geometry. In this section, we instead consider the imbalanced data setting and derive the first geometry analysis under this setting for MSE loss. Furthermore, we extend our study from the plain UFM setting, which includes only one layer of weight after the unconstrained features, to the deep linear network one.

4.1. Plain UFM Setting with No Bias

The bias-free plain UFM with MSE loss is given by:

$$\min_{\mathbf{W},\mathbf{H}} \frac{1}{2N}\|\mathbf{W}\mathbf{H}-\mathbf{Y}\|_F^2 + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2, \quad (5)$$

where $\mathbf{W} \in \mathbb{R}^{K\times d}$, $\mathbf{H} \in \mathbb{R}^{d\times N}$, and $\mathbf{Y} \in \mathbb{R}^{K\times N}$ is the one-hot vectors matrix consisting n_k one-hot vectors for each class $k, \forall\, k \in [K]$. We now state the $\mathcal{NC}$ properties of the global solutions of (5) under the imbalanced data setting when the feature dimension d is at least the number of classes K.

Theorem 4.1. *Let $d \geq K$ and $(\mathbf{W}^*, \mathbf{H}^*)$ be any global minimizer of problem (5). Then, we have:*

$(\mathcal{NC}1)$ $\quad \mathbf{H}^* = \overline{\mathbf{H}}^*\mathbf{Y} \Leftrightarrow \mathbf{h}_{k,i}^* = \mathbf{h}_k^* \ \forall\, k \in [K], i \in [n_k]$, *where* $\overline{\mathbf{H}}^* = [\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*] \in \mathbb{R}^{d\times K}$.

$(\mathcal{NC}2)$ *Let $a := N^2\lambda_W\lambda_H$, we have:*

$$\mathbf{W}^*\mathbf{W}^{*\top} = \operatorname{diag}\left\{s_k^2\right\}_{k=1}^K,$$
$$\overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* = \operatorname{diag}\left\{\frac{s_k^2}{(s_k^2+N\lambda_H)^2}\right\}_{k=1}^K,$$

$$\mathbf{W}^*\mathbf{H}^* = \operatorname{diag}\left\{\frac{s_k^2}{s_k^2+N\lambda_H}\right\}_{k=1}^K \mathbf{Y} = \begin{bmatrix} \frac{s_1^2}{s_1^2+N\lambda_H}\mathbf{1}_{n_1}^\top & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \frac{s_K^2}{s_K^2+N\lambda_H}\mathbf{1}_{n_K}^\top \end{bmatrix}.$$

where:

- *If $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_K} \leq 1$:*

$$s_k = \sqrt{\sqrt{\frac{n_k\lambda_H}{\lambda_W}} - N\lambda_H} \quad \forall\, k \in [K]$$

- *If there exists a $j \in [K-1]$ s.t. $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_j} \leq 1 < \frac{a}{n_{j+1}} \leq \ldots \leq \frac{a}{n_K}$:*

$$s_k = \begin{cases} \sqrt{\sqrt{\frac{n_k\lambda_H}{\lambda_W}} - N\lambda_H} & \forall\, k \leq j \\ 0 & \forall\, k > j \end{cases}.$$

- *If $1 < \frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_K}$:*

$$(s_1, s_2, \ldots, s_K) = (0, 0, \ldots, 0),$$

and $(\mathbf{W}^*, \mathbf{H}^*) = (\mathbf{0}, \mathbf{0})$ in this case.

For any k such that $s_k = 0$, we have:

$$\mathbf{w}_k^* = \mathbf{h}_k^* = \mathbf{0}.$$

$(\mathcal{NC}3)$ $\quad \mathbf{w}_k^* = \sqrt{\frac{n_k\lambda_H}{\lambda_W}}\mathbf{h}_k^* \quad \forall\, k \in [K]$.

The detailed proofs are provided in the Appendix E. We use the same approach as the proofs of Theorem 3.1 to prove this result, with challenge arises in the process of lower bounding the loss function w.r.t. the singular vectors of $\mathbf{W}$. Interestingly, the left singular matrix of $\mathbf{W}^*$ consists multiple orthogonal blocks on its diagonal, with each block corresponds with a group of classes having the same number of training samples. This property creates the orthogonality of $(\mathcal{NC}2)$ geometries.

Theorem 4.1 implies the following interesting results:

- **Features collapse:** The features in the same class also converge to their class-mean, similar as balanced case.
- **Convergence to GOF:** When the condition $N^2\lambda_W\lambda_H/n_K < 1$ is hold, the class-means matrix and the last-layer classifiers converge to GOF (see Definition 2.1). This geometry includes orthogonal vectors, but their length depends on the number of training samples in the class. The above condition implies that the imbalance and the regularization level should not be too heavy to avoid trivial solutions that may harm the model performances. We will discuss more about this phenomenon in Section 4.2.
- **Alignment between linear classifiers and last-layer features:** The last-layer linear classifier is aligned with the class-mean of the same class, but with a different ratio across classes. These ratios are proportional to the square root of the number of training samples, and thus different compared to the balanced case where $\mathbf{W}^*/\|\mathbf{W}^*\|_F = \overline{\mathbf{H}}^{*\top}/\|\overline{\mathbf{H}}^{*\top}\|_F$.

Remark 4.2. We study the case $d < K$ in Theorem E.2. In this case, while $(\mathcal{NC}1)$ and $(\mathcal{NC}3)$ are exactly similar as the case $d \geq K$, the $(\mathcal{NC}2)$ geometries are different if $a/n_d < 1$ and $n_d = n_{d+1}$, where a square block on the diagonal is replaced by its low-rank approximation. This square block corresponds to classes with the number of training samples equal n_d. Also, we have $\mathbf{w}_k^* = \mathbf{h}_k^* = \mathbf{0}$ for any class k with the amount of data is less than n_d.

4.2. GOF Structure with Different Imbalance Levels and Minority Collapse

Given the exact closed forms of the singular values of $\mathbf{W}^*$ stated in Theorem 4.1, we derive the norm ratios between the classifiers and between features across classes as follows:

Lemma 4.3. *Suppose $(\mathbf{W}^*, \mathbf{H}^*)$ is a global minimizer of problem (5) such that $d \geq K$ and $N^2\lambda_W\lambda_H/n_K < 1$, so that all the s_k's are positive. The following results hold:*

$$\frac{\|\mathbf{w}_i^*\|^2}{\|\mathbf{w}_j^*\|^2} = \frac{\sqrt{\frac{n_i\lambda_H}{\lambda_W}} - N\lambda_H}{\sqrt{\frac{n_j\lambda_H}{\lambda_W}} - N\lambda_H}, \quad \frac{\|\mathbf{h}_i^*\|^2}{\|\mathbf{h}_j^*\|^2} = \frac{n_j}{n_i}\frac{\sqrt{\frac{n_j\lambda_H}{\lambda_W}} - N\lambda_H}{\sqrt{\frac{n_i\lambda_H}{\lambda_W}} - N\lambda_H}.$$

If $n_i \geq n_j$, we have $\|\mathbf{w}_i^*\| \geq \|\mathbf{w}_j^*\|$ and $\|\mathbf{h}_i^*\| \leq \|\mathbf{h}_j^*\|$.

It has been empirically observed that the classifiers of the majority classes have greater norms (Kang et al., 2019). Our result is in agreement with this observation. Moreover, it has been shown that class imbalance impairs the model's accuracy on minority classes (Kang et al., 2019; Cao et al., 2019). Recently, (Fang et al., 2021) discover the "Minority Collapse" phenomenon. In particular, they show that there exists a finite threshold for imbalance level beyond which all the minority classifiers collapse to a single vector, resulting in the model's poor performance on these classes. *Theorem 4.1 is not only aligned with the "Minority Collapse" phenomenon, but also provides the imbalance threshold for the collapse of minority classes to vector* $\mathbf{0}$, *i.e.,* $N^2\lambda_W\lambda_H/n_K > 1$.

4.3. Bias-free Deep Linear Network under the UFM setting

We now generalize (5) to bias-free deep linear networks with $M \geq 2$ and arbitrary widths. We study the following optimization problem with imbalanced data:

$$\min_{\mathbf{W}_M,\mathbf{W}_{M-1},\ldots,\mathbf{W}_1,\mathbf{H}_1} \frac{1}{2N}\|\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1\mathbf{H}_1 - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_M}}{2}\|\mathbf{W}_M\|_F^2 + \ldots + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2, \tag{6}$$

where the target matrix $\mathbf{Y}$ is the one-hot vectors matrix defined in (5). We now state the $\mathcal{NC}$ properties of the global solutions of (6) when the dimensions of the hidden layers are at least the number of classes K.

Theorem 4.4. *Let $d_m \geq K$, $\forall\, m \in [M]$, and $(\mathbf{W}_M^*, \mathbf{W}_{M-1}^*, \ldots, \mathbf{W}_1^*, \mathbf{H}_1^*)$ be any global minimizer of problem (6). We have the following results:*

$(\mathcal{NC}1)$ $\quad \mathbf{H}_1^* = \overline{\mathbf{H}}^*\mathbf{Y} \Leftrightarrow \mathbf{h}_{k,i}^* = \mathbf{h}_k^* \;\forall\, k \in [K], i \in [n_k]$, *where* $\overline{\mathbf{H}}^* = [\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*] \in \mathbb{R}^{d_1\times K}$.

$(\mathcal{NC}2)$ Let $c := \frac{\lambda_{W_1}^{M-1}}{\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_2}}$, $a := N\sqrt[M]{N\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_1}\lambda_{H_1}}$ *and* $\forall k \in [K]$, *x_k^* is the largest positive solution of the equation* $\frac{a}{n_k} - \frac{x^{M-1}}{(x^M+1)^2} = 0$, *we have the following:*

$$\mathbf{W}_M^*\mathbf{W}_M^{*\top} = \frac{\lambda_{W_1}}{\lambda_{W_M}}\operatorname{diag}\left\{s_k^2\right\}_{k=1}^K,$$
$$(\mathbf{W}_M^*\ldots\mathbf{W}_1^*)(\mathbf{W}_M^*\ldots\mathbf{W}_1^*)^\top = \operatorname{diag}\left\{cs_k^{2M}\right\}_{k=1}^K,$$
$$\overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* = \operatorname{diag}\left\{\frac{cs_k^{2M}}{(cs_k^{2M}+N\lambda_{H_1})^2}\right\}_{k=1}^K,$$
$$\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_1^*\mathbf{H}_1^* = \left\{\frac{cs_k^{2M}}{cs_k^{2M}+N\lambda_{H_1}}\right\}_{k=1}^K\mathbf{Y},$$

($\mathcal{NC}3$) We have, $\forall\, k \in [K]$:

$$(\mathbf{W}_M^* \mathbf{W}_{M-1}^* \dots \mathbf{W}_1^*)_k = (c s_k^{2M} + N\lambda_{H_1}) \mathbf{h}_k^*,$$

where:

- *If $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \dots \leq \frac{a}{n_K} < \frac{(M-1)^{\frac{M-1}{M}}}{M^2}$, we have:*

$$s_k = \sqrt[2M]{\frac{N\lambda_{H_1} x_k^{*M}}{c}} \quad \forall\, k \in [K].$$

- *If there exists a $j \in [K-1]$ s.t. $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \dots \leq \frac{a}{n_j} < \frac{(M-1)^{\frac{M-1}{M}}}{M^2} < \frac{a}{n_{j+1}} \leq \dots \leq \frac{a}{n_K}$, we have:*

$$s_k = \begin{cases} \sqrt[2M]{\frac{N\lambda_{H_1} x_k^{*M}}{c}} & \forall\, k \leq j \\ 0 & \forall\, k > j \end{cases}.$$

For any k such that $s_k = 0$, we have:

$$(\mathbf{W}_M^*)_k = \mathbf{h}_k^* = \mathbf{0}.$$

- *If $\frac{(M-1)^{\frac{M-1}{M}}}{M^2} < \frac{a}{n_1} \leq \frac{a}{n_2} \leq \dots \leq \frac{a}{n_K}$, we have:*

$$(s_1, s_2, \dots, s_K) = (0, 0, \dots, 0),$$

and $(\mathbf{W}_M^*, \dots, \mathbf{W}_1^*, \mathbf{H}_1^*) = (\mathbf{0}, \dots, \mathbf{0}, \mathbf{0})$ in this case.

The detailed proofs of Theorem 4.4 and the remaining case where there are some $\frac{a}{n_k}$'s equal to $\frac{(M-1)^{\frac{M-1}{M}}}{M^2}$ are provided in Appendix F.

Remark 4.5. The equation that solves for the optimal singular value, $\frac{a}{n} - \frac{x^{M-1}}{(x^M+1)^2} = 0$, has exactly two positive solutions when $a < (M-1)^{\frac{M-1}{M}}/M^2$ (see Section D.2.1). Solving this equation leads to cumbersome solutions of a high-degree polynomial. Even without the exact closed-form formula for the solution, the ($\mathcal{NC}2$) geometries can still be easily computed by numerical methods.

Remark 4.6. We study the case $R := \min(d_M, \dots, d_1, K) < K$ in Theorem F.2. In this case, while ($\mathcal{NC}1$) and ($\mathcal{NC}3$) are exactly similar as the case $R = K$ in Theorem 4.4, the ($\mathcal{NC}2$) geometries are different if $a/n_R \leq 1$ and $n_R = n_{R+1}$, where a square block on the diagonal is replaced by its low-rank approximation. This square block corresponds to classes with the number of training samples equal n_R. Also, we have $(\mathbf{W}_M)_k^* = \mathbf{h}_k^* = \mathbf{0}$ for any class k with the amount of data is less than n_R.

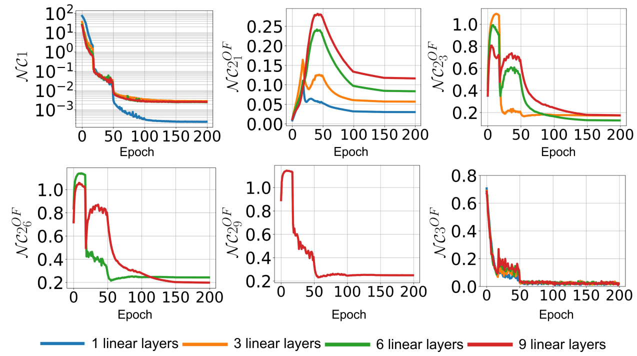


Figure 3. Illustration of $\mathcal{NC}$ with 6-layer MLP backbone on CIFAR10 for MSE loss, balanced data and bias-free setting.

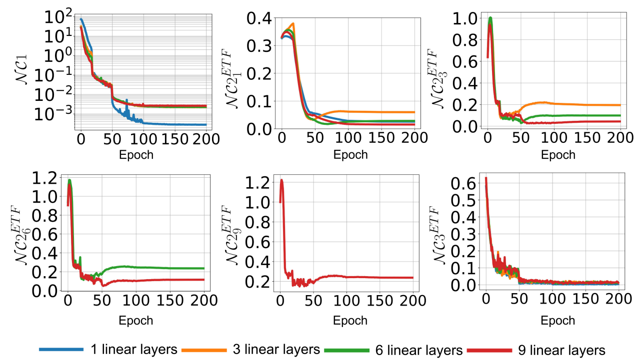


Figure 4. Same setup as Fig. 3 but having last-layer bias.

5. Experimental Results

In this section, we empirically verify our theoretical results in multiple settings for both balanced and imbalanced data settings. In particular, we observe the evolution of NC properties in the training of deep linear networks with a prior backbone feature extractor to create the "unconstrained" features (see Fig. 1 for a sample visualization). The experiments are performed on CIFAR10 (Krizhevsky, 2009) dataset for the image classification task. Moreover, we also perform direct optimization experiments, which follows the setting in (3) to guarantee our theoretical analysis.

The hyperparameters of the optimizers are tuned to reach the global optimizer in all experiments. The definitions of the $\mathcal{NC}$ metrics, hyperparameters details, and additional numerical results can be found in Appendix C.

5.1. Balanced Data

Under the balanced data setting, we alternatively substitute between multilayer perceptron (MLP), ResNet18 (He et al., 2016) and VGG16 (Simonyan & Zisserman, 2014) in place of the backbone feature extractor. For all experiments with MLP backbone model, we perform the regularization on the "unconstrained" features $\mathbf{H}_1$ and on subsequent weight layers to replicate the UFM setting in (3). For deep learn-

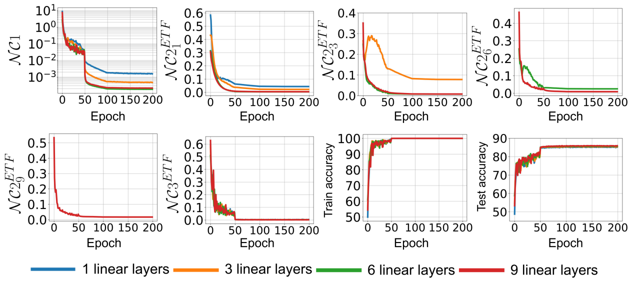


Figure 5. Training results with ResNet18 backbone on CIFAR10 for MSE loss, balanced data and last-layer bias setting.

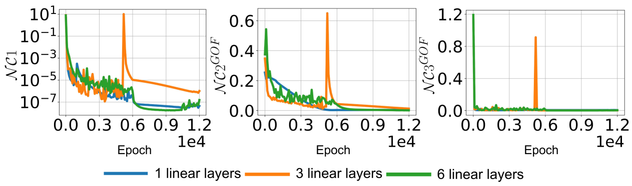


Figure 6. Illustration of $\mathcal{NC}$ with 6-layer MLP backbone on an imbalanced subset of CIFAR10 for MSE loss and bias-free setting.

ing experiments with ResNet18 and VGG16 backbone, we enforce the weight decay on all parameters of the network, which aligns to the typical training protocol.

Multilayer perceptron experiment: We use a 6-layer MLP model with ReLU activation as the backbone feature extractor in this experiment. For deep linear layers, we cover all depth-width combinations with depth $\in \{1, 3, 6, 9\}$ and width $\in \{512, 1024, 2048\}$. We run both bias-free and last-layer bias cases to demonstrate the convergence to OF and ETF geometry, with the models trained by Adam optimizer (Kingma & Ba, 2014) for 200 epochs. For a concrete illustration, the results of width-1024 MLP backbone and linear layers for MSE loss are shown in Fig. 3 and Fig. 4. We consistently observe the convergence of $\mathcal{NC}$ metrics to small values as training progresses for various depths of the linear networks. Additional results with MLP backbone for other widths and for CE loss can be found in Appendix C.1.

Deep learning experiment: We use ResNet18 and VGG16 as the deep learning backbone for extracting $\mathbf{H}_1$ in this experiment. The depths of the deep linear network are selected from the set $\{1, 3, 6, 9\}$ and the widths are chosen to equal the last-layer dimension of the backbone model (i.e., 512). The models are trained with the MSE loss without data augmentation for 200 epochs using stochastic gradient descent (SGD). As shown in Fig. 5 above and Fig.7 in the Appendix C.1.2, $\mathcal{NC}$ properties are obtained for widely used architectures in deep learning contexts. Furthermore, the results empirically confirm the occurrences of $\mathcal{NC}$ across deep linear classifiers described in Theorem 3.1.

Direct optimization experiment: To exactly replicate the problem (3), $\mathbf{W}_M, \ldots, \mathbf{W}_1$ and $\mathbf{H}_1$ are initialized with standard normal distribution scaled by 0.1 and optimized with gradient descent with step-size 0.1 for MSE loss. In this experiment, we set $K = 4, n = 100, d_M = d_{M-1} = \ldots = d_1 = 64$ and all λ's are set to be 5×10^{-4}. We cover multiple depth settings with M chosen from the set $\{1, 3, 6, 9\}$. Fig. 8 and Fig. 9 in Appendix C.1.2 shows the convergence to 0 of $\mathcal{NC}$ metrics for bias-free and last-layer bias settings, respectively. The convergence errors are less than 1e-3 at the final iteration, which corroborates Theorem 3.1.

5.2. Imbalanced Data

For imbalanced data setting, we perform two experiments: CIFAR10 image classification with an MLP backbone and direct optimization with a similar setup as in Section 5.1.

Multilayer perceptron experiment: In this experiment, we use a 6-layer MLP network with ReLU activation as the backbone model with removed batch normalization. We choose a random subset of CIFAR10 dataset with number of training samples of each class chosen from the list $\{500, 500, 400, 400, 300, 300, 200, 200, 100, 100\}$. The network is trained with batch gradient descent for 12000 epochs. Both the feature extraction model and deep linear model share the hidden width $d = 2048$. This experiment is performed with multiple linear model depths $M = 1, 3, 6$ and the results are shown in Fig. 6. The converge of $\mathcal{NC}$ metrics to 0 (errors are at most 5e-2 at the final epoch) strongly validates Theorem 4.1 and 4.4 with the convergence to GOF structure of learned classifiers and features.

Direct optimization experiment: In this experiment, except for the imbalanced data of $K = 4$ and $n_1 = 200, n_2 = 100, n_3 = n_4 = 50$, the settings are identical to the direct optimization experiment in balanced case for MSE loss. Fig. 12 in Appendix C.2.2 corroborates Theorems 4.1 and 4.4 for various depths $M = 1, 3, 6$ and 9.

6. Concluding Remarks

In this work, we extend the global optimal analysis of the deep linear networks trained with the mean squared error (MSE) and cross entropy (CE) losses under the unconstrained features model. We prove that NC phenomenon is exhibited by the global solutions across layers. Moreover, we extend our theoretical analysis to the UFM imbalanced data settings for the MSE loss, which are much less studied in the current literature, and thoroughly analyze NC properties under this scenario. In our work, we do not include biases in the training problem under imbalanced setting. We leave the study of the collapsed structure with the presence of biases as future work. As the next natural development of our results, characterizing NC for deep networks with non-linear activations under unconstrained features model is a highly interesting direction for future research.

References

Baldi, P. and Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90014-2. URL https://www.sciencedirect.com/science/article/pii/0893608089900142.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality?, 2018. URL https://arxiv.org/abs/1806.09471.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, jul 2019. doi: 10.1073/pnas.1903070116. URL https://doi.org/10.1073%2Fpnas.1903070116.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

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.

Demirkaya, A., Chen, J., and Oymak, S. Exploring the role of loss functions in multiclass classification. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–5, 2020. doi: 10.1109/CISS48834.2020.1570627167.

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.

Goodfellow, I. J., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

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, 2015. URL https://arxiv.org/abs/1512.03385.

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.

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, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Hui, L. and Belkin, M. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks, 2020. URL https://arxiv.org/abs/2006.07322.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. 2019. doi: 10.48550/ARXIV.1910.09217. URL https://arxiv.org/abs/1910.09217.

Kawaguchi, K. Deep learning without poor local minima, 2016. URL https://arxiv.org/abs/1605.07110.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

Laurent, T. and von Brecht, J. Deep linear neural networks with arbitrary loss: All local minima are global, 2017. URL https://arxiv.org/abs/1712.01473.

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

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning, 2017. URL https://arxiv.org/abs/1712.06559.

Mixon, D. G., Parshall, H., and Pi, J. Neural collapse with unconstrained features, 2020. URL https://arxiv.org/abs/2011.11619.

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.

Ruder, S. An overview of gradient descent optimization algorithms, 2016. URL https://arxiv.org/abs/1609.04747.

Safran, I. and Shamir, O. Spurious local minima are common in two-layer relu neural networks, 2017. URL https://arxiv.org/abs/1712.08968.

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

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.

Xie, L., Yang, Y., Cai, D., and He, X. Neural collapse inspired attraction-repulsion-balanced loss for imbalanced learning, 2022. URL https://arxiv.org/abs/2204.08735.

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.

Yun, C., Sra, S., and Jadbabaie, A. Global optimality conditions for deep neural networks, 2017. URL https://arxiv.org/abs/1707.02444.

Yun, C., Sra, S., and Jadbabaie, A. Small nonlinearities in activation functions create bad local minima in neural networks, 2018. URL https://arxiv.org/abs/1802.03487.

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., Soudry, D., Eldar, Y. C., and Wakin, M. B. The global optimization geometry of shallow linear neural networks, 2018. URL https://arxiv.org/abs/1805.04938.

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 in Deep Linear Networks: From Balanced to Imbalanced Data"

Firstly, we study $\mathcal{NC}$ characteristics for cross-entropy loss function in deep linear networks in Appendix A. The delayed related works discussion are provided in Appendix B. Next, we present additional numerical results and experiments, details of training hyperparameters and describe $\mathcal{NC}$ metrics used for experiments in Appendix C. Finally, detailed proofs for Theorems 3.1, 4.1, 4.4 and A.1 are provided in Appendix D, E, F and G, respecively.

A. Neural Collapse in Deep Linear Networks under UFM Setting for CE with Balanced Data

In this section, we turn to cross-entropy loss and generalize $\mathcal{NC}$ for deep linear networks with last-layer bias under balanced setting, and a mild assumption that all the hidden layers dimension are at least $K-1$ is required. We consider the training problem (3) with CE loss as following:

$$\min_{\mathbf{W}_M,\dots,\mathbf{W}_1,\mathbf{H}_1,\mathbf{b}} \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathcal{L}_{CE}(\mathbf{W}_M\dots\mathbf{W}_1\mathbf{h}_{k,i}+\mathbf{b},\mathbf{y}_k)+\frac{\lambda_{W_M}}{2}\|\mathbf{W}_M\|_F^2+\dots+\frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2+\frac{\lambda_b}{2}\|\mathbf{b}\|_2^2, \quad (7)$$

where:

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

Theorem A.1. *Assume $d_k \geq K-1\,\forall\, k\in[M]$, then any global minimizer $(\mathbf{W}_M^*,\dots,\mathbf{W}_1^*,\mathbf{H}_1^*,\mathbf{b}^*)$ of problem (7) satisfies:*

- $(\mathcal{NC}1)+(\mathcal{NC}3)$*:*

$$\mathbf{h}_{k,i}^* = \frac{\lambda_{H_1}^M}{\lambda_{W_M}\lambda_{W_{M-1}}\dots\lambda_{W_1}}\frac{\sum_{k=1}^{K-1}s_k^2}{\sum_{k=1}^{K-1}s_k^{2M}}(\mathbf{W}_M^*\mathbf{W}_{M-1}^*\dots\mathbf{W}_1^*)_k \quad \forall k\in[K], i\in[n]$$
$$\Rightarrow \mathbf{h}_{k,i}^* = \mathbf{h}_k^* \quad \forall\, i\in[n], k\in[K],$$

  where $\{s_k\}_{k=1}^{K-1}$ are the singular values of $\mathbf{H}_1^*$.

- $(\mathcal{NC}2)$: $\mathbf{H}_1^*$ *and* $\mathbf{W}_M^*\mathbf{W}_{M-1}^*\cdots\mathbf{W}_1^*$ *will converge to a simplex ETF when training progresses:*

$$(\mathbf{W}_M^*\mathbf{W}_{M-1}^*\cdots\mathbf{W}_1^*)(\mathbf{W}_M^*\mathbf{W}_{M-1}^*\cdots\mathbf{W}_1^*)^\top = \frac{\lambda_{H_1}^M\sum_{k=1}^{K-1}s_k^{2M}}{(K-1)\lambda_{W_M}\lambda_{W_{M-1}}\dots\lambda_{W_1}}\left(\mathbf{I}_K-\frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right).$$

- *We have $\mathbf{b}^*=b^*\mathbf{1}$ where either $b^*=0$ or $\lambda_b=0$.*

The proof is delayed until Section G and some of the key techniques are extended from the proof for the plain UFM in (Zhu et al., 2021). Comparing with the plain UFM with one layer of weight only, we have for deep linear case similar results as the plain UFM case, with the $(\mathcal{NC}2)$ and $(\mathcal{NC}3)$ property now hold for the product $\mathbf{W}_M\mathbf{W}_{M-1}\dots\mathbf{W}_1$ instead of $\mathbf{W}$.

B. Related Works

In recent years, there has been a rapid increase in interest in Neural Collapse, resulting in a decent amount of papers within a short period of time. Under the unconstrained feature model, (Zhu et al., 2021; Tirer & Bruna, 2022; Zhou et al., 2022a;b; Thrampoulidis et al., 2022; Fang et al., 2021; Lu & Steinerberger, 2020; Ergen & Pilanci, 2020; Yang et al., 2022) studied different training problems, proving simplex ETF and $\mathcal{NC}$ properties are exhibited by any global solutions of the loss functions. In particular, (Zhu et al., 2021; Fang et al., 2021; Lu & Steinerberger, 2020) uses UFM with CE training to analyze theoretical abstractions of Neural Collapse. Other works study UFM with MSE loss (Tirer & Bruna, 2022; Zhou et al., 2022a; Ergen & Pilanci, 2020; Rangamani & Banburski-Fahey, 2022), and recent extensions to account for one additional layer and nonlinearity (with an extra assumption) are studied in (Tirer & Bruna, 2022) or with batch normalization (Ergen & Pilanci, 2020). The work (Rangamani & Banburski-Fahey, 2022) studies deep homogeneous networks with MSE

	Loss	Train model	Setting	Consider $d < K-1$?	Extra assumption	$\mathcal{NC}2$ geometry
(Zhu et al., 2021)	CE	Plain UFM	Balanced	No	N/a	Simplex ETF
(Fang et al., 2021)	CE	Layer-peeled	Balanced	No	N/a	Simplex ETF
(Zhou et al., 2022a)	MSE	Plain UFM	Balanced	Yes	N/a	Simplex ETF
(Tirer & Bruna, 2022)	MSE	Plain UFM, no bias	Balanced	No	N/a	OF
	MSE	Plain UFM, un-reg. bias	Balanced	No	N/a	Simplex ETF
	MSE	Extended UFM 2 linear layers, no bias	Balanced	No	N/a	OF
	MSE	Extended UFM 2 layers with ReLU, no bias	Balanced	No	Nuclear norm equality [1]	OF
(Rangamani & Banburski-Fahey, 2022)	MSE	Deep ReLU network, no bias	Balanced	No	Symmetric Quasi-interpolation [2]	Simplex ETF
(Thrampoulidis et al., 2022)	CE	UFM Support Vector Machine	Imbalanced	No	N/a	SELI
This work	MSE	Extended UFM M linear layers, no bias (Theorem 3.1)	Balanced	Yes	N/a	OF
	MSE	Extended UFM M linear layers, un-reg. last bias (Theorem 3.1)	Balanced	Yes	N/a	Simplex ETF
	MSE	Plain UFM, no bias (Theorem 4.1)	Imbalanced	Yes	N/a	GOF
	MSE	Extended UFM M linear layers, no bias (Theorem 4.4)	Imbalanced	Yes	N/a	GOF
	CE	Extended UFM M linear layers (Theorem A.1)	Balanced	No	N/a	Simplex ETF

Table 1. Selected comparision of theoretical results on global optimality conditions with $\mathcal{NC}$ occurrence.

loss and trained with stochastic gradient descent. Specifically, the critical points of gradient flow satisfying the so-called symmetric quasi-interpolation assumption are proved to exhibit $\mathcal{NC}$ properties, but the other solutions are not investigated. (Zhou et al., 2022b) recently extended the global optimal characteristics to other loss functions, such as focal loss and label smoothing. Moreover, (Zhu et al., 2021; Zhou et al., 2022a;b) provide the benign optimization landscape for different loss functions under plain UFM, demonstrating that critical points can only be global minima or strict saddle points. Another line of work, for example (Zhu et al., 2021; Yang et al., 2022), exploits the simplex ETF structure to improve the network design, such as initially fixing the last-layer linear classifier as a simplex ETF and not performing any subsequent learning.

Most recent papers study Neural Collapse under a balanced setting, i.e., the number of training samples in every class is the same. This setting is vital for the existence of the simplex ETF structure. To the best of our knowledge, Neural Collapse with imbalanced data is studied in (Fang et al., 2021; Thrampoulidis et al., 2022; Yang et al., 2022; Xie et al., 2022). In particular, (Fang et al., 2021) is the first to observe that for imbalanced setting, the collapse of features within the same class $\mathcal{NC}1$ is preserved, but the geometry skew away from ETF. They also present a phenomenon called "Minority Collapse": for large levels of imbalance, the minorities' classifiers collapse to the same vector. (Thrampoulidis et al., 2022) theoretically studies the SVM problem, whose global minima follows a more general geometry than the ETF, called "SELI". However, this work also makes clear that the unregularized and bias-free (i.e., no bias) version of CE loss only converges to KKT points of the SVM problem, which are not necessarily global minima, and thus the geometry of the global minima of CE loss is not guaranteed to be the "SELI" geometry. (Yang et al., 2022) studies the imbalanced data setting but with fixed last-layer linear classifiers initialized as a simplex ETF right at the beginning. (Xie et al., 2022) proposed a novel loss function for balancing different components of the gradients for imbalanced learning. Therefore, $\mathcal{NC}$ characterizations with imbalanced data for commonly used loss functions in deep learning regimes such as CE, MSE, etc., still remain open. A comparison of our results with some existing works regarding the study of global optimality conditions is shown in Table 1.

This work also relates to recent advances in studying the optimization landscape in deep neural network training. As pointed out in (Zhu et al., 2021), the UFM takes a top-down approach to the analysis of deep neural networks, where last-layer features are treated as free optimization variables, in contrast to the conventional bottom-up approach that studies the problem starting from the input (Baldi & Hornik, 1989; Zhu et al., 2018; Kawaguchi, 2016; Yun et al., 2017; Laurent & von Brecht, 2017; Safran & Shamir, 2017; Yun et al., 2018). These works studies the optimization landscape of two-layer linear network (Baldi & Hornik, 1989; Zhu et al., 2018), deep linear network (Kawaguchi, 2016; Yun et al., 2017; Laurent & von Brecht, 2017) and non-linear network (Safran & Shamir, 2017; Yun et al., 2018). (Zhu et al., 2021) provides an interesting perspective about the differences between this top-down and bottom-up approach, with how results stemmed from UFM can provide more insights to the network design and the generalization of deep learning while requiring fewer unrealistic assumptions than the counterpart.

[1](Tirer & Bruna, 2022) assumes the nuclear norm of $\mathbf{W}_1^*\mathbf{H}_1^*$ and $\text{ReLU}(\mathbf{W}_1^*\mathbf{H}_1^*)$ are equal for any global solution $(\mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^*)$.

[2](Rangamani & Banburski-Fahey, 2022) assumes having a classifer $f : \mathbb{R}^D \to \mathbb{R}^K$ where $[f(\mathbf{x}_{k,i})]_k = 1 - \epsilon$ and $[f(\mathbf{x}_{k,i})]_{k'} = \epsilon/(K-1)\ \forall\, k' \neq k$ for all training samples

C. Additional Experiments, Network Training and Metrics

C.1. Balanced Data

C.1.1. Metric for measuring $\mathcal{NC}$ in balanced settings

For balanced data, we use similar metrics to those presented in (Zhu et al., 2021) and (Tirer & Bruna, 2022), but also extend them to the multilayer network setting:

- **Features collapse.** Since the collapse of the features of the backbone extractors implies the collapse of the features in subsequent linear layers, we only consider $\mathcal{NC}1$ metric for the output features of the backbone model. We recall the definition of the class-means and global-mean of the features $\{\mathbf{h}_{k,i}\}$ as:

$$\mathbf{h}_k := \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{k,i}, \quad \mathbf{h}_G := \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbf{h}_{k,i}.$$

  We also define the within-class, between-class covariance matrices, and $\mathcal{NC}1$ metric as following:

$$\mathbf{\Sigma}_W := \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}(\mathbf{h}_{k,i}-\mathbf{h}_{k,i})(\mathbf{h}_{k,i}-\mathbf{h}_{k,i})^\top, \quad \mathbf{\Sigma}_B := \frac{1}{K}\sum_{k=1}^{K}(\mathbf{h}_k-\mathbf{h}_G)(\mathbf{h}_k-\mathbf{h}_G)^\top,$$

$$\mathcal{NC}1 := \frac{1}{K}\text{trace}(\mathbf{\Sigma}_W\mathbf{\Sigma}_B^\dagger).$$

  where $\mathbf{\Sigma}_B^\dagger$ denotes the pseudo inverse of $\mathbf{\Sigma}_B$.

- **Convergence to OF/Simplex ETF.** To capture the $\mathcal{NC}$ behaviors across layers, we denote $\mathbf{W}^m := \mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_{M-m+1}$ as the product of last m weight matrices of the deep linear network. We define $\mathcal{NC}2_m^{OF}$ and $\mathcal{NC}2_m^{ETF}$ to measure the similarity of the learned classifiers $\mathbf{W}^m$ to OF (bias-free case) and ETF (last-layer bias case) as:

$$\mathcal{NC}2_m^{OF} := \left\|\frac{\mathbf{W}^m\mathbf{W}^{m\top}}{\|\mathbf{W}^m\mathbf{W}^{m\top}\|_F} - \frac{1}{\sqrt{K}}\mathbf{I}_K\right\|_F,$$
$$\mathcal{NC}2_m^{ETF} := \left\|\frac{\mathbf{W}^m\mathbf{W}^{m\top}}{\|\mathbf{W}^m\mathbf{W}^{m\top}\|_F} - \frac{1}{\sqrt{K-1}}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\right\|_F.$$

- **Convergence to self-duality.** We measure the alignment between the learned classifier $\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1$ and the learned class-means $\overline{\mathbf{H}}$ via:

$$\mathcal{NC}3^{OF} := \left\|\frac{\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1\overline{\mathbf{H}}}{\|\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1\overline{\mathbf{H}}\|_F} - \frac{1}{\sqrt{K}}\mathbf{I}_K\right\|_F,$$
$$\mathcal{NC}3^{ETF} := \left\|\frac{\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1\overline{\mathbf{H}}}{\|\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1\overline{\mathbf{H}}\|_F} - \frac{1}{\sqrt{K-1}}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\right\|_F,$$

  where $\overline{\mathbf{H}} = [\mathbf{h}_1,\ldots,\mathbf{h}_K]$ is the class-means matrix.

C.1.2. Additional numerical results for balanced data

This subsection expands upon the experiment results for balanced data in subsection 5.1 by the following points: i) For MLP experiment, we provide $\mathcal{NC}$ metrics measured at the last epoch for the remaining depth-widths combinations mentioned in subsection 5.1 and ii) Empirically verify Theorem A.1 of the $\mathcal{NC}$ existence for cross-entropy loss in deep linear network setting.

Last-epoch $\mathcal{NC}$ metrics for multilayer perceptron and deep learning experiments . We include the full set of last-epoch $\mathcal{NC}$ metrics for mentioned MLP depth-width combinations in Table 2 and 3. In which, Table 2 corresponds to the bias-free

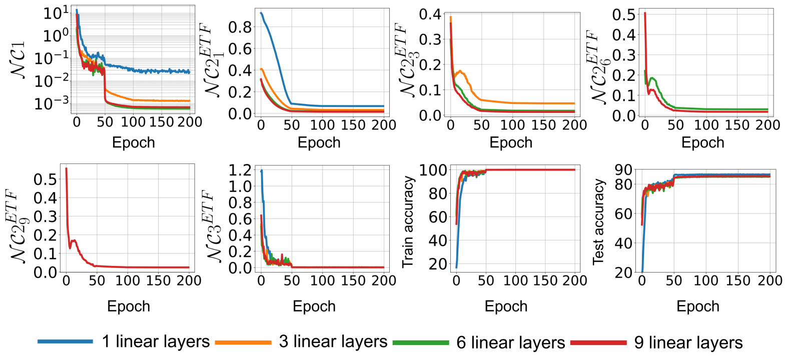


Figure 7. Illustration of $\mathcal{NC}$ for VGG16 backbone with MSE loss, balanced data and last-layer bias setting.

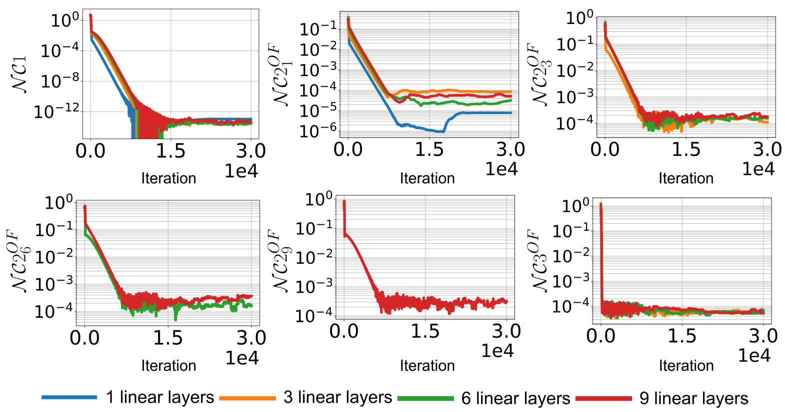


Figure 8. Illustration of $\mathcal{NC}$ for direct optimization experiment with MSE loss, balanced data and bias-free setting.

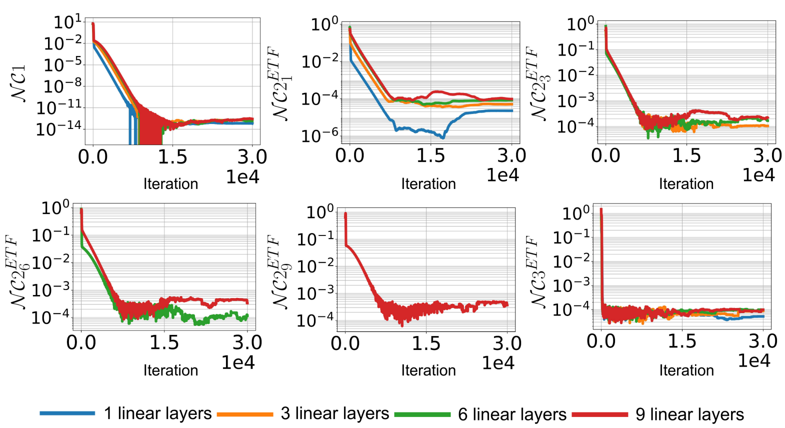


Figure 9. Illustration of $\mathcal{NC}$ for direct optimization experiment with MSE loss, balanced data and last-layer bias setting.

setting and Table 3 corresponds to the last-layer bias setting. Similarly, the full set of last-epoch $\mathcal{NC}$ metrics for deep learning experiments with ResNet18 and VGG19 models are also presented in Table 4.

Verification of Theorem A.1 for CE loss: We run two experiments to verify neural collapse for CE loss described in Theorem A.1 in two settings: MLP backbone model and direct optimization. Our network training procedure is similar to multilayer perceptron experiment and direct optimization experiment for last-layer bias setting described in subsection 5.1. For MLP experiment, we only change the learning rate to 0.0002 and substitute cross entropy loss in place of MSE loss. We run the experiment with all depth-width combinations with linear layer depth $\in \{1, 3\}$ and width $\in \{512, 1024, 2048\}$. For direct optimization experiment, we change learning rate to 0.02, width to 256 and keep other settings to be the same.

No. layer	Hidden dim	$\mathcal{NC}1$	$\mathcal{NC}2_1^{OF}$	$\mathcal{NC}2_2^{OF}$	$\mathcal{NC}2_3^{OF}$	$\mathcal{NC}2_4^{OF}$	$\mathcal{NC}2_5^{OF}$	$\mathcal{NC}2_6^{OF}$	$\mathcal{NC}2_7^{OF}$	$\mathcal{NC}2_8^{OF}$	$\mathcal{NC}2_9^{OF}$	$\mathcal{NC}3^{OF}$
1	512	1.819×10^{-3}	5.856×10^{-2}									1.769×10^{-2}
	1024	2.437×10^{-4}	3.024×10^{-2}									1.528×10^{-2}
	2048	1.259×10^{-4}	1.467×10^{-2}									1.712×10^{-2}
3	512	8.992×10^{-3}	5.09×10^{-2}	1.057×10^{-1}	1.486×10^{-1}							2.958×10^{-2}
	1024	2.843×10^{-3}	5.697×10^{-2}	1.009×10^{-1}	1.731×10^{-1}							2.368×10^{-2}
	2048	5.165×10^{-4}	3.857×10^{-2}	5.799×10^{-2}	8.648×10^{-2}							2.797×10^{-2}
6	512	8.701×10^{-3}	7.833×10^{-2}	1.009×10^{-1}	1.186×10^{-1}	1.340×10^{-1}	1.511×10^{-1}	1.824×10^{-1}				3.478×10^{-2}
	1024	2.578×10^{-3}	8.356×10^{-2}	1.066×10^{-1}	1.283×10^{-1}	1.489×10^{-1}	1.725×10^{-1}	2.429×10^{-1}				1.928×10^{-2}
	2048	8.231×10^{-4}	7.187×10^{-2}	9.224×10^{-2}	1.078×10^{-1}	1.160×10^{-1}	1.214×10^{-1}	1.386×10^{-1}				3.430×10^{-2}
9	512	9.359×10^{-3}	1.149×10^{-1}	1.480×10^{-1}	1.703×10^{-1}	1.824×10^{-1}	1.868×10^{-1}	1.855×10^{-1}	1.821×10^{-1}	1.823×10^{-1}	2.033×10^{-1}	3.074×10^{-2}
	1024	2.615×10^{-3}	1.165×10^{-1}	1.488×10^{-1}	1.745×10^{-1}	1.893×10^{-1}	1.961×10^{-1}	1.975×10^{-1}	1.972×10^{-1}	2.013×10^{-1}	2.492×10^{-1}	2.089×10^{-2}
	2048	7.694×10^{-4}	1.070×10^{-1}	1.402×10^{-1}	1.701×10^{-1}	1.864×10^{-1}	1.929×10^{-1}	1.892×10^{-1}	1.763×10^{-1}	1.592×10^{-1}	1.371×10^{-1}	2.141×10^{-2}

Table 2. Full set of metrics $\mathcal{NC}1$, $\mathcal{NC}2$, and $\mathcal{NC}3$ described in multilayer perceptron experiment in section 5.1 with bias-free setting.

No. layer	Hidden dim	$\mathcal{NC}1$	$\mathcal{NC}2_1^{ETF}$	$\mathcal{NC}2_2^{ETF}$	$\mathcal{NC}2_3^{ETF}$	$\mathcal{NC}2_4^{ETF}$	$\mathcal{NC}2_5^{ETF}$	$\mathcal{NC}2_6^{ETF}$	$\mathcal{NC}2_7^{ETF}$	$\mathcal{NC}2_8^{ETF}$	$\mathcal{NC}2_9^{ETF}$	$\mathcal{NC}3^{ETF}$
1	512	2.058×10^{-3}	4.936×10^{-2}									5.406×10^{-3}
	1024	2.791×10^{-4}	2.540×10^{-2}									3.862×10^{-3}
	2048	1.434×10^{-4}	9.418×10^{-3}									1.750×10^{-3}
3	512	7.601×10^{-3}	5.147×10^{-2}	1.124×10^{-1}	1.586×10^{-1}							1.972×10^{-2}
	1024	2.194×10^{-3}	5.967×10^{-2}	1.071×10^{-1}	1.949×10^{-1}							1.155×10^{-2}
	2048	6.397×10^{-4}	3.447×10^{-2}	5.795×10^{-2}	9.811×10^{-2}							5.311×10^{-3}
6	512	8.308×10^{-3}	2.006×10^{-2}	5.110×10^{-2}	8.624×10^{-2}	1.221×10^{-1}	1.587×10^{-1}	1.997×10^{-1}				1.757×10^{-2}
	1024	2.258×10^{-3}	2.818×10^{-2}	6.244×10^{-1}	9.861×10^{-2}	1.350×10^{-1}	1.710×10^{-1}	2.350×10^{-1}				1.320×10^{-2}
	2048	5.653×10^{-4}	1.848×10^{-2}	3.409×10^{-2}	5.134×10^{-2}	6.849×10^{-2}	8.570×10^{-2}	1.279×10^{-1}				4.522×10^{-3}
9	512	9.745×10^{-3}	1.608×10^{-2}	2.040×10^{-2}	3.916×10^{-2}	6.095×10^{-2}	8.494×10^{-2}	1.107×10^{-1}	1.383×10^{-1}	1.679×10^{-1}	2.102×10^{-1}	1.772×10^{-2}
	1024	2.587×10^{-3}	1.522×10^{-2}	2.462×10^{-2}	4.350×10^{-2}	6.525×10^{-2}	8.910×10^{-2}	1.147×10^{-1}	1.422×10^{-1}	1.711×10^{-1}	2.370×10^{-1}	1.245×10^{-2}
	2048	6.943×10^{-4}	1.217×10^{-2}	2.043×10^{-2}	3.218×10^{-2}	4.517×10^{-2}	5.899×10^{-1}	7.350×10^{-2}	8.881×10^{-2}	1.042×10^{-1}	1.414×10^{-1}	7.937×10^{-3}

Table 3. Full set of metrics $\mathcal{NC}1$, $\mathcal{NC}2$, and $\mathcal{NC}3$ in multilayer perceptron experiment in section 5.1 with last-layer bias setting.

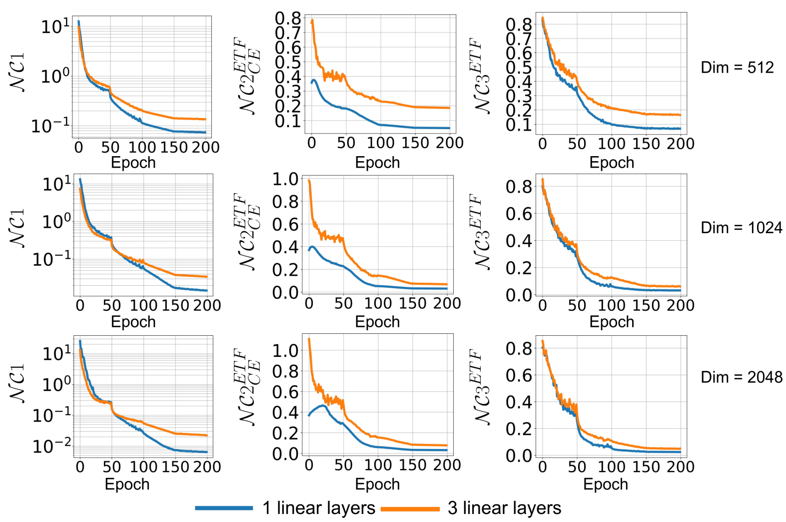


Figure 10. Illustration of $\mathcal{NC}$ with 6-layer MLP backbone on CIFAR10 for cross entropy loss, balanced data and last-layer bias setting.

Theorem A.1 indicates that all the features of the same class converge to a single vector, and the alignment between the learned classifier $\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1$ and the learned class-means $\overline{\mathbf{H}}$ has ETF form. Therefore, we use the same $\mathcal{NC}1$ and $\mathcal{NC}3$ as in the balanced data, last-layer bias case. Theorem A.1 also indicates that $\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1$ converges to ETF form. Hence, the metric used for CE loss to measure the convergence of $\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1$ is defined as $\mathcal{NC}2_{CE}^{ETF} := \mathcal{NC}2_M^{ETF}$, where $\mathcal{NC}2_M^{ETF}$ is defined in C.1.1. Fig. 10 and Fig. 11 demonstrate the convergence of $\mathcal{NC}$ for MLP and direct optimization experiments, respectively. The convergence to 0 of the $\mathcal{NC}$ metrics verifies theorem A.1.

C.1.3. Details of network training and hyperparameters for balanced data experiments

Multilayer perceptron experiment: In this experiment, we use a 6-layer MLP model with ReLU activation as the backbone feature extractor. Hidden width of the backbone model and the deep linear network are set to be equal. We cover all

Model name	No.layer	$\mathcal{NC}1$	$\mathcal{NC}2_1^{ETF}$	$\mathcal{NC}2_2^{ETF}$	$\mathcal{NC}2_3^{ETF}$	$\mathcal{NC}2_4^{ETF}$	$\mathcal{NC}2_5^{ETF}$	$\mathcal{NC}2_6^{ETF}$	$\mathcal{NC}2_7^{ETF}$	$\mathcal{NC}2_8^{ETF}$	$\mathcal{NC}2_9^{ETF}$	$\mathcal{NC}3^{ETF}$
ResNet18	1	1.556×10^{-3}	4.376×10^{-2}									3.598×10^{-3}
	3	4.713×10^{-4}	2.191×10^{-2}	4.714×10^{-2}	7.813×10^{-2}							2.131×10^{-3}
	6	1.824×10^{-4}	4.295×10^{-3}	4.868×10^{-3}	7.651×10^{-3}	1.156×10^{-2}	1.681×10^{-2}	2.459×10^{-2}				1.817×10^{-3}
	9	2.156×10^{-4}	3.609×10^{-3}	6.459×10^{-3}	7.835×10^{-3}	8.056×10^{-3}	8.096×10^{-3}	8.362×10^{-3}	9.400×10^{-3}	1.212×10^{-2}	1.683×10^{-2}	2.210×10^{-3}
VGG16	1	2.447×10^{-2}	6.689×10^{-2}									1.977×10^{-3}
	3	1.347×10^{-3}	3.120×10^{-2}	3.035×10^{-2}	4.606×10^{-2}							2.767×10^{-3}
	6	5.959×10^{-4}	1.645×10^{-2}	1.266×10^{-2}	1.703×10^{-2}	2.183×10^{-2}	2.473×10^{-2}	3.015×10^{-2}				2.483×10^{-3}
	9	6.893×10^{-4}	1.438×10^{-2}	9.511×10^{-3}	1.198×10^{-2}	1.314×10^{-2}	1.619×10^{-2}	1.774×10^{-2}	2.030×10^{-2}	2.218×10^{-2}	2.445×10^{-2}	2.434×10^{-3}

Table 4. Full set of metrics $\mathcal{NC}1$, $\mathcal{NC}2$, and $\mathcal{NC}3$ described in deep learning experiment in section 5.1 for ResNet18 and VGG16 backbones with last-layer bias setting.

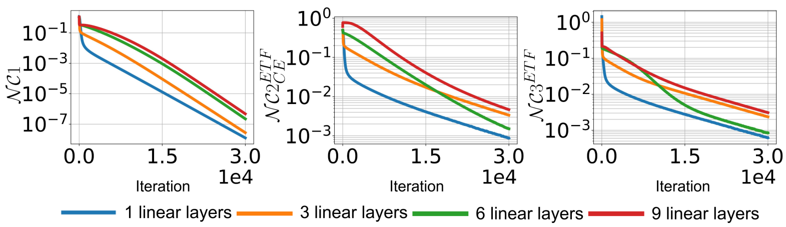


Figure 11. Illustration of $\mathcal{NC}$ for direct optmization experiment with cross-entropy loss, balanced data and last-layer bias setting.

depth-width combinations with depth $\in \{1, 3, 6, 9\}$ and width $\in \{512, 1024, 2048\}$ for two settings, bias-free and last-layer bias. All models are trained with Adam optimizer with MSE loss for 200 epochs with batch size 128 and learning rate 0.0001 (divided by 10 every 50 epochs). Weight decay and feature decay are set to 1×10^{-4}.

Deep learning experiment: In deep learning experiment, we use ResNet18 and VGG16 as backbones feature extractors. We train both models with SGD optimizer with batch size 128 for MSE loss. Data augmentation is not used in this experiment. The learning rate decays 0.1 every 50 epochs for 200 epochs. Depth of the deep linear layers are selected from the set $\{1, 3, 6, 9\}$. Width of the deep linear layers are set to 512 to be equal to the last-layer dimension of the backbone model. Weight decay in both models is enforced on all network parameters to align with the typical training protocol. For ResNet18 backbone models, we use the learning rate of 0.05 and weight decay of 2×10^{-4}. For VGG16 backbone, the learning rate is 0.02. Except for VGG16-backbone with 1 linear layer using weight decay of 5×10^{-4}, all other VGG16-backbone models shares the weight decay of 3×10^{-4}.

Direct optimization experiment: In this experiment, we replicate the optimization problem (3). $\mathbf{W}_M, \ldots, \mathbf{W}_1$ and $\mathbf{H}_1$ are initialized with standard normal distribution scaled by 0.1. We set $K = 4, n = 100, d_M = \ldots = d_1 = 64$ and all λ's are set to be 5×10^{-4}. Depth of the linear layers are selected from the set $\{1, 3, 6, 9\}$. $\mathbf{W}_M, \ldots, \mathbf{W}_1$ and $\mathbf{H}_1$ are optimized by gradient descent for 30000 iterations with learning rate 0.1.

C.2. Imbalanced Data

C.2.1. Metric for measuring $\mathcal{NC}$ in imbalanced data

For imbalanced setting, $\mathcal{NC}1$ metric is identical to the balanced setting's. While for $\mathcal{NC}2$ and $\mathcal{NC}3$, we measure the closeness of learned classifiers and features to GOF structure as follows:

$$\mathcal{NC}2^{GOF} := \left\| \frac{(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)^\top}{\|(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)^\top\|_F} - \frac{\operatorname{diag}\{c s_k^{2M}\}_{k=1}^K}{\|\operatorname{diag}\{c s_k^{2M}\}_{k=1}^K\|_F} \right\|_F,$$

$$\mathcal{NC}3^{GOF} := \left\| \frac{\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1 \overline{\mathbf{H}}}{\|\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1 \overline{\mathbf{H}}\|_F} - \frac{\operatorname{diag}\left\{\frac{c s_k^{2M}}{c s_k^{2M} + N\lambda_{H_1}}\right\}_{k=1}^K}{\left\|\operatorname{diag}\left\{\frac{c s_k^{2M}}{c s_k^{2M} + N\lambda_{H_1}}\right\}_{k=1}^K\right\|_F} \right\|_F,$$

where $\overline{\mathbf{H}} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$ is the class-means matrix, c and $\{s_k\}_{k=1}^K$ are as defined in Theorem 4.4.

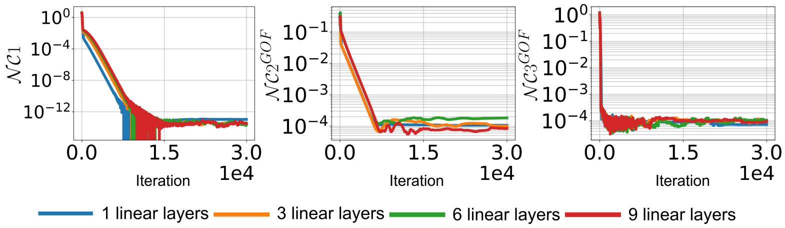


Figure 12. Illustration of $\mathcal{NC}$ for direct optimization experiment with MSE loss, imbalanced data and bias-free setting.

C.2.2. Additional numerical results for imbalanced data

Continue from subsection 5.2, to empirically validate the Minority Collapse of the problems (5) and (6), we run two direct optimization schemes similar as Section 5.2 with heavy imbalanced data of $K = 4$ and $n_1 = 2000, n_2 = n_3 = 495$ and $n_4 = 10$ for $M = 1$ ($d = 16$) and $M = 3$ ($d = 40$). Both models are trained by gradient descent for 30000 iterations. The final weight matrices of these models are as following (results are rounded to 2 decimal places):

$$\mathbf{W}_1 = \begin{bmatrix} -1.55 & 1.50 & 2.19 & -1.36 & -0.65 & 3.08 & -0.81 & -1.76 & -0.96 & -0.48 & -1.21 & -1.06 & 1.01 & 1.72 & 0.30 & -1.73 \\ -1.26 & -0.56 & -0.94 & -1.24 & 0.11 & -1.46 & -0.51 & -1.75 & -0.69 & 0.11 & 1.09 & -0.89 & -0.56 & 0.57 & 0.48 & 0.27 \\ 0.76 & -0.31 & 0.32 & -1.30 & -0.42 & 0.09 & 2.22 & -1.07 & 1.15 & -0.58 & -0.28 & -0.88 & -0.03 & -0.40 & -1.29 & 0.43 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix},$$

for case $M = 1$. For case $M = 3$, we have:

$$\mathbf{W}_3 = \begin{bmatrix} 0.65 & -0.96 & 0.49 & -0.15 & 0.50 & -0.11 & -0.14 & 0.40 & \dots & 0.02 & 0.05 & 0.27 & 0.13 & 0.71 & -0.29 & 0.14 & -0.30 \\ -0.25 & 0.13 & -0.40 & -0.33 & 0.14 & 0.11 & -0.32 & 0.15 & \dots & 0.40 & -0.10 & -0.86 & 0.34 & 0.20 & 0.54 & 0.66 & 0.18 \\ 0.36 & -0.15 & -0.04 & -0.23 & -0.66 & -0.04 & -0.51 & -0.33 & \dots & -0.07 & -0.52 & 0.15 & -0.03 & 0.04 & -0.36 & 0.35 & 0.02 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & \dots & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}. \tag{8}$$

As can be seen from both cases, the classifier of the fourth class converges to zero vector (with the convergence error are less than 1e-8), due to the heavy imbalance level of the dataset, which align to Theorem 4.1 and Theorem 4.4.

C.2.3. Details of network training and hyperparameters for imbalanced data experiments

Multilayer perceptron experiment: In this experiment, we use a subset of CIFAR10 dataset with training samples of each class in the list $\{500, 500, 400, 400, 300, 300, 200, 200, 100, 100\}$. We use a 6-layer MLP model with ReLU activation with removed activation as the backbone feature extractor. Hidden width of both the backbone model and the deep linear networks are set to be 2048. Depth of the linear layers are selected from the set $\{1, 3, 6\}$. All models are trained with Adam optimizer and MSE loss for 12000 epochs, no data augmentation, full batch gradient descent, learning rate 1×10^{-4} (divided by 10 every 6000 epochs), feature decay and weight decay are set to be 1×10^{-5}.

Direct optimization experiment: In this experiment, we replicate the optimization problem (3) in imbalance data setting. We set $K = 4$ and $n_1 = 200, n_2 = 100, n_3 = n_4 = 50, d_M = \ldots = d_1 = 64$. Similar to the direct optimization experiment in balance case, all λ's are set to be 5×10^{-4}. $\mathbf{W}_M, \ldots, \mathbf{W}_1$ and $\mathbf{H}_1$ are optimized by stochastic gradient descent for 30000 iterations, with learning rate 0.1.

D. Proof of Theorem 3.1

First we state the proof for UFM bias-free with three layers of weights with same width across layers, as a warm-up for our approach in the next proofs.

D.1. Warm-up Case: UFM with Three Layers of Weights

Consider the following bias-free optimization problem:

$$\min_{\mathbf{W}_3, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1} \frac{1}{2N} \|\mathbf{W}_3 \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1 - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_3}}{2} \|\mathbf{W}_3\|_F^2 + \frac{\lambda_{W_2}}{2} \|\mathbf{W}_2\|_F^2 + \frac{\lambda_{W_1}}{2} \|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2} \|\mathbf{H}_1\|_F^2 \tag{9}$$

where $\lambda_{W_3}, \lambda_{W_2}, \lambda_{W_1}, \lambda_{H_1}$ are regularization hyperparameters, and $\mathbf{W}_3 \in \mathbb{R}^{K \times d}$, $\mathbf{W}_2 \in \mathbb{R}^{d \times d}$, $\mathbf{W}_1 \in \mathbb{R}^{d \times d}$, $\mathbf{H}_1 \in \mathbb{R}^{d \times N}$ and $\mathbf{Y} \in \mathbb{R}^{K \times N}$. We assume $d \geq K$ for this problem.

Proof of Theorem 3.1 with 3 layers of weight and $d \geq K$. By definition, any critical point $(\mathbf{W}_3, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1)$ of the loss function (9) satisfies the following :

$$\frac{\partial f}{\partial \mathbf{W}_3} = \frac{1}{N}(\mathbf{W}_3\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1 - \mathbf{Y})\mathbf{H}_1^\top \mathbf{W}_1^\top \mathbf{W}_2^\top + \lambda_{W_3}\mathbf{W}_3 = \mathbf{0}, \tag{10}$$

$$\frac{\partial f}{\partial \mathbf{W}_2} = \frac{1}{N}\mathbf{W}_3^\top(\mathbf{W}_3\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1 - \mathbf{Y})\mathbf{H}_1^\top \mathbf{W}_1^\top + \lambda_{W_2}\mathbf{W}_2 = \mathbf{0}, \tag{11}$$

$$\frac{\partial f}{\partial \mathbf{W}_1} = \frac{1}{N}\mathbf{W}_2^\top\mathbf{W}_3^\top(\mathbf{W}_3\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1 - \mathbf{Y})\mathbf{H}_1^\top + \lambda_{W_1}\mathbf{W}_1 = \mathbf{0}, \tag{12}$$

$$\frac{\partial f}{\partial \mathbf{H}_1} = \frac{1}{N}\mathbf{W}_1^\top\mathbf{W}_2^\top\mathbf{W}_3^\top(\mathbf{W}_3\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1 - \mathbf{Y}) + \lambda_{H_1}\mathbf{H}_1 = \mathbf{0}. \tag{13}$$

Next, from $\mathbf{W}_3^\top \frac{\partial f}{\partial \mathbf{W}_3} - \frac{\partial f}{\partial \mathbf{W}_2}\mathbf{W}_2^\top = \mathbf{0}$, we have:

$$\lambda_{W_3}\mathbf{W}_3^\top\mathbf{W}_3 = \lambda_{W_2}\mathbf{W}_2\mathbf{W}_2^\top. \tag{14}$$

Similarly, we also have:

$$\lambda_{W_2}\mathbf{W}_2^\top\mathbf{W}_2 = \lambda_{W_1}\mathbf{W}_1\mathbf{W}_1^\top, \tag{15}$$

$$\lambda_{W_1}\mathbf{W}_1^\top\mathbf{W}_1 = \lambda_{H_1}\mathbf{H}_1\mathbf{H}_1^\top. \tag{16}$$

Also, from equation (13), by solving for $\mathbf{H}_1$, we have:

$$\begin{aligned}
\mathbf{H}_1 &= (\mathbf{W}_1^\top\mathbf{W}_2^\top\mathbf{W}_3^\top\mathbf{W}_3\mathbf{W}_2\mathbf{W}_1 + N\lambda_{H_1}\mathbf{I})^{-1}\mathbf{W}_1^\top\mathbf{W}_2^\top\mathbf{W}_3^\top\mathbf{Y} \\
&= \left(\frac{\lambda_{W_2}}{\lambda_{W_3}}\mathbf{W}_1^\top(\mathbf{W}_2^\top\mathbf{W}_2)^2\mathbf{W}_1 + N\lambda_{H_1}\mathbf{I}\right)^{-1}\mathbf{W}_1^\top\mathbf{W}_2^\top\mathbf{W}_3^\top\mathbf{Y} \\
&= \left(\frac{\lambda_{W_1}^2}{\lambda_{W_3}\lambda_{W_2}}(\mathbf{W}_1^\top\mathbf{W}_1)^3 + N\lambda_{H_1}\mathbf{I}\right)^{-1}\mathbf{W}_1^\top\mathbf{W}_2^\top\mathbf{W}_3^\top\mathbf{Y},
\end{aligned} \tag{17}$$

where we use equations (14) and (15) for the derivation.

Now, let $\mathbf{W}_1 = \mathbf{U}_{W_1}\mathbf{S}_{W_1}\mathbf{V}_{W_1}^\top$ be the SVD decomposition of $\mathbf{W}_1$ with $\mathbf{U}_{W_1}, \mathbf{V}_{W_1} \in \mathbb{R}^{d\times d}$ are orthonormal matrix and $\mathbf{S}_{W_1} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with **decreasing** non-negative singular values. We note that from equations (14)-(16), we have $\operatorname{rank}(\mathbf{W}_3^\top\mathbf{W}_3) = \operatorname{rank}(\mathbf{W}_3) = \operatorname{rank}(\mathbf{W}_2) = \operatorname{rank}(\mathbf{W}_1) = \operatorname{rank}(\mathbf{H}_1)$ and is at most K. We denote the K singular values (some of them can be 0's) of $\mathbf{W}_1$ as $\{s_k\}_{k=1}^K$.

From equation (15), we have:

$$\mathbf{W}_2^\top\mathbf{W}_2 = \frac{\lambda_{W_1}}{\lambda_{W_2}}\mathbf{W}_1\mathbf{W}_1^\top = \frac{\lambda_{W_1}}{\lambda_{W_2}}\mathbf{U}_{W_1}\mathbf{S}_{W_1}^2\mathbf{U}_{W_1}^\top = \mathbf{U}_{W_1}\mathbf{S}_{W_2}^2\mathbf{U}_{W_1}^\top,$$

where $\mathbf{S}_{W_2} = \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_2}}}\mathbf{S}_{W_1} \in \mathbb{R}^{d\times d}$. This means that $\mathbf{S}_{W_2}^2$ contains the eigenvalues and the columns of $\mathbf{U}_{W_1}$ are the eigenvectors of $\mathbf{W}_2^\top\mathbf{W}_2$. Hence, we can write the SVD decomposition of $\mathbf{W}_2$ as $\mathbf{W}_2 = \mathbf{U}_{W_2}\mathbf{S}_{W_2}\mathbf{U}_{W_1}^\top$ with orthonormal matrix $\mathbf{U}_{W_2} \in \mathbb{R}^{d\times d}$.

By making similar arguments for $\mathbf{W}_3$, from equation (14):

$$\mathbf{W}_3^\top\mathbf{W}_3 = \frac{\lambda_{W_2}}{\lambda_{W_3}}\mathbf{W}_2\mathbf{W}_2^\top = \frac{\lambda_{W_2}}{\lambda_{W_3}}\mathbf{U}_{W_2}\mathbf{S}_{W_2}^2\mathbf{U}_{W_2}^\top = \frac{\lambda_{W_1}}{\lambda_{W_3}}\mathbf{U}_{W_2}\mathbf{S}_{W_1}^2\mathbf{U}_{W_2}^\top = \mathbf{U}_{W_2}\mathbf{S}_{W_3}^\top\mathbf{S}_{W_3}\mathbf{U}_{W_2}^\top,$$

with $\mathbf{S}_{W_3} = \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_3}}} \begin{bmatrix} \operatorname{diag}(s_1, s_2, \ldots, s_K) & \mathbf{0}_{K \times (d-K)} \end{bmatrix} \in \mathbb{R}^{K \times d}$, we can write SVD decomposition of $\mathbf{W}_3$ as $\mathbf{W}_3 = \mathbf{U}_{W_3} \mathbf{S}_{W_3} \mathbf{U}_{W_2}^\top$ with orthonormal matrix $\mathbf{U}_{W_3} \in \mathbb{R}^{d \times d}$.

Using these SVD in the RHS of equation (17) yields:

$$\begin{aligned}
\mathbf{H}_1 &= \left(\frac{\lambda_{W_1}^2}{\lambda_{W_3}\lambda_{W_2}} (\mathbf{W}_1^\top \mathbf{W}_1)^3 + N\lambda_{H_1} \mathbf{I} \right)^{-1} \mathbf{W}_1^\top \mathbf{W}_2^\top \mathbf{W}_3^\top \mathbf{Y} \\
&= \left(\frac{\lambda_{W_1}^2}{\lambda_{W_3}\lambda_{W_2}} \mathbf{V}_{W_1} \mathbf{S}_{W_1}^6 \mathbf{V}_{W_1}^\top + N\lambda_{H_1} \mathbf{I} \right)^{-1} \mathbf{W}_1^\top \mathbf{W}_2^\top \mathbf{W}_3^\top \mathbf{Y} \\
&= \left(\frac{\lambda_{W_1}^2}{\lambda_{W_3}\lambda_{W_2}} \mathbf{V}_{W_1} \mathbf{S}_{W_1}^6 \mathbf{V}_{W_1}^\top + N\lambda_{H_1} \mathbf{I} \right)^{-1} \mathbf{V}_{W_1} \mathbf{S}_{W_1} \mathbf{S}_{W_2} \mathbf{S}_{W_3}^\top \mathbf{U}_{W_3}^\top \mathbf{Y} \\
&= \mathbf{V}_{W_1} \left(\frac{\lambda_{W_1}^2}{\lambda_{W_3}\lambda_{W_2}} \mathbf{S}_{W_1}^6 + N\lambda_{H_1} \mathbf{I} \right)^{-1} \mathbf{S}_{W_1} \mathbf{S}_{W_2} \mathbf{S}_{W_3}^\top \mathbf{U}_{W_3}^\top \mathbf{Y} \\
&= \mathbf{V}_{W_1} \left(\frac{\lambda_{W_1}^2}{\lambda_{W_3}\lambda_{W_2}} \mathbf{S}_{W_1}^6 + N\lambda_{H_1} \mathbf{I} \right)^{-1} \sqrt{\frac{\lambda_{W_1}^2}{\lambda_{W_3}\lambda_{W_2}}} \begin{bmatrix} \operatorname{diag}(s_1^3, s_2^3, \ldots, s_K^3) \\ \mathbf{0}_{(d-K) \times K} \end{bmatrix} \mathbf{U}_{W_3}^\top \mathbf{Y} \\
&= \mathbf{V}_{W_1} \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{\sqrt{c} s_1^3}{c s_1^6 + N\lambda_{H_1}}, \ldots, \frac{\sqrt{c} s_K^3}{c s_K^6 + N\lambda_{H_1}} \right) \\ \mathbf{0} \end{bmatrix}}_{\mathbf{C} \in \mathbb{R}^{d \times K}} \mathbf{U}_{W_3}^\top \mathbf{Y} \\
&= \mathbf{V}_{W_1} \mathbf{C} \mathbf{U}_{W_3}^\top \mathbf{Y},
\end{aligned} \tag{18}$$

with $c := \frac{\lambda_{W_1}^2}{\lambda_{W_3}\lambda_{W_2}}$. We further have:

$$\begin{aligned}
\mathbf{W}_3 \mathbf{W}_2 \mathbf{W}_1 \mathbf{H} &= \mathbf{U}_{W_3} \mathbf{S}_{W_3} \mathbf{S}_{W_2} \mathbf{S}_{W_1} \mathbf{V}_{W_1}^\top \mathbf{V}_{W_1} \mathbf{C} \mathbf{U}_{W_3}^\top \mathbf{Y} \\
&= \mathbf{U}_{W_3} \operatorname{diag}\left(\frac{c s_1^6}{c s_1^6 + N\lambda_{H_1}}, \ldots, \frac{c s_K^6}{c s_K^6 + N\lambda_{H_1}} \right) \mathbf{U}_{W_3}^\top \mathbf{Y}
\end{aligned} \tag{19}$$

$$\begin{aligned}
\Rightarrow \mathbf{W}_3 \mathbf{W}_2 \mathbf{W}_1 \mathbf{H} - \mathbf{Y} &= \mathbf{U}_{W_3} \left(\operatorname{diag}\left(\frac{c s_1^6}{c s_1^6 + N\lambda_{H_1}}, \ldots, \frac{c s_K^6}{c s_K^6 + N\lambda_{H_1}} \right) - \mathbf{I}_K \right) \mathbf{U}_{W_3}^\top \mathbf{Y} \\
&= \mathbf{U}_{W_3} \underbrace{\operatorname{diag}\left(\frac{-N\lambda_{H_1}}{c s_1^6 + N\lambda_{H_1}}, \ldots, \frac{-N\lambda_{H_1}}{c s_K^6 + N\lambda_{H_1}} \right)}_{\mathbf{D} \in \mathbb{R}^{K \times K}} \mathbf{U}_{W_3}^\top \mathbf{Y} \\
&= \mathbf{U}_{W_3} \mathbf{D} \mathbf{U}_{W_3}^\top \mathbf{Y}.
\end{aligned} \tag{20}$$

Next, we will calculate the Frobenius norm of $\mathbf{W}_3 \mathbf{W}_2 \mathbf{W}_1 \mathbf{H} - \mathbf{Y}$:

$$\begin{aligned}
\|\mathbf{W}_3 \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1 - \mathbf{Y}\|_F^2 &= \|\mathbf{U}_{W_3} \mathbf{D} \mathbf{U}_{W_3}^\top \mathbf{Y}\|_F^2 = \operatorname{trace}(\mathbf{U}_{W_3} \mathbf{D} \mathbf{U}_{W_3}^\top \mathbf{Y} (\mathbf{U}_{W_3} \mathbf{D} \mathbf{U}_{W_3}^\top \mathbf{Y})^\top) \\
&= \operatorname{trace}(\mathbf{U}_{W_3} \mathbf{D} \mathbf{U}_{W_3}^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{U}_{W_3} \mathbf{D} \mathbf{U}_{W_3}^\top) = \operatorname{trace}(\mathbf{D}^2 \mathbf{U}_{W_3}^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{U}_{W_3}) \\
&= n \operatorname{trace}(\mathbf{D}^2) = n \sum_{k=1}^{K} \left(\frac{-N\lambda_{H_1}}{c s_k^6 + N\lambda_{H_1}} \right)^2.
\end{aligned} \tag{21}$$

where we use the fact $\mathbf{Y}\mathbf{Y}^\top = n\mathbf{I}_K$ and $\mathbf{U}_{W_3}$ is orthonormal matrix.

Similarly, from the RHS of equation (18), we have:

$$\begin{aligned}
\|\mathbf{H}_1\|_F^2 &= \operatorname{trace}(\mathbf{V}_{W_1} \mathbf{C} \mathbf{U}_{W_3}^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{U}_{W_3} \mathbf{C}^\top \mathbf{V}_{W_1}^\top) = \operatorname{trace}(\mathbf{C}^\top \mathbf{C} \mathbf{U}_{W_3}^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{U}_{W_3}) \\
&= n \operatorname{trace}(\mathbf{C}^\top \mathbf{C}) = n \sum_{k=1}^{K} \left(\frac{\sqrt{c} s_k^3}{c s_k^6 + N\lambda_{H_1}} \right)^2.
\end{aligned} \tag{22}$$

Now, we will plug equations (21), (22), and the SVD decomposition of $\mathbf{W}_2, \mathbf{W}_1, \mathbf{H}$ into the function (9) and note that orthonormal matrix does not change the Frobenius form:

$$
\begin{aligned}
&f(\mathbf{W}_3, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1) = \frac{1}{2N} \left\| \mathbf{W}_3 \mathbf{W}_2 \mathbf{W}_1 \mathbf{H} - \mathbf{I}_K \right\|_F^2 + \frac{\lambda_{W_3}}{2} \left\| \mathbf{W}_3 \right\|_F^2 + \frac{\lambda_{W_2}}{2} \left\| \mathbf{W}_2 \right\|_F^2 + \frac{\lambda_{W_1}}{2} \left\| \mathbf{W}_1 \right\|_F^2 + \frac{\lambda_{H_1}}{2} \left\| \mathbf{H_1} \right\|_F^2 \\
&= \frac{1}{2K} \sum_{k=1}^K \left(\frac{-N\lambda_{H_1}}{cs_k^6 + N\lambda_{H_1}} \right)^2 + \frac{\lambda_{W_3}}{2} \sum_{k=1}^K \frac{\lambda_{W_1}}{\lambda_{W_3}} s_k^2 + \frac{\lambda_{W_2}}{2} \sum_{k=1}^K \frac{\lambda_{W_1}}{\lambda_{W_2}} s_k^2 + \frac{\lambda_{W_1}}{2} \sum_{k=1}^K s_k^2 + \frac{n\lambda_{H_1}}{2} \sum_{k=1}^K \frac{cs_k^6}{(cs_k^6 + N\lambda_{H_1})^2} \\
&= \frac{n\lambda_{H_1}}{2} \sum_{k=1}^K \frac{1}{cs_k^6 + N\lambda_{H_1}} + \frac{3\lambda_{W_1}}{2} \sum_{k=1}^K s_k^2 \\
&= \frac{1}{2K} \sum_{k=1}^K \left(\frac{1}{\frac{cs_k^6}{N\lambda_{H_1}} + 1} + 3K\lambda_{W_1} \frac{\sqrt[3]{N\lambda_{H_1}}}{\sqrt[3]{c}} \frac{\sqrt[3]{c} s_k^2}{\sqrt[3]{N\lambda_{H_1}}} \right) \\
&= \frac{1}{2K} \sum_{k=1}^K \left(\frac{1}{x_k^3 + 1} + bx_k \right), \qquad (23)
\end{aligned}
$$

with $x_k := \frac{\sqrt[3]{c} s_k^2}{\sqrt[3]{N\lambda_{H_1}}}$ and $b := 3K\lambda_{W_1} \frac{\sqrt[3]{N\lambda_{H_1}}}{\sqrt[3]{c}} = 3K \sqrt[3]{N\lambda_{W_3}\lambda_{W_2}\lambda_{W_1}\lambda_{H_1}}$.

Next, we consider the function:

$$
g(x) = \frac{1}{x^3 + 1} + bx \text{ with } x \geq 0, b > 0. \qquad (24)
$$

Clearly, $g(0) = 1$. As in equation (23), $f(\mathbf{W}_3, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H})$ is the sum of $g(x_k)$ (with separable x_k). Hence, if we can minimize $g(x)$, we will finish lower bounding $f(\mathbf{W}_3, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H})$. We consider the following cases for $g(x)$:

- If $b > \frac{\sqrt[3]{4}}{3}$: For $x > 0$, we always have $g(x) > \frac{1}{x^3+1} + \frac{\sqrt[3]{4}}{3} x \geq 1 = g(0)$. Indeed, the second inequality is equivalent to:

$$
\begin{aligned}
&\frac{1}{x^3 + 1} + \frac{\sqrt[3]{4}}{3} x \geq 1 \\
\Leftrightarrow \quad &\frac{\sqrt[3]{4}}{3} x^4 - x^3 + \frac{\sqrt[3]{4}}{3} x \geq 0 \\
\Leftrightarrow \quad &x(x + \frac{1}{\sqrt[3]{4}})(x - \sqrt[3]{2})^2 \geq 0.
\end{aligned}
$$

  Therefore, in this case, $g(x)$ is minimized at $x = 0$ with minimal value of 1.

- If $b = \frac{\sqrt[3]{4}}{3}$: Similar as above, we have:

$$
\begin{aligned}
&g(x) \geq 1 \\
\Leftrightarrow \quad &x(x + \frac{1}{\sqrt[3]{4}})(x - \sqrt[3]{2})^2 \geq 0.
\end{aligned}
$$

  In this case, $g(x)$ is minimized at $x = 0$ or $x = \sqrt[3]{2}$.

- If $b < \frac{\sqrt[3]{4}}{3}$: We take the first and second derivatives of $g(x)$:

$$
\begin{aligned}
g'(x) &= b - \frac{3x^2}{(x^3 + 1)^2}, \\
g''(x) &= \frac{12x^4 - 6x}{(x^3 + 1)^3}.
\end{aligned}
$$

We have: $g''(x) = 0 \Leftrightarrow x = 0$ or $x = \sqrt[3]{\frac{1}{2}}$. Therefore, with $x \geq 0$, $g'(x) = 0$ has at most two solutions. We also have $g'\left(\sqrt[3]{\frac{1}{2}}\right) = b - \frac{2\sqrt[3]{2}}{3} < 0$ (since $b < \frac{\sqrt[3]{4}}{3}$). Thus, together with the fact that $g'(0) = b > 0$ and $g(+\infty) > 0$, $g'(x) = 0$ has exactly two solutions, we call it x_1 and x_2 ($x_1 < \sqrt[3]{\frac{1}{2}} < x_2$). Next, we note that $g'(x_2) = 0$ and $g'(x) > 0 \quad \forall x > x_2$ (since $g''(x) > 0 \quad \forall x > x_2$). In the meanwhile, $g'(\sqrt[3]{2}) = b - \frac{\sqrt[3]{4}}{3} < 0$. Hence, we must have $x_2 > \sqrt[3]{2}$.

From the variation table, we can see that $g(x_2) < g(\sqrt[3]{2}) = \frac{1}{3} + b\sqrt[3]{2} < \frac{1}{3} + \frac{2}{3} = 1 = g(0)$. Hence, the minimizer in this case is the largest solution $x > \sqrt[3]{2}$ of the equation $g'(x) = 0$.

x	0	x_1	$\sqrt[3]{\frac{1}{2}}$	$\sqrt[3]{2}$	x_2	∞
g''	0	-	0	+	+	+
g'	+	0	-	-	0	+
g	1	$g(x_1)$	$g\left(\sqrt[3]{\frac{1}{2}}\right)$	$\frac{1}{3} + b\sqrt[3]{2}$	$g(x_2)$	∞

From the above result, we can summarize the original problem as follows:

- If $b = 3K\sqrt[3]{Kn\lambda_{W_3}\lambda_{W_2}\lambda_{W_1}\lambda_{H_1}} > \frac{\sqrt[3]{4}}{3}$: all the singular values of $\mathbf{W}_1^*$ are 0's. Therefore, the singular values of $\mathbf{W}_3^*, \mathbf{W}_1^*, \mathbf{H}^*$ are also all 0's. In this case, $f(\mathbf{W}_3, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1)$ is minimized at $(\mathbf{W}_3^*, \mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^*) = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0})$.
- If $b = 3K\sqrt[3]{Kn\lambda_{W_3}\lambda_{W_2}\lambda_{W_1}\lambda_{H_1}} < \frac{\sqrt[3]{4}}{3}$: In this case, $\mathbf{W}_1^*$ has K singular values, all of which are multiplier of the largest positive solution of the equation $b - \frac{3x^2}{(x^3+1)^2} = 0$, denoted as s. Hence, we have the compact SVD form (with a bit of notation abuse) of $\mathbf{W}_1^*$ as $\mathbf{W}_1^* = s\mathbf{U}_{W_1}\mathbf{V}_{W_1}^\top$ with semi-orthonormal matrices $\mathbf{U}_{W_1}, \mathbf{V}_{W_1} \in \mathbb{R}^{d\times K}$. We also have $\mathbf{U}_{W_1}^\top\mathbf{U}_{W_1} = \mathbf{I}_K$ and $\mathbf{V}_{W_1}^\top\mathbf{V}_{W_1} = \mathbf{I}_K$.

  Similarly, since the singular matrices of $\mathbf{W}_3, \mathbf{W}_1$ are aligned to $\mathbf{W}_1$'s, we also have:

$$\mathbf{W}_3^* = \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_3}}} s\mathbf{U}_{W_3}\mathbf{U}_{W_2}^T,$$
$$\mathbf{W}_2^* = \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_2}}} s\mathbf{U}_{W_2}\mathbf{U}_{W_1}^\top,$$
$$\mathbf{W}_1^* = s\mathbf{U}_{W_1}\mathbf{V}_{W_1}^\top,$$
$$\mathbf{H}_1^* = \frac{\sqrt{c}s^3}{cs^6 + N\lambda_{H_1}}\mathbf{V}_{W_1}\mathbf{U}_{W_3}^\top\mathbf{Y},$$

  with orthonormal matrices $\mathbf{U}_{W_3} \in \mathbb{R}^{K\times K}$, semi-orthonormal matrix $\mathbf{U}_{W_2}, \mathbf{U}_{W_1}, \mathbf{V}_{W_1} \in \mathbb{R}^{d\times K}$. Let $\overline{\mathbf{H}}^* = \frac{\sqrt{c}s^3}{cs^6+N\lambda_{H_1}}\mathbf{V}_{W_1}\mathbf{U}_{W_3}^\top \in \mathbb{R}^{K\times K}$, we have: $\mathbf{H}_1^* = \overline{\mathbf{H}}^*\mathbf{Y} = \overline{\mathbf{H}}^* \otimes \mathbf{1}_n^\top$.

  We have the geometry of the global solutions as follows:

$$\begin{aligned}
\mathbf{W}_3^*\mathbf{W}_3^{\top *} &\propto \mathbf{U}_{W_3}\mathbf{U}_{W_2}^\top\mathbf{U}_{W_2}\mathbf{U}_{W_3}^\top \propto \mathbf{I}_K, \\
\overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* &\propto \mathbf{U}_{W_3}\mathbf{V}_{W_1}^\top\mathbf{V}_{W_1}\mathbf{U}_{W_3}^\top \propto \mathbf{I}_K, \\
(\mathbf{W}_3^*\mathbf{W}_2^*)(\mathbf{W}_3^*\mathbf{W}_2^*)^\top &\propto (\mathbf{U}_{W_3}\mathbf{U}_{W_2}^T\mathbf{U}_{W_2}\mathbf{U}_{W_1}^\top)(\mathbf{U}_{W_3}\mathbf{U}_{W_2}^T\mathbf{U}_{W_2}\mathbf{U}_{W_1}^\top)^\top \propto \mathbf{I}_K, \\
(\mathbf{W}_1^*\overline{\mathbf{H}}^*)^\top(\mathbf{W}_1^*\overline{\mathbf{H}}^*) &\propto (\mathbf{U}_{W_1}\mathbf{V}_{W_1}^\top\mathbf{V}_{W_1}\mathbf{U}_{W_3}^\top)^\top(\mathbf{U}_{W_1}\mathbf{V}_{W_1}^\top\mathbf{V}_{W_1}\mathbf{U}_{W_3}^\top) \propto \mathbf{I}_K, \\
(\mathbf{W}_3^*\mathbf{W}_2^*\mathbf{W}_1^*)(\mathbf{W}_3^*\mathbf{W}_2^*\mathbf{W}_1^*)^\top &\propto (\mathbf{U}_{W_3}\mathbf{V}_{W_1}^\top)(\mathbf{U}_{W_3}\mathbf{V}_{W_1}^\top)^\top \propto \mathbf{I}_K, \\
(\mathbf{W}_2^*\mathbf{W}_1^*\overline{\mathbf{H}}^*)^\top(\mathbf{W}_2^*\mathbf{W}_1^*\overline{\mathbf{H}}^*) &\propto (\mathbf{U}_{W_2}\mathbf{U}_{W_3}^\top)^\top(\mathbf{U}_{W_2}\mathbf{U}_{W_3}^\top) \propto \mathbf{I}_K,
\end{aligned} \tag{25}$$

and,

$$\mathbf{W}_3^*\mathbf{W}_2^*\mathbf{W}_1^*\overline{\mathbf{H}}^* \propto \mathbf{U}_{W_3}\mathbf{U}_{W_2}^\top\mathbf{U}_{W_2}\mathbf{V}_{W_2}^\top\mathbf{V}_{W_2}\mathbf{V}_{W_1}^\top\mathbf{V}_{W_1}\mathbf{U}_{W_3}^\top \propto \mathbf{I}_K. \tag{26}$$

Next, we can derive the alignments between weights and features as following:

$$\begin{aligned}
\mathbf{W}_3^*\mathbf{W}_2^*\mathbf{W}_1^* &\propto \mathbf{U}_{W_3}\mathbf{V}_{W_1}^\top \propto \overline{\mathbf{H}}^{*\top}, \\
\mathbf{W}_2^*\mathbf{W}_1^*\overline{\mathbf{H}}^* &\propto \mathbf{U}_{W_2}\mathbf{U}_{W_3}^\top \propto \mathbf{W}_3^{*\top}, \\
\mathbf{W}_3^*\mathbf{W}_2^* &\propto \mathbf{U}_{W_3}\mathbf{V}_{W_2}^\top \propto (\mathbf{W}_1^*\overline{\mathbf{H}}^*)^\top.
\end{aligned} \tag{27}$$

- If $b = 3K\sqrt[3]{Kn\lambda_{W_3}\lambda_{W_2}\lambda_{W_1}\lambda_{H_1}} = \frac{\sqrt[3]{4}}{3}$: For this case, x_k^* can either be 0 or $\sqrt[3]{2}$, as long as $\{x_k^*\}_{k=1}^K$ is a decreasing sequence. If all the singular values are 0's, we have the trivial global minima $(\mathbf{W}_3^*, \mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^*) = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0})$. If there are exactly $r \leq K$ positive singular values $s_1 = s_2 = \ldots = s_r := s > 0$ and $s_{r+1} = \ldots = s_K = 0$, then we can write the compact SVD form of weight matrices and $\mathbf{H}_1^*$ as following:

$$\begin{aligned}
\mathbf{W}_3^* &= \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_3}}} s\mathbf{U}_{W_3}\mathbf{U}_{W_2}^T, \\
\mathbf{W}_2^* &= \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_2}}} s\mathbf{U}_{W_2}\mathbf{U}_{W_1}^\top, \\
\mathbf{W}_1^* &= s\mathbf{U}_{W_1}\mathbf{V}_{W_1}^\top, \\
\mathbf{H}_1^* &= \frac{\sqrt{c}s^3}{cs^6 + N\lambda_{H_1}}\mathbf{V}_{W_1}\mathbf{U}_{W_3}^\top\mathbf{Y} = \overline{\mathbf{H}}^*\mathbf{Y},
\end{aligned}$$

where $\mathbf{U}_{W_3}, \mathbf{U}_{W_2}, \mathbf{U}_{W_1}, \mathbf{V}_{W_1}$ are semi-orthonormal matrices consist r orthogonal columns. Additionally, we note that $\mathbf{U}_{W_3} \in \mathbb{R}^{K\times r}$ are created from orthonormal matrices size $K \times K$ with the removal of columns corresponding with singular values equal 0. Thus, $\mathbf{U}_{W_3}\mathbf{U}_{W_3}^\top$ is the best rank-r approximation of $\mathbf{I}_K$. From here, we can deduce the geometry of the following:

$$\begin{aligned}
\mathbf{W}_3^*\mathbf{W}_3^{*\top} &\propto \overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* \propto \mathbf{W}_3^*\mathbf{W}_2^*\mathbf{W}_1^*\overline{\mathbf{H}}^* \\
&\propto (\mathbf{W}_3^*\mathbf{W}_2^*)(\mathbf{W}_3^*\mathbf{W}_2^*)^\top \propto (\mathbf{W}_1^*\overline{\mathbf{H}})^\top(\mathbf{W}_1^*\overline{\mathbf{H}}) \\
&\propto (\mathbf{W}_3^*\mathbf{W}_2^*\mathbf{W}_1^*)(\mathbf{W}_3^*\mathbf{W}_2^*\mathbf{W}_1^*)^\top \propto (\mathbf{W}_2^*\mathbf{W}_1^*\overline{\mathbf{H}})^\top(\mathbf{W}_2^*\mathbf{W}_1^*\overline{\mathbf{H}}) \propto \mathcal{P}_r(\mathbf{I}_K),
\end{aligned}$$

where $\mathcal{P}_r(\mathbf{I}_K)$ denotes the best rank-r approximation of $\mathbf{I}_K$. The collapse of features ($\mathcal{NC}1$) and the alignments between weights and features ($\mathcal{NC}3$) are identical as the case $b < \frac{\sqrt[3]{4}}{3}$.

□

D.2. Supporting Lemmas for UFM Deep Linear Networks with M Layers of Weights

Before deriving the proof for M layers linear network, from the proof of three layers of weights, we generalize some useful results that support the main proof.

Consider MSE loss function with M layers linear network and arbitrary target matrix $\mathbf{Y} \in \mathbb{R}^{K\times N}$:

$$\begin{aligned}
f(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1) = \frac{1}{2N}\|\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1 - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_M}}{2}\|\mathbf{W}_M\|_F^2 \\
+\frac{\lambda_{W_{M-1}}}{2}\|\mathbf{W}_{M-1}\|_F^2 + \ldots + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2,
\end{aligned} \tag{28}$$

with $\mathbf{W}_M \in \mathbb{R}^{K\times d_M}$, $\mathbf{W}_{M-1} \in \mathbb{R}^{d_M\times d_{M-1}}$, $\mathbf{W}_{M-2} \in \mathbb{R}^{d_{M-1}\times d_{M-2}}, \ldots, \mathbf{W}_2 \in \mathbb{R}^{d_3\times d_2}$, $\mathbf{W}_1 \in \mathbb{R}^{d_2\times d_1}$, $\mathbf{H}_1 \in \mathbb{R}^{d_1\times K}$ with $d_M, d_{M-1}, \ldots, d_2, d_1$ are arbitrary positive integers.

Lemma D.1. *The partial derivative of* $\|\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1-\mathbf{Y}\|_F^2$ *w.r.t* $\mathbf{W}_i$ $(i=1,2,\ldots,M)$*:*

$$\frac{1}{2}\frac{\partial\|\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_i\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1-\mathbf{Y}\|_F^2}{\partial\mathbf{W}_i}=$$
$$\mathbf{W}_{i+1}^\top\mathbf{W}_{i+2}^\top\ldots\mathbf{W}_M^\top(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_i\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1-\mathbf{Y})\mathbf{H}_1^\top\mathbf{W}_1^\top\ldots\mathbf{W}_{i-1}^\top.$$

This result is common and the proof can be found in (Yun et al., 2017), for example.

Lemma D.2. *For any critical point* $(\mathbf{W}_M,\mathbf{W}_{M-1},\ldots,\mathbf{W}_2,\mathbf{W}_1,\mathbf{H}_1)$ *of* f*, we have the following:*

$$\begin{aligned}
\lambda_{W_M}\mathbf{W}_M^\top\mathbf{W}_M&=\lambda_{W_{M-1}}\mathbf{W}_{M-1}\mathbf{W}_{M-1}^\top,\\
\lambda_{W_{M-1}}\mathbf{W}_{M-1}^\top\mathbf{W}_{M-1}&=\lambda_{W_{M-2}}\mathbf{W}_{M-2}\mathbf{W}_{M-2}^\top,\\
&\ldots,\\
\lambda_{W_2}\mathbf{W}_2^\top\mathbf{W}_2&=\lambda_{W_1}\mathbf{W}_1\mathbf{W}_1^\top,\\
\lambda_{W_1}\mathbf{W}_1^\top\mathbf{W}_1&=\lambda_{H_1}\mathbf{H}_1\mathbf{H}_1^\top,
\end{aligned}$$

and:

$$\mathbf{H}_1=(c(\mathbf{W}_1^\top\mathbf{W}_1)^M+N\lambda_{H_1}\mathbf{I})^{-1}\mathbf{W}_1^\top\mathbf{W}_2^\top\ldots\mathbf{W}_M^\top\mathbf{Y},\tag{29}$$

with $c:=\frac{\lambda_{W_1}^{M-1}}{\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_2}}$.

Proof of Lemma D.2. By definition and using Lemma D.1, any critical point $(\mathbf{W}_M,\mathbf{W}_{M-1},\ldots,\mathbf{W}_2,\mathbf{W}_1,\mathbf{H}_1)$ satisfies the following :

$$\begin{aligned}
&\frac{\partial f}{\partial\mathbf{W}_M}=\frac{1}{N}(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1-\mathbf{Y})\mathbf{H}_1^\top\mathbf{W}_1^\top\ldots\mathbf{W}_{M-1}^\top+\lambda_{W_M}\mathbf{W}_M=\mathbf{0},\\
&\frac{\partial f}{\partial\mathbf{W}_{M-1}}=\frac{1}{N}\mathbf{W}_M^\top(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1-\mathbf{Y})\mathbf{H}_1^\top\mathbf{W}_1^\top\ldots\mathbf{W}_{M-2}^\top+\lambda_{W_{M-1}}\mathbf{W}_{M-1}=\mathbf{0},\\
&\ldots,\\
&\frac{\partial f}{\partial\mathbf{W}_1}=\frac{1}{N}\mathbf{W}_2^\top\mathbf{W}_3^\top\ldots\mathbf{W}_M^\top(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1-\mathbf{Y})\mathbf{H}_1^\top+\lambda_{W_1}\mathbf{W}_1=\mathbf{0},\\
&\frac{\partial f}{\partial\mathbf{H}_1}=\frac{1}{N}\mathbf{W}_1^\top\mathbf{W}_2^\top\ldots\mathbf{W}_M^\top(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1-\mathbf{Y})+\lambda_{H_1}\mathbf{H}_1=\mathbf{0}.
\end{aligned}$$

Next, we have:

$$\begin{aligned}
&\mathbf{0}=\mathbf{W}_M^\top\frac{\partial f}{\partial\mathbf{W}_M}-\frac{\partial f}{\partial\mathbf{W}_{M-1}}\mathbf{W}_{M-1}^\top=\lambda_{W_M}\mathbf{W}_M^\top\mathbf{W}_M-\lambda_{W_{M-1}}\mathbf{W}_{M-1}\mathbf{W}_{M-1}^\top\\
&\Rightarrow\lambda_{W_M}\mathbf{W}_M^\top\mathbf{W}_M=\lambda_{W_{M-1}}\mathbf{W}_{M-1}\mathbf{W}_{M-1}^\top.\\
&\mathbf{0}=\mathbf{W}_{M-1}^\top\frac{\partial f}{\partial\mathbf{W}_{M-1}}-\frac{\partial f}{\partial\mathbf{W}_{M-2}}\mathbf{W}_{M-2}^\top=\lambda_{W_{M-1}}\mathbf{W}_{M-1}^\top\mathbf{W}_{M-1}-\lambda_{W_{M-2}}\mathbf{W}_{M-2}\mathbf{W}_{M-2}^\top\\
&\Rightarrow\lambda_{W_{M-1}}\mathbf{W}_{M-1}^\top\mathbf{W}_{M-1}=\lambda_{W_{M-2}}\mathbf{W}_{M-2}\mathbf{W}_{M-2}^\top.
\end{aligned}$$

Making similar argument for the other derivatives, we have:

$$\begin{aligned}
\lambda_{W_M}\mathbf{W}_M^\top\mathbf{W}_M&=\lambda_{W_{M-1}}\mathbf{W}_{M-1}\mathbf{W}_{M-1}^\top,\\
\lambda_{W_{M-1}}\mathbf{W}_{M-1}^\top\mathbf{W}_{M-1}&=\lambda_{W_{M-2}}\mathbf{W}_{M-2}\mathbf{W}_{M-2}^\top,\\
&\ldots,\\
\lambda_{W_2}\mathbf{W}_2^\top\mathbf{W}_2&=\lambda_{W_1}\mathbf{W}_1\mathbf{W}_1^\top,\\
\lambda_{W_1}\mathbf{W}_1^\top\mathbf{W}_1&=\lambda_{H_1}\mathbf{H}_1\mathbf{H}_1^\top.
\end{aligned}$$

Also, from $\frac{\partial f}{\partial \mathbf{H}_1} = \mathbf{0}$, solving for $\mathbf{H}_1$ yields:

$$\begin{aligned}
\mathbf{H}_1 &= (\mathbf{W}_1^\top \mathbf{W}_2^\top \ldots \mathbf{W}_{M-1}^\top \mathbf{W}_M^\top \mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 + N\lambda_{H_1}\mathbf{I})^{-1} \mathbf{W}_1^\top \mathbf{W}_2^\top \ldots \mathbf{W}_M^\top \mathbf{Y} \\
&= \left(\frac{\lambda_{W_{M-1}}}{\lambda_{W_M}} \mathbf{W}_1^\top \mathbf{W}_2^\top \ldots (\mathbf{W}_{M-1}^\top \mathbf{W}_{M-1})^2 \ldots \mathbf{W}_2 \mathbf{W}_1 + N\lambda_{H_1}\mathbf{I} \right)^{-1} \mathbf{W}_1^\top \mathbf{W}_2^\top \ldots \mathbf{W}_M^\top \mathbf{Y} \\
&= \ldots \\
&= \left(\underbrace{\frac{\lambda_{W_1}^{M-1}}{\lambda_{W_M}\lambda_{W_{M-1}} \ldots \lambda_{W_2}}}_{c} (\mathbf{W}_1^\top \mathbf{W}_1)^M + N\lambda_{H_1} \right)^{-1} \mathbf{W}_1^\top \mathbf{W}_2^\top \ldots \mathbf{W}_M^\top \mathbf{Y} \\
&= (c(\mathbf{W}_1^\top \mathbf{W}_1)^M + N\lambda_{H_1}\mathbf{I})^{-1} \mathbf{W}_1^\top \mathbf{W}_2^\top \ldots \mathbf{W}_M^\top \mathbf{Y}.
\end{aligned}$$

□

Lemma D.3. *For any critical point* $(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1)$*, we have* $r := \operatorname{rank}(\mathbf{W}_M) = \operatorname{rank}(\mathbf{W}_{M-1}) = \operatorname{rank}(\mathbf{W}_{M-2}) = \ldots = \operatorname{rank}(\mathbf{W}_1) = \operatorname{rank}(\mathbf{H}_1) \leq \min(K, d_M, d_{M-1}, \ldots, d_1) := R$*.*

Proof of Lemma D.3. The result is deduced from Lemma D.2 and the matrix rank property $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}^\top \mathbf{A}) = \operatorname{rank}(\mathbf{A}\mathbf{A}^\top)$. □

Lemma D.4. *For any critical point* $(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1)$ *of* f*, let* $\mathbf{W}_1 = \mathbf{U}_{W_1}\mathbf{S}_{W_1}\mathbf{V}_{W_1}^\top$ *be the SVD decomposition of* $\mathbf{W}_1$ *with* $\mathbf{U}_{W_1} \in \mathbb{R}^{d_2 \times d_2}, \mathbf{V}_{W_1} \in \mathbb{R}^{d_1 \times d_1}$ *are orthonormal matrices and* $\mathbf{S}_{W_1} \in \mathbb{R}^{d_2 \times d_1}$ *is a diagonal matrix with* ***decreasing*** *non-negative singular values. We denote the* $r := \operatorname{rank}(\mathbf{W}_1)$ *singular values of* $\mathbf{W}_1$ *as* $\{s_k\}_{k=1}^r$ *(*$r \leq R := \min(K, d_M, \ldots, d_1)$*, from Lemma D.3).*

Then, we can write the SVD of weight matrices as:

$$\begin{aligned}
\mathbf{W}_M &= \mathbf{U}_{W_M}\mathbf{S}_{W_M}\mathbf{U}_{W_{M-1}}^\top, \\
\mathbf{W}_{M-1} &= \mathbf{U}_{W_{M-1}}\mathbf{S}_{W_{M-1}}\mathbf{U}_{W_{M-2}}^\top, \\
\mathbf{W}_{M-2} &= \mathbf{U}_{W_{M-2}}\mathbf{S}_{W_{M-2}}\mathbf{U}_{W_{M-3}}^\top, \\
\mathbf{W}_{M-3} &= \mathbf{U}_{W_{M-3}}\mathbf{S}_{W_{M-3}}\mathbf{U}_{W_{M-4}}^\top, \\
&\ldots, \\
\mathbf{W}_2 &= \mathbf{U}_{W_2}\mathbf{S}_{W_2}\mathbf{U}_{W_1}^\top, \\
\mathbf{W}_1 &= \mathbf{U}_{W_1}\mathbf{S}_{W_1}\mathbf{V}_{W_1}^\top,
\end{aligned}$$

with:

$$\mathbf{S}_{W_j} = \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_j}}} \begin{bmatrix} \operatorname{diag}(s_1, \ldots, s_r) & \mathbf{0}_{r \times (d_j - r)} \\ \mathbf{0}_{(d_{j+1}-r) \times r} & \mathbf{0}_{(d_{j+1}-r) \times (d_j - r)} \end{bmatrix} \in \mathbb{R}^{d_{j+1} \times d_j}, \quad \forall\, j \in [M],$$

and $\mathbf{U}_{W_M}, \mathbf{U}_{W_{M-1}}, \mathbf{U}_{W_{M-2}}, \mathbf{U}_{W_{M-3}}, \ldots, \mathbf{U}_{W_1}, \mathbf{V}_{W_1}$ *are all orthonormal matrices.*

Proof of Lemma D.4. From Lemma D.2, we have:

$$\mathbf{W}_2^\top \mathbf{W}_2 = \frac{\lambda_{W_1}}{\lambda_{W_2}} \mathbf{W}_1 \mathbf{W}_1^\top = \frac{\lambda_{W_1}}{\lambda_{W_2}} \mathbf{U}_{W_1}\mathbf{S}_{W_1}\mathbf{S}_{W_1}^\top \mathbf{U}_{W_1}^\top = \mathbf{U}_{W_1}\mathbf{S}_{W_2}^\top \mathbf{S}_{W_2}\mathbf{U}_{W_1}^\top,$$

where:

$$\mathbf{S}_{W_2} := \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_2}}} \begin{bmatrix} \operatorname{diag}(s_1, \ldots, s_r) & \mathbf{0}_{r \times (d_2 - r)} \\ \mathbf{0}_{(d_3-r) \times r} & \mathbf{0}_{(d_3-r) \times (d_2 - r)} \end{bmatrix} \in \mathbb{R}^{d_3 \times d_2}.$$

This means the diagonal matrix $\mathbf{S}_{W_2}^\top \mathbf{S}_{W_2}$ contains the eigenvalues and the columns of $\mathbf{U}_{W_1}$ are the eigenvectors of $\mathbf{W}_2^\top \mathbf{W}_2$. Hence, we can write the SVD decomposition of $\mathbf{W}_2$ as $\mathbf{W}_2 = \mathbf{U}_{W_2}\mathbf{S}_{W_2}\mathbf{U}_{W_1}^\top$ with orthonormal matrix $\mathbf{U}_{W_2} \in \mathbb{R}^{d_3 \times d_3}$.

By making similar arguments as above for $\mathbf{W}_3$, from:

$$\mathbf{W}_3^\top \mathbf{W}_3 = \frac{\lambda_{W_2}}{\lambda_{W_3}} \mathbf{W}_2 \mathbf{W}_2^\top = \frac{\lambda_{W_2}}{\lambda_{W_3}} \mathbf{U}_{W_2}\mathbf{S}_{W_2}\mathbf{S}_{W_2}^\top \mathbf{U}_{W_2}^\top = \mathbf{U}_{W_2}\mathbf{S}_{W_3}^\top \mathbf{S}_{W_3}\mathbf{U}_{W_2}^\top,$$

where:

$$\mathbf{S}_{W_3} := \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_3}}} \begin{bmatrix} \operatorname{diag}(s_1, \ldots, s_r) & \mathbf{0}_{r \times (d_3 - r)} \\ \mathbf{0}_{(d_4 - r) \times r} & \mathbf{0}_{(d_4 - r) \times (d_3 - r)} \end{bmatrix} \in \mathbb{R}^{d_4 \times d_3},$$

and thus, we can write SVD decomposition of $\mathbf{W}_3$ as $\mathbf{W}_3 = \mathbf{U}_{W_3}\mathbf{S}_{W_3}\mathbf{U}_{W_2}^\top$ with orthonormal matrix $\mathbf{U}_{W_3} \in \mathbb{R}^{d_4 \times d_4}$. Repeating the process for other weight matrices, we got the desired result. □

Lemma D.5. *Continue from the setting and result of Lemma D.4, we have:*

$$\mathbf{H}_1 = \mathbf{V}_{W_1} \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{\sqrt{c} s_1^M}{c s_1^{2M} + N\lambda_{H_1}}, \ldots, \frac{\sqrt{c} s_r^M}{c s_r^{2M} + N\lambda_{H_1}} \right) & \mathbf{0}_{r \times (K-r)} \\ \mathbf{0}_{(d_1 - r) \times r} & \mathbf{0}_{(d_1 - r) \times (K - r)} \end{bmatrix}}_{\mathbf{C} \in \mathbb{R}^{d_1 \times K}} \mathbf{U}_{W_M}^\top \mathbf{Y},$$

$$\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H} - \mathbf{Y} = \mathbf{U}_{W_M} \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{-N\lambda_{H_1}}{c s_1^{2M} + N\lambda_{H_1}}, \ldots, \frac{-N\lambda_{H_1}}{c s_r^{2M} + N\lambda_{H_1}} \right) & \mathbf{0}_{r \times (K-r)} \\ \mathbf{0}_{(K - r) \times r} & -\mathbf{I}_{K-r} \end{bmatrix}}_{\mathbf{D} \in \mathbb{R}^{K \times K}} \mathbf{U}_{W_M}^\top \mathbf{Y},$$

with $c := \frac{\lambda_{W_1}^{M-1}}{\lambda_{W_M} \lambda_{W_{M-1}} \ldots \lambda_{W_2}}$.

Proof of Lemma D.5. From Lemma D.2, together with the SVD of weight matrices and the form of singular matrix $\mathbf{S}_{W_j}$ derived in Lemma D.4, we have:

$$\begin{aligned}
\mathbf{H}_1 &= (c(\mathbf{W}_1^\top \mathbf{W}_1)^M + N\lambda_{H_1}\mathbf{I})^{-1} \mathbf{W}_1^\top \mathbf{W}_2^\top \ldots \mathbf{W}_M^\top \mathbf{Y} \\
&= (c\mathbf{V}_{W_1}(\mathbf{S}_{W_1}^\top \mathbf{S}_{W_1})^M \mathbf{V}_{W_1}^\top + N\lambda_{H_1}\mathbf{I})^{-1} \mathbf{V}_{W_1}\mathbf{S}_{W_1}^\top \mathbf{S}_{W_2}^\top \ldots \mathbf{S}_{W_M}^\top \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \mathbf{V}_{W_1}(c(\mathbf{S}_{W_1}^\top \mathbf{S}_{W_1})^M + N\lambda_{H_1}\mathbf{I})^{-1} \mathbf{S}_{W_1}^\top \mathbf{S}_{W_2}^\top \ldots \mathbf{S}_{W_M}^\top \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \mathbf{V}_{W_1}(c(\mathbf{S}_{W_1}^\top \mathbf{S}_{W_1})^M + N\lambda_{H_1}\mathbf{I})^{-1} \sqrt{c} \begin{bmatrix} \operatorname{diag}(s_1^M, \ldots, s_r^M) & \mathbf{0}_{r \times (K-r)} \\ \mathbf{0}_{(d_1 - r) \times r} & \mathbf{0}_{(d_1 - r) \times (K-r)} \end{bmatrix} \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \mathbf{V}_{W_1} \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{\sqrt{c} s_1^M}{c s_1^{2M} + N\lambda_{H_1}}, \ldots, \frac{\sqrt{c} s_r^M}{c s_r^{2M} + N\lambda_{H_1}} \right) & \mathbf{0}_{r \times (K-r)} \\ \mathbf{0}_{(d_1 - r) \times r} & \mathbf{0}_{(d_1 - r) \times (K - r)} \end{bmatrix}}_{\mathbf{C} \in \mathbb{R}^{d_1 \times K}} \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \mathbf{V}_{W_1} \mathbf{C} \mathbf{U}_{W_M}^\top \mathbf{Y}
\end{aligned}$$

$$\begin{aligned}
&\Rightarrow \mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1 = \mathbf{U}_{W_M} \mathbf{S}_{W_M} \mathbf{S}_{W_{M-1}} \ldots \mathbf{S}_{W_1} \mathbf{C} \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_M}}} \mathbf{U}_{W_M} \begin{bmatrix} \operatorname{diag}(s_1, \ldots, s_r) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{S}_{W_{M-1}} \ldots \mathbf{S}_{W_1} \mathbf{C} \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \ldots \\
&= \mathbf{U}_{W_M} \sqrt{c} \begin{bmatrix} \operatorname{diag}(s_1^M, \ldots, s_r^M) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{C} \mathbf{U}_{W_M}^\top \mathbf{Y}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{U}_{W_M} \begin{bmatrix} \operatorname{diag}\left(\frac{cs_1^{2M}}{cs_1^{2M}+N\lambda_{H_1}}, \ldots, \frac{cs_r^{2M}}{cs_r^{2M}+N\lambda_{H_1}}\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&\Rightarrow \mathbf{W}_M \ldots \mathbf{W}_1 \mathbf{H}_1 - \mathbf{Y} = \mathbf{U}_{W_M} \left(\begin{bmatrix} \operatorname{diag}\left(\frac{cs_1^{2M}}{cs_1^{2M}+N\lambda_{H_1}}, \ldots, \frac{cs_r^{2M}}{cs_r^{2M}+N\lambda_{H_1}}\right) & \mathbf{0}_{r\times(K-r)} \\ \mathbf{0}_{(K-r)\times r} & \mathbf{0}_{(K-r)\times(K-r)} \end{bmatrix} - \mathbf{I}_K \right) \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \mathbf{U}_{W_M} \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{-N\lambda_{H_1}}{cs_1^{2M}+N\lambda_{H_1}}, \ldots, \frac{-N\lambda_{H_1}}{cs_r^{2M}+N\lambda_{H_1}}\right) & \mathbf{0}_{r\times(K-r)} \\ \mathbf{0}_{(K-r)\times r} & -\mathbf{I}_{K-r} \end{bmatrix}}_{\mathbf{D}\in\mathbb{R}^{K\times K}} \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top \mathbf{Y}.
\end{aligned}
$$

□

D.2.1. Minimizer of the function $g(x) = \frac{1}{x^M+1} + bx$

Next, we study the minimization problem of the following function, this result will be used frequently in proofs of theorems in the main paper:

$$
g(x) = \frac{1}{x^M+1} + bx \text{ with } x \geq 0, b > 0, M \geq 2.
$$

Clearly, $g(0) = 1$. We consider the following cases for parameter b:

- If $b > \frac{(M-1)^{\frac{M-1}{M}}}{M}$: We have with $x > 0$: $g(x) > \frac{1}{x^M+1} + \frac{(M-1)^{\frac{M-1}{M}}}{M} x$. We will prove:

$$
\begin{aligned}
&\frac{1}{x^M+1} + \frac{(M-1)^{\frac{M-1}{M}}}{M} x \geq 1 \\
&\Leftrightarrow \frac{(M-1)^{\frac{M-1}{M}}}{M} x^{M+1} - x^M + \frac{(M-1)^{\frac{M-1}{M}}}{M} x \geq 0 \\
&\Leftrightarrow x(x^M - \frac{M}{(M-1)^{\frac{M-1}{M}}} x^{M-1} + 1) \geq 0 \\
&\Leftrightarrow x^M - \frac{M}{(M-1)^{\frac{M-1}{M}}} x^{M-1} + 1 \geq 0.
\end{aligned}
\tag{30}
$$

  Let $h(x) = x^M - \frac{M}{(M-1)^{\frac{M-1}{M}}} x^{M-1} + 1$ with $x \geq 0$, we have:

$$
\begin{aligned}
h'(x) &= Mx^{M-1} - M(M-1)^{1/M} x^{M-2}, \\
h'(x) &= 0 \Leftrightarrow x = 0 \text{ or } x = (M-1)^{1/M}.
\end{aligned}
\tag{31}
$$

  We also have: $h(0) = 1$ and $h((M-1)^{1/M}) = M - 1 - M + 1 = 0$. From the variation table, we clearly have $h(x) \geq 0 \, \forall \, x \geq 0$.

x	0	$(M-1)^{1/M}$	∞
$h'(x)$	-	0	+
$h(x)$	1	0	∞

  Hence, in this case, $g(x) > 1 \, \forall \, x > 0$, therefore, $g(x)$ is minimized at $x = 0$.

- If $b = \frac{(M-1)^{\frac{M-1}{M}}}{M}$: We have $g(x) = \frac{1}{x^M+1} + \frac{(M-1)^{\frac{M-1}{M}}}{M} x \geq 1$. Thus, $g(x)$ is minimized at $x = 0$ or $x = (M-1)^{1/M}$.

- If $b < \frac{(M-1)^{\frac{M-1}{M}}}{M}$: We take the first and second derivatives of $g(x)$:

$$
\begin{aligned}
g'(x) &= b - \frac{Mx^{M-1}}{(x^M+1)^2}, \\
g''(x) &= -M\left(\frac{(M-1)x^{M-2}}{(x^M+1)^2} - \frac{2Mx^{2M-2}}{(x^M+1)^3}\right). \\
&= \frac{(M^2+M)x^{2M-2} - (M^2-M)x^{M-2}}{(x^M+1)^3}
\end{aligned}
$$

We have: $g''(x) = 0 \Leftrightarrow x = 0$ or $x = \sqrt[M]{\frac{M-1}{M+1}}$. Therefore, with $x \geq 0$, $g'(x) = 0$ has at most 2 solutions. We further have $g'(\sqrt[M]{\frac{M-1}{M+1}}) = b - M(\frac{M-1}{M+1})^{\frac{M-1}{M}}/(\frac{M-1}{M+1}+1)^2 < (M-1)^{\frac{M-1}{M}}/M - M(\frac{M-1}{M+1})^{\frac{M-1}{M}}/(\frac{M-1}{M+1}+1)^2$. Actually, we have:

$$
\begin{aligned}
&\frac{(M-1)^{\frac{M-1}{M}}}{M} < \frac{M(\frac{M-1}{M+1})^{\frac{M-1}{M}}}{(\frac{M-1}{M+1}+1)^2} \\
\Leftrightarrow &\left(\frac{M-1}{M+1}+1\right)^2 < \frac{M^2}{(M+1)^{\frac{M-1}{M}}} \\
\Leftrightarrow &\frac{4M^2}{(M+1)^2} < \frac{M^2}{(M+1)^{\frac{M-1}{M}}} \\
\Leftrightarrow &\, 4 < (M+1)^{2-\frac{M-1}{M}} \\
\Leftrightarrow &\, 4 < (M+1)^{1+\frac{1}{M}} \quad \text{(true } \forall M \geq 2).
\end{aligned}
$$

Therefore, $g'(\sqrt[M]{\frac{M-1}{M+1}}) < 0$. Together with the fact that $g'(0) = b > 0$ and $g'(+\infty) > 0$, $g'(x) = 0$ has exactly two solutions, we call it x_1 and x_2 ($x_1 < \sqrt[M]{\frac{M-1}{M+1}} < x_2$). Next, we note that $g'(x_2) = 0$ and $g'(x) > 0 \quad \forall x > x_2$ (since $g''(x) > 0 \quad \forall x > x_2$). In the meanwhile, $g'(\sqrt[M]{M-1}) = b - \frac{M(M-1)^{\frac{M-1}{M}}}{M^2} = b - \frac{(M-1)^{\frac{M-1}{M}}}{M} < 0$. Hence, we must have $x_2 > \sqrt[M]{M-1}$.

x	0	x_1	$\sqrt[M]{\frac{M-1}{M+1}}$	$\sqrt[M]{M-1}$	x_2	$+\infty$
$g''(x)$	0	-	0	+	+	+
$g'(x)$	+	0	-	-	0	+
$g(x)$	1	$g(x_1)$	$g(\sqrt[M]{\frac{M-1}{M+1}})$	$\frac{1}{M} + b\sqrt[M]{M-1}$	$g(x_2)$	$+\infty$

From the variation table, we can see that $g(x_2) < g(\sqrt[M]{M-1}) = \frac{1}{M} + b\sqrt[M]{M-1} < \frac{1}{M} + \frac{(M-1)^{\frac{M-1}{M}}}{M}\sqrt[M]{M-1} = \frac{1}{M} + \frac{M-1}{M} = 1 = g(0)$.

In conclusion, in this case, $g(x)$ is minimized at $x_2 > \sqrt[M]{M-1}$, i.e. the largest solution of the equation $b - \frac{Mx^{M-1}}{(x^M+1)^2} = 0$.

D.3. Full Proof of Theorem 3.1 with Bias-Free

Now, we state the proof of Theorem 3.1 for general setting with M layers of weight with no bias (i.e., excluding $\mathbf{b}$) with arbitrary widths $d_M, d_{M-1}, \ldots, d_1$.

Proof of Theorem 3.1 (bias-free). First, by using Lemma D.2, we have for any critical point

$(\mathbf{W}_M, \mathbf{W}_{M-1}, \dots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1)$ of f, we have the following:

$$\begin{aligned}
\lambda_{W_M} \mathbf{W}_M^\top \mathbf{W}_M &= \lambda_{W_{M-1}} \mathbf{W}_{M-1} \mathbf{W}_{M-1}^\top, \\
\lambda_{W_{M-1}} \mathbf{W}_{M-1}^\top \mathbf{W}_{M-1} &= \lambda_{W_{M-2}} \mathbf{W}_{M-2} \mathbf{W}_{M-2}^\top, \\
&\dots, \\
\lambda_{W_2} \mathbf{W}_2^\top \mathbf{W}_2 &= \lambda_{W_1} \mathbf{W}_1 \mathbf{W}_1^\top, \\
\lambda_{W_1} \mathbf{W}_1^\top \mathbf{W}_1 &= \lambda_{H_1} \mathbf{H}_1 \mathbf{H}_1^\top.
\end{aligned}$$

Let $\mathbf{W}_1 = \mathbf{U}_{W_1} \mathbf{S}_{W_1} \mathbf{V}_{W_1}^\top$ be the SVD decomposition of $\mathbf{W}_1$ with $\mathbf{U}_{W_1} \in \mathbb{R}^{d_2 \times d_2}, \mathbf{V}_{W_1} \in \mathbb{R}^{d_1 \times d_1}$ are orthonormal matrices and $\mathbf{S}_{W_1} \in \mathbb{R}^{d_2 \times d_1}$ is a diagonal matrix with **decreasing** non-negative singular values. We denote the r singular values of $\mathbf{W}_1$ as $\{s_k\}_{k=1}^r$ ($r \leq R := \min(K, d_M, \dots, d_1)$, from Lemma D.3). From Lemma D.4, we have the SVD of other weight matrices as:

$$\begin{aligned}
\mathbf{W}_M &= \mathbf{U}_{W_M} \mathbf{S}_{W_M} \mathbf{U}_{W_{M-1}}^\top, \\
\mathbf{W}_{M-1} &= \mathbf{U}_{W_{M-1}} \mathbf{S}_{W_{M-1}} \mathbf{U}_{W_{M-2}}^\top, \\
\mathbf{W}_{M-2} &= \mathbf{U}_{W_{M-2}} \mathbf{S}_{W_{M-2}} \mathbf{U}_{W_{M-3}}^\top, \\
\mathbf{W}_{M-3} &= \mathbf{U}_{W_{M-3}} \mathbf{S}_{W_{M-3}} \mathbf{U}_{W_{M-4}}^\top, \\
&\dots \\
\mathbf{W}_2 &= \mathbf{U}_{W_2} \mathbf{S}_{W_2} \mathbf{U}_{W_1}^\top, \\
\mathbf{W}_1 &= \mathbf{U}_{W_1} \mathbf{S}_{W_1} \mathbf{V}_{W_1}^\top,
\end{aligned}$$

where:

$$\mathbf{S}_{W_j} = \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_j}}} \begin{bmatrix} \operatorname{diag}(s_1, \dots, s_r) & \mathbf{0}_{r \times (d_j - r)} \\ \mathbf{0}_{(d_{j+1} - r) \times r} & \mathbf{0}_{(d_{j+1} - r) \times (d_j - r)} \end{bmatrix} \in \mathbb{R}^{d_{j+1} \times d_j}, \quad \forall\, j \in [M],$$

and $\mathbf{U}_{W_M}, \mathbf{U}_{W_{M-1}}, \mathbf{U}_{W_{M-2}}, \mathbf{U}_{W_{M-3}}, \dots, \mathbf{U}_{W_1}, \mathbf{V}_{W_1}$ are all orthonormal matrices.

From Lemma D.5, denote $c := \frac{\lambda_{W_1}^{M-1}}{\lambda_{W_M} \lambda_{W_{M-1}} \dots \lambda_{W_2}}$, we have:

$$\begin{aligned}
\mathbf{H}_1 &= \mathbf{V}_{W_1} \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{\sqrt{c} s_1^M}{c s_1^{2M} + N \lambda_{H_1}}, \dots, \frac{\sqrt{c} s_r^M}{c s_r^{2M} + N \lambda_{H_1}} \right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\mathbf{C} \in \mathbb{R}^{d_1 \times K}} \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \mathbf{V}_{W_1} \mathbf{C} \mathbf{U}_{W_M}^\top \mathbf{Y},
\end{aligned} \tag{32}$$

$$\begin{aligned}
\mathbf{W}_M \mathbf{W}_{M-1} \dots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H} - \mathbf{Y} &= \mathbf{U}_{W_M} \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{-N \lambda_{H_1}}{c s_1^{2M} + N \lambda_{H_1}}, \dots, \frac{-N \lambda_{H_1}}{c s_r^{2M} + N \lambda_{H_1}} \right) & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_{K-r} \end{bmatrix}}_{\mathbf{D} \in \mathbb{R}^{K \times K}} \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top \mathbf{Y}.
\end{aligned} \tag{33}$$

Next, we will calculate the Frobenius norm of $\mathbf{W}_M \mathbf{W}_{M-1} \dots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H} - \mathbf{Y}$:

$$\begin{aligned}
\|\mathbf{W}_M \mathbf{W}_{M-1} \dots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1 - \mathbf{Y}\|_F^2 &= \|\mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top \mathbf{Y}\|_F^2 \\
&= \operatorname{trace}(\mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top \mathbf{Y} (\mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top \mathbf{Y})^\top) \\
&= \operatorname{trace}(\mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top) \\
&= \operatorname{trace}(\mathbf{D}^2 \mathbf{U}_{W_M}^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{U}_{W_M}) \\
&= n \operatorname{trace}(\mathbf{D}^2) = n \left[\sum_{k=1}^r \left(\frac{-N \lambda_{H_1}}{c s_1^{2M} + N \lambda_{H_1}} \right)^2 + K - r \right].
\end{aligned} \tag{34}$$

where we use the fact $\mathbf{Y}\mathbf{Y}^\top = (\mathbf{I}_K \otimes \mathbf{1}_n^\top)(\mathbf{I}_K \otimes \mathbf{1}_n^\top)^\top = n\mathbf{I}_K$ and $\mathbf{U}_{W_M}$ is an orthonormal matrix.

Similarly, for $\mathbf{H}_1$, we have:

$$\begin{aligned} \|\mathbf{H}_1\|_F^2 &= \operatorname{trace}(\mathbf{V}_{W_1}\mathbf{C}\mathbf{U}_{W_M}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_{W_M}\mathbf{C}^\top \mathbf{V}_{W_1}^\top) = \operatorname{trace}(\mathbf{C}^\top\mathbf{C}\mathbf{U}_{W_M}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_{W_M}) \\ &= n\sum_{k=1}^{r} \frac{cs_k^{2M}}{cs_k^{2M} + N\lambda_{H_1}}. \end{aligned} \tag{35}$$

Now, we plug equations (34), (35) and the SVD of weight matrices into the function f and note that orthonormal matrix does not change Frobenius norm, we got:

$$\begin{aligned} f(\mathbf{W}_M, \dots, \mathbf{W}_1, \mathbf{H}_1) &= \frac{1}{2N}\|\mathbf{W}_M\mathbf{W}_{M-1}\dots\mathbf{W}_2\mathbf{W}_1\mathbf{H} - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_M}}{2}\|\mathbf{W}_M\|_F^2 + \dots + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2 \\ &= \frac{1}{2K}\sum_{k=1}^{r}\frac{(-N\lambda_{H_1})^2}{(cs_k^{2M} + N\lambda_{H_1})^2} + \frac{K-r}{2K} + \frac{\lambda_{W_M}}{2}\sum_{k=1}^{r}\frac{\lambda_{W_1}}{\lambda_{W_M}}s_k^2 + \frac{\lambda_{W_{M-1}}}{2}\sum_{k=1}^{r}\frac{\lambda_{W_1}}{\lambda_{W_{M-1}}}s_k^2 \\ &+ \dots + \frac{\lambda_{W_1}}{2}\sum_{k=1}^{r}s_k^2 + \frac{n\lambda_{H_1}}{2}\sum_{k=1}^{r}\frac{cs_k^{2M}}{(cs_k^{2M} + N\lambda_{H_1})^2} \\ &= \frac{n\lambda_{H_1}}{2}\sum_{k=1}^{r}\frac{1}{cs_k^{2M} + N\lambda_{H_1}} + \frac{K-r}{2K} + \frac{M\lambda_{W_1}}{2}\sum_{k=1}^{r}s_k^2 \\ &= \frac{1}{2K}\sum_{k=1}^{r}\left(\frac{1}{\frac{cs_k^{2M}}{N\lambda_{H_1}} + 1} + MN\lambda_{W_1}\sqrt[M]{\frac{N\lambda_{H_1}}{c}}\left(\sqrt[M]{\frac{cs_k^{2M}}{N\lambda_{H_1}}}\right)\right) + \frac{K-r}{2K} \\ &= \frac{1}{2K}\sum_{k=1}^{r}\left(\frac{1}{x_k^M + 1} + bx_k\right) + \frac{K-r}{2K}, \end{aligned} \tag{36}$$

with $x_k := \sqrt[M]{\frac{cs_k^{2M}}{N\lambda_{H_1}}}$ and $b := MK\lambda_{W_1}\sqrt[M]{\frac{N\lambda_{H_1}}{c}} = MK\lambda_{W_1}\sqrt[M]{\frac{N\lambda_{W_M}\lambda_{W_{M-2}}\dots\lambda_{W_1}\lambda_{H_1}}{\lambda_{W_1}^{M-1}}} = MK\sqrt[M]{Kn\lambda_{W_M}\lambda_{W_{M-1}}\dots\lambda_{W_1}\lambda_{H_1}}$.

Recall that we have studied the minimizer of function $g(x) = \frac{1}{x^M+1} + bx$ in Section D.2.1. From equation (36), f can be written as $\frac{1}{2K}\sum_{k=1}^{r} g(x_k) + \frac{K-r}{2N}$. By applying the result from Section D.2.1 for each $g(x_k)$, we finish bounding f and the equality conditions are as following:

- If $b = MK\sqrt[M]{Kn\lambda_{W_M}\lambda_{W_{M-1}}\dots\lambda_{W_1}\lambda_{H_1}} > \frac{(M-1)^{\frac{M-1}{M}}}{M}$: all the singular values of $\mathbf{W}_1$ are zeros. Therefore, the singular values of $\mathbf{W}_M, \mathbf{W}_{M-1}, \dots, \mathbf{H}_1$ are also all zeros. In this case, $f(\mathbf{W}_M, \mathbf{W}_{M-1}, \dots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1)$ is minimized at $(\mathbf{W}_M^*, \mathbf{W}_{M-1}^*, \dots, \mathbf{W}_1^*, \mathbf{H}_1^*) = (\mathbf{0}, \mathbf{0}, \dots \mathbf{0}, \mathbf{0})$.

- If $b = MK\sqrt[M]{Kn\lambda_{W_M}\lambda_{W_{M-1}}\dots\lambda_{W_1}\lambda_{H_1}} < \frac{(M-1)^{\frac{M-1}{M}}}{M}$: In this case, $\mathbf{W}_1^*$ have r singular values, all of which are equal a multiplier of the largest positive solution of the equation $b - \frac{Mx^{M-1}}{(x^M+1)^2} = 0$, we denote that singular value as s. Hence, we can write the compact SVD form (with a bit of notation abuse) of $\mathbf{W}_{M-1}^*$ as $\mathbf{W}_1^* = s\mathbf{U}_{W_1}\mathbf{V}_{W_1}^\top$ with semi-orthonormal matrices $\mathbf{U}_{W_1} \in \mathbb{R}^{d_2\times r}, \mathbf{V}_{W_1} \in \mathbb{R}^{d_1\times r}$. (note that $\mathbf{U}_{W_1}^\top\mathbf{U}_{W_1} = \mathbf{I}$ and $\mathbf{V}_{W_1}^\top\mathbf{V}_{W_1} = \mathbf{I}$). Since $\frac{1}{x^{*M}+1} + bx^* < 1$, we have $r = R = \min(K, d_M, \dots, d_1)$ in this case.

Similarly, we also have the compact SVD form of other weight matrices and feature matrix as:

$$\begin{aligned}
\mathbf{W}_M^* &= \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_M}}} s \mathbf{U}_{W_M} \mathbf{U}_{W_{M-1}}^T, \\
\mathbf{W}_{M-1}^* &= \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_{M-1}}}} s \mathbf{U}_{W_{M-1}} \mathbf{U}_{W_{M-2}}^\top, \\
&\dots \\
\mathbf{W}_1^* &= s \mathbf{U}_{W_1} \mathbf{V}_{W_1}^\top, \\
\mathbf{H}_1^* &= \frac{\sqrt{c} s^M}{c s^{2M} + N\lambda_{H_1}} \mathbf{V}_{W_1} \mathbf{U}_{W_M}^\top \mathbf{Y} \quad \text{(from equation (35))},
\end{aligned}$$

with semi-orthonormal matrices $\mathbf{U}_{W_M}, \mathbf{U}_{W_{M-1}}, \mathbf{U}_{W_{M-2}}, \dots, \mathbf{U}_{W_1}, \mathbf{V}_{W_1}$ that each has R orthogonal columns, i.e. $\mathbf{U}_{W_M}^\top \mathbf{U}_{W_M} = \mathbf{U}_{W_{M-1}}^\top \mathbf{U}_{W_{M-1}} = \dots = \mathbf{U}_{W_1}^\top \mathbf{U}_{W_1} = \mathbf{V}_{W_1}^\top \mathbf{V}_{W_1} = \mathbf{I}_R$. Furthermore, $\mathbf{U}_{W_M}, \mathbf{U}_{W_{M-1}}, \dots, \mathbf{U}_{W_1}, \mathbf{V}_{W_1}$ are truncated matrices from orthonormal matrices (remove columns that do not correspond with non-zero singular values), hence $\mathbf{U}_{W_M} \mathbf{U}_{W_M}^\top, \mathbf{U}_{W_{M-1}} \mathbf{U}_{W_{M-1}}^\top, \dots, \mathbf{U}_{W_1} \mathbf{U}_{W_1}^\top, \mathbf{V}_{W_1} \mathbf{V}_{W_1}^\top$ are the best rank-R approximations of the identity matrix of the same size.

Let $\overline{\mathbf{H}}^* = \frac{\sqrt{c} s^M}{c s^{2M} + N\lambda_{H_1}} \mathbf{V}_{W_1} \mathbf{U}_{W_M}^\top \in \mathbb{R}^{d_1 \times K}$, then we have $(\mathcal{NC}1)$ $\mathbf{H}_1^* = \overline{\mathbf{H}}^* \mathbf{Y} = \overline{\mathbf{H}}^* \otimes \mathbf{1}_n^\top$, thus we conclude the features within the same class collapse to their class-mean and $\overline{\mathbf{H}}^*$ is the class-means matrix.

From above arguments, we can deduce the geometry of the following $(\mathcal{NC}2)$:

$$\begin{aligned}
\mathbf{W}_M^* \mathbf{W}_M^{\top *} &\propto \mathbf{U}_{W_M} \mathbf{U}_{W_M}^\top \propto \mathcal{P}_R(\mathbf{I}_K), \\
\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* &\propto \mathbf{U}_{W_M} \mathbf{U}_{W_M}^\top \propto \mathcal{P}_R(\mathbf{I}_K), \\
\mathbf{W}_M^* \mathbf{W}_{M-1}^* \mathbf{W}_{M-2}^* \dots \mathbf{W}_2^* \mathbf{W}_1^* \overline{\mathbf{H}}^* &\propto \mathbf{U}_{W_M} \mathbf{U}_{W_M}^\top \propto \mathcal{P}_R(\mathbf{I}_K), \\
(\mathbf{W}_M^* \mathbf{W}_{M-1}^* \dots \mathbf{W}_j^*)(\mathbf{W}_M^* \mathbf{W}_{M-1}^* \dots \mathbf{W}_j^*)^\top &\propto \mathbf{U}_{W_M} \mathbf{U}_{W_M}^\top \propto \mathcal{P}_R(\mathbf{I}_K), \quad \forall\, j \in [M].
\end{aligned} \tag{37}$$

Note that if $R = K$, we have $\mathcal{P}_R(\mathbf{I}_K) = \mathbf{I}_K$.

Also, the product of each weight matrix or features with its transpose will be the multiplier of one of the best rank-r approximations of the identity matrix of the same size. For example, $\mathbf{W}_{M-1}^{*\top} \mathbf{W}_{M-1}^* \propto \mathbf{U}_{W_{M-2}} \mathbf{U}_{W_{M-2}}^\top$ and $\mathbf{W}_{M-1}^* \mathbf{W}_{M-1}^{*\top} \propto \mathbf{U}_{W_{M-1}} \mathbf{U}_{W_{M-1}}^\top$ are two best rank-R approximations of $\mathbf{I}_{d_{M-1}}$ and $\mathbf{I}_{d_M}$, respectively.

Next, we can derive the alignments between weights and features as following $(\mathcal{NC}3)$:

$$\begin{aligned}
\mathbf{W}_M^* \mathbf{W}_{M-1}^* \dots \mathbf{W}_1^* &\propto \mathbf{U}_{W_M} \mathbf{V}_{W_1}^\top \propto \overline{\mathbf{H}}^{*\top}, \\
\mathbf{W}_{M-1}^* \mathbf{W}_{M-2}^* \dots \mathbf{W}_1^* \overline{\mathbf{H}}^* &\propto \mathbf{U}_{W_{M-1}} \mathbf{U}_{W_M}^\top \propto \mathbf{W}_M^{*\top}, \\
\mathbf{W}_M^* \mathbf{W}_{M-1}^* \dots \mathbf{W}_j^* &\propto \mathbf{U}_{W_M} \mathbf{U}_{W_{j-1}}^\top \propto (\mathbf{W}_{j-1}^* \dots \mathbf{W}_1^* \overline{\mathbf{H}}^*)^\top.
\end{aligned} \tag{38}$$

- If $b = MK \sqrt[M]{Kn\lambda_{W_M} \lambda_{W_{M-1}} \dots \lambda_{W_1} \lambda_{H_1}} = \frac{(M-1)^{\frac{M-1}{M}}}{M}$: In this case, x_k^* can either be 0 or the largest positive solution of the equation $b - \frac{M x^{M-1}}{(x^M+1)^2} = 0$. If all the singular values are 0's, we have the trivial global minima $(\mathbf{W}_M^*, \dots, \mathbf{W}_1^*, \mathbf{H}_1^*) = (\mathbf{0}, \dots, \mathbf{0}, \mathbf{0})$.

  If there are exactly $0 < r \leq R$ positive singular values $s_1 = s_2 = \dots = s_r := s > 0$ and $s_{r+1} = \dots = s_R = 0$, then similar as the case $b < \frac{(M-1)^{\frac{M-1}{M}}}{M}$, we also have similar compact SVD form (with exactly r singular vectors, instead of R as the above case). Thus, the nontrivial solutions exhibit $(\mathcal{NC}1)$ and $(\mathcal{NC}3)$ property similarly as the case

$b < \frac{(M-1)^{\frac{M-1}{M}}}{M}$ above.

For $(\mathcal{NC}2)$ property, for $j = 1, \ldots, M$, we have:

$$\begin{aligned} \mathbf{W}_M^* \mathbf{W}_M^{*\top} &\propto \overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* \propto \mathbf{W}_M^* \mathbf{W}_{M-1}^* \mathbf{W}_{M-2}^* \ldots \mathbf{W}_2^* \mathbf{W}_1^* \overline{\mathbf{H}}^* \\ &\propto (\mathbf{W}_M^* \mathbf{W}_{M-1}^* \ldots \mathbf{W}_j^*)(\mathbf{W}_M^* \mathbf{W}_{M-1}^* \ldots \mathbf{W}_j^*)^\top \propto \mathcal{P}_r(\mathbf{I}_K). \end{aligned}$$

We finish the proof of Theorem 3.1 for bias-free case. □

D.4. Full Proof of Theorem 3.1 with Last-layer Unregularized Bias

Now, we state the proof of Theorem 3.1 for general setting with M layers of weight with last-layer bias (i.e., including $\mathbf{b}$) with arbitrary widths $d_M, d_{M-1}, \ldots, d_1$.

Proof of Theorem 3.1 (last-layer bias). First, we have that the objective function f is convex w.r.t $\mathbf{b}$. Hence, we can derive the optimal $\mathbf{b}^*$ through its derivative w.r.t $\mathbf{b}$ (note that $N = Kn$):

$$\begin{aligned} &\frac{1}{N}(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1 + \mathbf{b}^* \mathbf{1}_N^\top - \mathbf{Y}) \mathbf{1}_N = \mathbf{0} \\ \Rightarrow \mathbf{b}^* &= \frac{1}{N}(\mathbf{Y} - \mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1) \mathbf{1}_N = \frac{1}{N} \sum_{k=1}^K \sum_{i=1}^n (\mathbf{y}_k - \mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{h}_{k,i}). \end{aligned} \tag{39}$$

Since $\{\mathbf{y}_k\}$ are one-hot vectors, we have:

$$\mathbf{b}_{k'}^* = \frac{n}{N} - \frac{1}{N} \sum_{k=1}^K \sum_{i=1}^n (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_{k'}^\top \mathbf{h}_{k,i} = \frac{1}{K} - (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_{k'}^\top \mathbf{h}_{\mathbf{G}}, \tag{40}$$

where $\mathbf{h}_G := \frac{1}{N} \sum_{k=1}^K \sum_{i=1}^n \mathbf{h}_{k,i}$ is the features' global-mean and $(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_{k'}$ is k'-th row of $\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1$.

Next, we plug $\mathbf{b}^*$ into f:

$$\begin{aligned} f &= \frac{1}{2Kn} \| \mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1 + \mathbf{b}^* \mathbf{1}_N^\top - \mathbf{Y} \|_F^2 + \frac{\lambda_{W_M}}{2} \| \mathbf{W}_M \|_F^2 + \ldots + \frac{\lambda_{W_2}}{2} \| \mathbf{W}_2 \|_F^2 + \frac{\lambda_{W_1}}{2} \| \mathbf{W}_1 \|_F^2 \\ &+ \frac{\lambda_{H_1}}{2} \| \mathbf{H}_1 \|_F^2 \\ &= \frac{1}{2Kn} \sum_{k=1}^K \sum_{i=1}^n \| \mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{h}_{k,i} + \mathbf{b}^* - \mathbf{y}_k \|_2^2 + \frac{\lambda_{W_M}}{2} \| \mathbf{W}_M \|_F^2 + \ldots + \frac{\lambda_{W_2}}{2} \| \mathbf{W}_2 \|_F^2 + \frac{\lambda_{W_1}}{2} \| \mathbf{W}_1 \|_F^2 \\ &+ \sum_{k=1}^K \sum_{i=1}^n \| \mathbf{h}_{k,i} \|_2^2 \\ &= \frac{1}{2Kn} \sum_{k=1}^K \sum_{i=1}^n \sum_{k'=1}^K \left((\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_{k'}^\top (\mathbf{h}_{k,i} - \mathbf{h}_G) + \frac{1}{K} - \mathbf{1}_{k=k'} \right)^2 + \frac{\lambda_{W_M}}{2} \| \mathbf{W}_M \|_F^2 + \ldots \\ &+ \frac{\lambda_{W_1}}{2} \| \mathbf{W}_1 \|_F^2 + \sum_{k=1}^K \sum_{i=1}^n \| \mathbf{h}_{k,i} \|_2^2 \end{aligned}$$

$$\geq \frac{1}{2Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{k'=1}^{K}\left((\mathbf{W}_M\mathbf{W}_{M-1}\dots\mathbf{W}_2\mathbf{W}_1)_{k'}^\top(\mathbf{h}_{k,i}-\mathbf{h}_G)+\frac{1}{K}-\mathbf{1}_{k=k'}\right)^2+\frac{\lambda_{W_M}}{2}\|\mathbf{W}_M\|_F^2+\dots$$
$$+\frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2+\sum_{k=1}^{K}\sum_{i=1}^{n}\|\mathbf{h}_{k,i}-\mathbf{h}_G\|_2^2$$
$$=\frac{1}{2Kn}\|\mathbf{W}_M\mathbf{W}_{M-1}\dots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1^{'}-(\mathbf{Y}-\frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top)\|_F^2+\frac{\lambda_{W_M}}{2}\|\mathbf{W}_M\|_F^2+\dots+\frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2$$
$$+\frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2+\frac{\lambda_{H_1}}{2}\|\mathbf{H}_1^{'}\|_F^2:=f^{'}(\mathbf{W}_M,\mathbf{W}_{M-1},\dots,\mathbf{W}_2,\mathbf{W}_1,\mathbf{H}_1^{'}),$$

where $\mathbf{H}_1^{'}=[\mathbf{h}_{1,1}-\mathbf{h}_G,\dots,\mathbf{h}_{K,n}-\mathbf{h}_G]\in\mathbb{R}^{d\times N}$ and the inequality is from:

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\|\mathbf{h}_{k,i}\|_2^2=\sum_{k=1}^{K}\sum_{i=1}^{n}\left(\|\mathbf{h}_{k,i}-\mathbf{h}_G\|_2^2+2(\mathbf{h}_{k,i}-\mathbf{h}_G)^\top\mathbf{h}_G+\|\mathbf{h}_G\|_2^2\right)$$
$$=\sum_{k=1}^{K}\sum_{i=1}^{n}\|\mathbf{h}_{k,i}-\mathbf{h}_G\|_2^2+N\|\mathbf{h}_G\|_2^2$$
$$\geq\sum_{k=1}^{K}\sum_{i=1}^{n}\|\mathbf{h}_{k,i}-\mathbf{h}_G\|_2^2, \tag{41}$$

where the equality happens when $\mathbf{h}_G=0$.

Noting that $f^{'}$ has similar form as function f for bias-free case (except the difference of the target matrix $\mathbf{Y}$), we can use the lemmas derived at Section D.2 for $f^{'}$. First, by using Lemma D.2, we have for any critical point $(\mathbf{W}_M,\mathbf{W}_{M-1},\dots,\mathbf{W}_2,\mathbf{W}_1,\mathbf{H}_1^{'})$ of $f^{'}$, we have the following:

$$\lambda_{W_M}\mathbf{W}_M^\top\mathbf{W}_M=\lambda_{W_{M-1}}\mathbf{W}_{M-1}\mathbf{W}_{M-1}^\top,$$
$$\lambda_{W_{M-1}}\mathbf{W}_{M-1}^\top\mathbf{W}_{M-1}=\lambda_{W_{M-2}}\mathbf{W}_{M-2}\mathbf{W}_{M-2}^\top,$$
$$\dots,$$
$$\lambda_{W_2}\mathbf{W}_2^\top\mathbf{W}_2=\lambda_{W_1}\mathbf{W}_1\mathbf{W}_1^\top,$$
$$\lambda_{W_1}\mathbf{W}_1^\top\mathbf{W}_1=\lambda_{H_1}\mathbf{H}_1^{'}\mathbf{H}_1^{'\top}.$$

Let $\mathbf{W}_1=\mathbf{U}_{W_1}\mathbf{S}_{W_1}\mathbf{V}_{W_1}^\top$ be the SVD decomposition of $\mathbf{W}_1$ with $\mathbf{U}_{W_1}\in\mathbb{R}^{d_2\times d_2},\mathbf{V}_{W_1}\in\mathbb{R}^{d_1\times d_1}$ are orthonormal matrices and $\mathbf{S}_{W_1}\in\mathbb{R}^{d_2\times d_1}$ is a diagonal matrix with **decreasing** non-negative singular values. We denote the r singular values of $\mathbf{W}_1$ as $\{s_k\}_{k=1}^r$ ($r\leq R:=\min(K,d_M,\dots,d_1)$, from Lemma D.3) . From Lemma D.4, we have the SVD of other weight matrices as:

$$\mathbf{W}_M=\mathbf{U}_{W_M}\mathbf{S}_{W_M}\mathbf{U}_{W_{M-1}}^\top,$$
$$\mathbf{W}_{M-1}=\mathbf{U}_{W_{M-1}}\mathbf{S}_{W_{M-1}}\mathbf{U}_{W_{M-2}}^\top,$$
$$\mathbf{W}_{M-2}=\mathbf{U}_{W_{M-2}}\mathbf{S}_{W_{M-2}}\mathbf{U}_{W_{M-3}}^\top,$$
$$\mathbf{W}_{M-3}=\mathbf{U}_{W_{M-3}}\mathbf{S}_{W_{M-3}}\mathbf{U}_{W_{M-4}}^\top,$$
$$\dots,$$
$$\mathbf{W}_2=\mathbf{U}_{W_2}\mathbf{S}_{W_2}\mathbf{U}_{W_1}^\top,$$
$$\mathbf{W}_1=\mathbf{U}_{W_1}\mathbf{S}_{W_1}\mathbf{V}_{W_1}^\top,$$

where:

$$\mathbf{S}_{W_j}=\sqrt{\frac{\lambda_{W_1}}{\lambda_{W_j}}}\begin{bmatrix}\operatorname{diag}(s_1,\dots,s_r) & \mathbf{0}_{r\times(d_j-r)}\\ \mathbf{0}_{(d_{j+1}-r)\times r} & \mathbf{0}_{(d_{j+1}-r)\times(d_j-r)}\end{bmatrix}\in\mathbb{R}^{d_{j+1}\times d_j},\quad\forall\, j\in[M],$$

and $\mathbf{U}_{W_M}, \mathbf{U}_{W_{M-1}}, \mathbf{U}_{W_{M-2}}, \mathbf{U}_{W_{M-3}}, \ldots, \mathbf{U}_{W_1}, \mathbf{V}_{W_1}$ are all orthonormal matrices.

From Lemma D.5, denote $c := \frac{\lambda_{W_1}^{M-1}}{\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_2}}$, we have:

$$\begin{aligned}\mathbf{H}_1^{'} &= \mathbf{V}_{W_1} \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{\sqrt{c}s_1^M}{cs_1^{2M}+N\lambda_{H_1}}, \ldots, \frac{\sqrt{c}s_r^M}{cs_r^{2M}+N\lambda_{H_1}}\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\mathbf{C}\in\mathbb{R}^{d_1\times K}} \mathbf{U}_{W_M}^\top \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right) \\ &= \mathbf{V}_{W_1}\mathbf{C}\mathbf{U}_{W_M}^\top \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right).\end{aligned} \tag{42}$$

$$\begin{aligned}&\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1^{'} - \mathbf{Y} \\ &= \mathbf{U}_{W_M} \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{-N\lambda_{H_1}}{cs_1^{2M}+N\lambda_{H_1}}, \ldots, \frac{-N\lambda_{H_1}}{cs_r^{2M}+N\lambda_{H_1}}\right) & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_{K-r} \end{bmatrix}}_{\mathbf{D}\in\mathbb{R}^{K\times K}} \mathbf{U}_{W_M}^\top \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right) \\ &= \mathbf{U}_{W_M}\mathbf{D}\mathbf{U}_{W_M}^\top \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right).\end{aligned}$$

Next, we will calculate the Frobenius norm of $\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1^{'} - \mathbf{Y}$:

$$\begin{aligned}&\|\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1^{'} - \mathbf{Y}\|_F^2 = \left\|\mathbf{U}_{W_M}\mathbf{D}\mathbf{U}_{W_M}^\top \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)\right\|_F^2 \\ &= \operatorname{trace}\left(\mathbf{U}_{W_M}\mathbf{D}\mathbf{U}_{W_M}^\top \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)\left(\mathbf{U}_{W_M}\mathbf{D}\mathbf{U}_{W_M}^\top \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)\right)^\top\right) \\ &= \operatorname{trace}\left(\mathbf{U}_{W_M}\mathbf{D}\mathbf{U}_{W_M}^\top \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)\left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)^\top \mathbf{U}_{W_M}\mathbf{D}\mathbf{U}_{W_M}^\top\right) \\ &= \operatorname{trace}\left(\mathbf{D}^2\mathbf{U}_{W_M}^\top \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)\left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)^\top \mathbf{U}_{W_M}\right).\end{aligned} \tag{43}$$

Note that:

$$\begin{aligned}\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top &= \left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\otimes \mathbf{1}_n^\top, \\ \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)\left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)^\top &= \left(\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\otimes \mathbf{1}_n^\top\right)\left(\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\otimes \mathbf{1}_n^\top\right)^\top \\ &= \left(\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\otimes \mathbf{1}_n^\top\right)\left(\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\otimes \mathbf{1}_n\right) \\ &= \left(\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\right)\otimes\left(\mathbf{1}_n^\top\mathbf{1}_n\right) \\ &= n\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right),\end{aligned}$$

since $\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$ is an idempotent matrix.

Next, we have:

$$\mathbf{U}_{W_M}^\top \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)\left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)^\top \mathbf{U}_{W_M} = n\mathbf{U}_{W_M}^\top\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\mathbf{U}_{W_M}$$

$$= n \left(\mathbf{I}_K - \frac{1}{K} \mathbf{U}_{W_M}^\top \mathbf{1}_K \mathbf{1}_K^\top \mathbf{U}_{W_M} \right).$$

We denote $\mathbf{q} = \mathbf{U}_{W_M}^\top \mathbf{1}_K = [q_1, \dots, q_K]^\top \in \mathbb{R}^K$, then q_k will equal the sum of entries of the k-th column of $\mathbf{U}_{W_M}$. Hence, $\mathbf{U}_{W_M}^\top \mathbf{1}_K \mathbf{1}_K^\top \mathbf{U}_{W_M} = \mathbf{q}\mathbf{q}^\top = (q_i q_j)_{i,j}$. Note that from the orthonormality of $\mathbf{U}_{W_M}$, we can deduce $\sum_{k=1}^K q_k^2 = K$. Thus, continue from equation (43):

$$\begin{aligned} \|\mathbf{W}_M \mathbf{W}_{M-1} \dots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1^{'} - \mathbf{Y}\|_F^2 &= n \operatorname{trace} \left(\mathbf{D}^2 \left(\mathbf{I}_K - \frac{1}{K} \mathbf{q}\mathbf{q}^\top \right) \right) \\ &= n \left(\sum_{k=1}^r \left(1 - \frac{1}{K} q_k^2 \right) \frac{(-N\lambda_{H_1})^2}{(cs_k^{2M} + N\lambda_{H_1})^2} + \sum_{h=r+1}^K \left(1 - \frac{1}{K} q_h^2 \right) \right). \end{aligned} \tag{44}$$

Similarly, we calculate the Frobenius norm for $\mathbf{H}_1^{'}$, continue from the RHS of equation (42):

$$\begin{aligned} \|\mathbf{H}_1^{'}\|_F^2 &= \operatorname{trace} \left(\mathbf{V}_{W_1} \mathbf{C} \mathbf{U}_{W_M}^\top \left(\mathbf{Y} - \frac{1}{K} \mathbf{1}_K \mathbf{1}_N^\top \right) \left(\mathbf{Y} - \frac{1}{K} \mathbf{1}_K \mathbf{1}_N^\top \right)^\top \mathbf{U}_{W_M} \mathbf{C}^\top \mathbf{V}_{W_1}^\top \right) \\ &= n \operatorname{trace} \left(\mathbf{C}^\top \mathbf{C} \left(\mathbf{I}_K - \frac{1}{K} \mathbf{q}\mathbf{q}^\top \right) \right) \\ &= n \sum_{k=1}^r \left(1 - \frac{1}{K} q_k^2 \right) \frac{cs_k^{2M}}{(cs_k^{2M} + N\lambda_{H_1})^2}. \end{aligned} \tag{45}$$

Plug the equations (44), (45) and the SVD of weight matrices into $f^{'}$ yields:

$$\begin{aligned} &\frac{1}{2Kn} \left\| \mathbf{W}_M \mathbf{W}_{M-1} \dots \mathbf{W}_1 \mathbf{H}_1^{'} - (\mathbf{Y} - \frac{1}{K} \mathbf{1}_K \mathbf{1}_N^T) \right\|_F^2 + \frac{\lambda_{W_M}}{2} \|\mathbf{W}_M\|_F^2 + \dots \frac{\lambda_{W_1}}{2} \|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2} \|\mathbf{H}_1^{'}\|_F^2 \\ &= \frac{1}{2K} \sum_{k=1}^r \left(1 - \frac{1}{K} q_k^2 \right) \left(\frac{-N\lambda_{H_1}}{cs_k^{2M} + N\lambda_{H_1}} \right)^2 + \frac{1}{2K} \sum_{h=r+1}^K \left(1 - \frac{1}{K} q_h^2 \right) + \frac{\lambda_{W_M}}{2} \sum_{k=1}^r \frac{\lambda_{W_1}}{\lambda_{W_M}} s_k^2 \\ &+ \frac{\lambda_{W_{M-1}}}{2} \sum_{k=1}^r \frac{\lambda_{W_1}}{\lambda_{W_{M-1}}} s_k^2 + \dots + \frac{\lambda_{W_1}}{2} \sum_{k=1}^r s_k^2 + \frac{n\lambda_{H_1}}{2} \sum_{k=1}^r \left(1 - \frac{1}{K} q_k^2 \right) \frac{cs_k^{2M}}{(cs_k^{2M} + N\lambda_{H_1})^2} \\ &= \frac{1}{2K} \sum_{k=1}^r \left(1 - \frac{1}{K} q_k^2 \right) \frac{(N\lambda_{H_1})^2}{(cs_k^{2M} + N\lambda_{H_1})^2} + \frac{n\lambda_{H_1}}{2} \sum_{k=1}^r \left(1 - \frac{1}{K} q_k^2 \right) \frac{cs_k^{2M}}{(cs_k^{2M} + N\lambda_{H_1})^2} + \frac{M\lambda_{W_1}}{2} \sum_{k=1}^r s_k^2 \\ &+ \frac{1}{2K} \sum_{h=r+1}^K \left(1 - \frac{1}{K} q_h^2 \right) \\ &= \frac{n\lambda_{H_1}}{2} \sum_{k=1}^r \frac{1 - \frac{1}{K} q_k^2}{cs_k^{2M} + N\lambda_{H_1}} + \frac{M\lambda_{W_1}}{2} \sum_{k=1}^r s_k^2 + \frac{1}{2K} \sum_{h=r+1}^K \left(1 - \frac{1}{K} q_h^2 \right) \\ &= \frac{1}{2K} \sum_{k=1}^r \left(\frac{1 - \frac{1}{K} q_k^2}{\frac{cs_k^{2M}}{N\lambda_{H_1}} + 1} + MK\lambda_{W_1} \sqrt[M]{\frac{N\lambda_{H_1}}{c}} \left(\sqrt[M]{\frac{cs_k^{2M}}{N\lambda_{H_1}}} \right) \right) + \frac{1}{2K} \sum_{h=r+1}^K \left(1 - \frac{1}{K} q_h^2 \right) \\ &= \frac{1}{2K} \sum_{k=1}^r \left(\frac{1 - \frac{1}{K} q_k^2}{x_k^M + 1} + b x_k \right) + \frac{1}{2K} \sum_{h=r+1}^K \left(1 - \frac{1}{K} q_h^2 \right), \end{aligned} \tag{46}$$

with $x_k := \sqrt[M]{\frac{cs_k^{2M}}{N\lambda_{H_1}}}$ and $b := MK\lambda_{W_1} \sqrt[M]{\frac{N\lambda_{H_1}}{c}} = MK\lambda_{W_1} \sqrt[M]{\frac{Kn\lambda_{W_M}\lambda_{W_{M-2}} \dots \lambda_{W_1}\lambda_{H_1}}{\lambda_{W_1}^{M-1}}} = MK \sqrt[M]{Kn\lambda_{W_M}\lambda_{W_{M-1}} \dots \lambda_{W_1}\lambda_{H_1}}$.

Before continue optimizing the RHS of equation (46), we first simplify it by proving if $s_k > 0$ then $q_k = 0$, i.e. sum of entries of k-th column of $\mathbf{U}_{W_M}$ equals 0. To prove this, we will utilize a property of $\mathbf{H}_1^{'} = [\mathbf{h}_{1,1} - \mathbf{h}_G, \ldots, \mathbf{h}_{K,n} - \mathbf{h}_G]$, which is the sum of entries on every row equals 0. First, we connect $\mathbf{W}_M$ and $\mathbf{H}_1^{'}$ through:

$$\frac{\partial f^{'}}{\partial \mathbf{W}_M} = \frac{1}{N}\left(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1\mathbf{H}_1^{'} - \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)\right)\mathbf{H}_1^{'\top}\mathbf{W}_1^\top\ldots\mathbf{W}_{M-1}^\top + \lambda_{W_M}\mathbf{W}_M = \mathbf{0}$$

$$\Rightarrow \mathbf{W}_M = \left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)\mathbf{H}_1^{'\top}\underbrace{\mathbf{W}_1^\top\ldots\mathbf{W}_{M-1}^\top\left(\mathbf{W}_{M-1}\ldots\mathbf{W}_1\mathbf{H}_1^{'}\mathbf{H}_1^{'\top}\mathbf{W}_1^\top\ldots\mathbf{W}_{M-1}^\top + N\lambda_{W_M}\mathbf{I}_K\right)^{-1}}_{\mathbf{G}}. \quad (47)$$

From the definition of $\mathbf{H}_1^{'}$, we know that the sum of entries of every column of $\mathbf{H}_1^{'\top}$ is 0. Recall the class-mean definition $\mathbf{h}_k = \frac{1}{n}\sum_{i=1}^n \mathbf{h}_{k,i}$, we have:

$$\left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right)\mathbf{H}_1^{'\top} = \mathbf{Y}\mathbf{H}_1^{'\top} = n\begin{bmatrix}(\mathbf{h}_1 - \mathbf{h}_G)^\top \\ (\mathbf{h}_2 - \mathbf{h}_G)^\top \\ \ldots \\ (\mathbf{h}_K - \mathbf{h}_G)^\top\end{bmatrix}$$

$$\Rightarrow \mathbf{W}_M = n\begin{bmatrix}(\mathbf{h}_1 - \mathbf{h}_G)^\top \\ (\mathbf{h}_2 - \mathbf{h}_G)^\top \\ \ldots \\ (\mathbf{h}_K - \mathbf{h}_G)^\top\end{bmatrix}\mathbf{G},$$

and thus, the sum of entries of every column of $\mathbf{W}_M$ equals 0. From the SVD $\mathbf{W}_M = \mathbf{U}_{W_M}\mathbf{S}_{W_M}\mathbf{V}_{W_M}^\top$, denote $\mathbf{u}_j$ and $\mathbf{v}_j$ the j-th column of $\mathbf{U}_{W_M}$ and $\mathbf{V}_{W_M}$, respectively. We have from the definition of left and right singular vectors:

$$\mathbf{W}_M\mathbf{v}_j = s_j\mathbf{u}_j, \quad (48)$$

and since the sum of entries of every column of $\mathbf{W}_M$ equals 0, we have the sum of entries of vector $\mathbf{W}_M\mathbf{v}_j$ equals 0. Thus, if $s_j > 0$, we have $q_j = 0$.

Return to the expression of $f^{'}$ as the RHS of equation (46), notice that it is separable w.r.t each singular value s_j, we will analyze how each singular value contribute to the value of the expression (46). For every singular value s_j with $j = 1, \ldots, r$, if $s_j > 0$, then $q_j = 0$, and its contribution to the expression (46) will be $\frac{1}{2K}(\frac{1}{x_j^M + 1} + bx_j) = \frac{1}{2K}g(x_j)$ (with the minimizer of $g(x)$ has been studied in Section D.2.1). Otherwise, if $s_j = 0$ (hence $x_j = 0$), its contribution to the value of the expression (46) will be $\frac{1 - \frac{1}{K}q_j^2}{2K}$, and it eventually be $\frac{1}{2K}$ because $\sum_{k=1}^K \frac{1}{K}q_j^2$ always equal 1, thus $\frac{1}{K}q_j^2$ has no additional contribution to the expression (46). Therefore, it is a comparision between $\frac{1}{2K}$ and $\frac{1}{2K}\min_{x_j > 0} g(x_j)$ to decide whether $s_j^* = 0$ or $s_j^* = \sqrt[2M]{\frac{N\lambda_{H_1}}{c}}\sqrt{x_j^*}$ with $x_j^* = \arg\min_{x>0} g(x)$. Therefore, we consider three cases:

- If $b > \frac{(M-1)^{\frac{M-1}{M}}}{M}$: In this case, $g(x)$ is minimized at $x = 0$ and $g(0) = 1$. Hence, $\frac{1}{2K} < \frac{1}{2K}\min_{x_j > 0} g(x_j)$ and thus, $s_j^* = 0 \,\forall j = 1, \ldots, r$.
- If $b < \frac{(M-1)^{\frac{M-1}{M}}}{M}$: In this case, $g(x)$ is minimized at some $x_0 > \sqrt[M]{M-1}$ and $g(x_0) < 1$. Hence, $\frac{1}{2K}\min_{x_j>0} g(x_j) < \frac{1}{2K}$ and thus, $s_j^* = \sqrt[2M]{\frac{N\lambda_{H_1}}{c}}\sqrt{x_0}\,\forall\, j = 1, \ldots, r$.
  We also note that in this case, we have $q_j = 0 \,\forall j = 1, \ldots, r$ (meaning the sum of entries of every column in the first r columns of $\mathbf{U}_{W_M}$ is equal 0).
- If $b = \frac{(M-1)^{\frac{M-1}{M}}}{M}$: In this case, $g(x)$ is minimized at $x = 0$ or some $x = x_0 > \sqrt[M]{M-1}$ with $g(0) = g(x_0) = 1$. Therefore, s_j^* can either be 0 or x_0 as long as $\{s_k\}_{k=1}^r$ is a decreasing sequence.

To help for the conclusion of the geometry properties of weight matrices and features, we state a lemma as following:

Lemma D.6. *Let $\mathbf{W} \in \mathbb{R}^{K \times d_M}$ be a matrix with $r \leq K-1$ singular values equal a positive constant $s > 0$. If there exists a compact SVD form of $\mathbf{W}$ as $\mathbf{W} = s\mathbf{U}\mathbf{V}^\top$ with semi-orthonormal matrices $\mathbf{U} \in \mathbb{R}^{K \times r}, \mathbf{V} \in \mathbb{R}^{d_M \times r}$ such that the sum of entries of every column of $\mathbf{U}$ equals 0. Then, $\mathbf{W}\mathbf{W}^\top \propto \mathbf{U}\mathbf{U}^\top$ and $\mathbf{U}\mathbf{U}^\top$ is a best rank-r approximation of the simplex ETF $(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)$.*

Proof. Let's denote $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$ with $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are r orthonormal vectors. Since the sum of entries in each $\mathbf{u}_i$ equals 0, $\frac{1}{\sqrt{K}}\mathbf{1}_K$ can be added to the set $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ to form $r+1$ orthonormal vectors. Let $\hat{\mathbf{U}} = [\mathbf{u}_1, \ldots, \mathbf{u}_r, \frac{1}{\sqrt{K}}\mathbf{1}_K]$, we have $\dim(\mathrm{Col}\,\hat{\mathbf{U}}) = r+1$. Hence, $\dim(\mathrm{Null}\,\hat{\mathbf{U}}^\top) = K - r - 1$ and thus, we can choose an orthonormal basis of $\mathrm{Null}\,\hat{\mathbf{U}}^\top$ including $K-r-1$ orthonormal vectors $\{\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_{K-1}\}$. And because these $K-r-1$ orthonormal vectors are in $\mathrm{Null}\,\hat{\mathbf{U}}^\top$, we can add these vectors to the set $\{\mathbf{u}_1, \ldots, \mathbf{u}_r, \frac{1}{\sqrt{K}}\mathbf{1}_K\}$ to form a basis of $\mathbb{R}^K$ including K orthonormal vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_{K-1}, \frac{1}{\sqrt{K}}\mathbf{1}_K\}$. We denote $\overline{\mathbf{U}} = [\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_{K-1}, \frac{1}{\sqrt{K}}\mathbf{1}_K] \in \mathbb{R}^{K \times K}$. We have $\overline{\mathbf{U}}^\top\overline{\mathbf{U}} = \mathbf{I}_K$. From the Inverse Matrix Theorem, we deduce that $\overline{\mathbf{U}}^{-1} = \overline{\mathbf{U}}^\top$ and thus, $\overline{\mathbf{U}}$ is an orthonormal matrix. We have $\overline{\mathbf{U}}$ is an orthonormal matrix with the last column $\frac{1}{\sqrt{K}}\mathbf{1}_K$, hence by simple matrix multiplication, we have:

$$
\begin{aligned}
&[\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_{K-1}][\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_{K-1}]^\top = \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top \\
&\Rightarrow \overline{\mathbf{U}} \begin{bmatrix} \mathbf{I}_{K-1} & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \overline{\mathbf{U}}^\top = \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top.
\end{aligned} \tag{49}
$$

Therefore, $\mathbf{U}\mathbf{U}^\top$ is the best rank-r approximation of $\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$, and the proof for the lemma is finished. □

Thus, we finish bounding f and the equality conditions are as following:

- If $b = MK\sqrt[M]{Kn\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_1}\lambda_{H_1}} > \frac{(M-1)^{\frac{M-1}{M}}}{M}$: all the singular values of $\mathbf{W}_1$ are zeros. Therefore, the singular values of $\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{H}_1^{'}$ are also all zeros. In this case, $f(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1, \mathbf{b})$ is minimized at $(\mathbf{W}_M^*, \mathbf{W}_{M-1}^*, \ldots, \mathbf{W}_1^*, \mathbf{H}_1^*, \mathbf{b}^*) = (\mathbf{0}, \mathbf{0}, \ldots \mathbf{0}, \mathbf{0}, \frac{1}{K}\mathbf{1}_K)$.

- If $b = MK\sqrt[M]{Kn\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_1}\lambda_{H_1}} < \frac{(M-1)^{\frac{M-1}{M}}}{M}$: In this case, $\mathbf{W}_1^*$ will have the its r (r will be specified later) singular values all equal a multiplier of the largest positive solution of the equation $b - \frac{Mx^{M-1}}{(x^M+1)^2} = 0$, denoted as s. Hence, we can write the compact SVD form (with a bit of notation abuse) of $\mathbf{W}_{M-1}^*$ as $\mathbf{W}_1^* = s\mathbf{U}_{W_1}\mathbf{V}_{W_1}^\top$ with semi-orthonormal matrices $\mathbf{U}_{W_1} \in \mathbb{R}^{d_2 \times r}, \mathbf{V}_{W_1} \in \mathbb{R}^{d_1 \times r}$ (note that $\mathbf{U}_{W_1}^\top\mathbf{U}_{W_1} = \mathbf{I}$ and $\mathbf{V}_{W_1}^\top\mathbf{V}_{W_1} = \mathbf{I}$).

  Similarly, we also have the compact SVD form of other weight matrices and feature matrix as:

$$
\begin{aligned}
\mathbf{W}_M^* &= \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_M}}}s\mathbf{U}_{W_M}\mathbf{U}_{W_{M-1}}^\top, \\
\mathbf{W}_{M-1}^* &= \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_{M-1}}}}s\mathbf{U}_{W_{M-1}}\mathbf{U}_{W_{M-2}}^\top, \\
&\ldots \\
\mathbf{W}_1^* &= s\mathbf{U}_{W_1}\mathbf{V}_{W_1}^\top, \\
\mathbf{H}_1^{'*} &= \frac{\sqrt{c}s^M}{cs^{2M} + N\lambda_{H_1}}\mathbf{V}_{W_1}\mathbf{U}_{W_M}^\top\left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right),
\end{aligned}
$$

  with semi-orthonormal matrices $\mathbf{U}_{W_M}, \mathbf{U}_{W_{M-1}}, \ldots, \mathbf{U}_{W_1}, \mathbf{V}_{W_1}$ that each has r orthogonal columns, i.e., $\mathbf{U}_{W_M}^\top\mathbf{U}_{W_M} = \mathbf{U}_{W_{M-1}}^\top\mathbf{U}_{W_{M-1}} = \ldots = \mathbf{U}_{W_1}^\top\mathbf{U}_{W_1} = \mathbf{V}_{W_1}^T\mathbf{V}_{W_1} = \mathbf{I}_r$. Furthermore, $\mathbf{U}_{W_M}, \mathbf{U}_{W_{M-1}}, \ldots, \mathbf{U}_{W_1}, \mathbf{V}_{W_1}$ are truncated matrices from orthonormal matrices (remove columns that does not correspond with non-zero singular values), hence $\mathbf{U}_{W_M}\mathbf{U}_{W_M}^\top, \mathbf{U}_{W_{M-1}}\mathbf{U}_{W_{M-1}}^\top, \ldots, \mathbf{U}_{W_1}\mathbf{U}_{W_1}^\top, \mathbf{V}_{W_1}\mathbf{V}_{W_1}^\top$ are the best rank-r approximations of the identity matrix of the same size.

Since $\left(\mathbf{Y} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_N^\top\right) = \left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\mathbf{Y} = \left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right) \otimes \mathbf{1}_n^\top$, let $\overline{\mathbf{H}}^* = \frac{\sqrt{c}s^M}{cs^{2M} + N\lambda_{H_1}}\mathbf{V}_{W_1}\mathbf{U}_{W_M}^\top\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right) \in \mathbb{R}^{d_1 \times K}$, then we have $(\mathcal{NC}1)$ $\mathbf{H}_1^{'*} = \overline{\mathbf{H}}^*\mathbf{Y} = \overline{\mathbf{H}}^* \otimes \mathbf{1}_n^\top$, thus we conclude the features within the same class collapse to their class-mean and $\overline{\mathbf{H}}^*$ is the class-means matrix. We also have $\mathbf{h}_G = \mathbf{0}$ (the equality condition of inequality (41)), hence $\mathbf{H}_1^* = \mathbf{H}_1^{'*}$. Furthermore, clearly we have $\operatorname{rank}(\mathbf{H}_1^{'*}) = \operatorname{rank}(\overline{\mathbf{H}}^*)$ and since $\mathbf{h}_G = 0$, we have $r = \operatorname{rank}(\mathbf{H}_1^{'*}) = \operatorname{rank}(\overline{\mathbf{H}}^*) \leq K - 1$. Hence, $r = \min(R, K-1)$.

By using Lemma D.6 for $\mathbf{W}_M$ with the note $q_j = 0 \,\forall\, j \leq r$, we have $\mathbf{U}_W\mathbf{U}_W^\top$ is a best rank-r approximation of the simplex ETF $\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$. Thus, we can deduce the geometry of the following $(\mathcal{NC}2)$:

$$\begin{aligned}
\mathbf{W}_M^*\mathbf{W}_M^{\top *} &\propto \mathbf{U}_{W_M}\mathbf{U}_{W_M}^\top \propto \mathcal{P}_r(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top), \\
\overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* &\propto (\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top)\mathbf{U}_{W_M}\mathbf{U}_{W_M}^\top(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top) \propto \mathbf{U}_{W_M}\mathbf{U}_{W_M}^\top \propto \mathcal{P}_r(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top), \\
\mathbf{W}_M^*\mathbf{W}_{M-1}^* \ldots \mathbf{W}_2^*\mathbf{W}_1^*\overline{\mathbf{H}}^* &\propto \mathbf{U}_{W_M}\mathbf{U}_{W_M}^\top(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top) \propto \mathbf{U}_{W_M}\mathbf{U}_{W_M}^\top \propto \mathcal{P}_r(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top), \\
(\mathbf{W}_M^*\mathbf{W}_{M-1}^* \ldots \mathbf{W}_j^*)(\mathbf{W}_M^*\mathbf{W}_{M-1}^* \ldots \mathbf{W}_j^*)^\top &\propto \mathbf{U}_{W_M}\mathbf{U}_{W_M}^\top \propto \mathcal{P}_r(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top) \quad \forall\, j \in [M].
\end{aligned} \tag{50}$$

Note that if $r = K - 1$, we have $\mathcal{P}_r(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top) = \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$.

Also, the product of each weight matrix or features with its transpose will be the multiplier of one of the best rank-r approximations of the identity matrix of the same size. For example, $\mathbf{W}_{M-1}^{*\top}\mathbf{W}_{M-1}^* \propto \mathbf{U}_{W_{M-2}}\mathbf{U}_{W_{M-2}}^\top$ and $\mathbf{W}_{M-1}^*\mathbf{W}_{M-1}^{*\top} \propto \mathbf{U}_{W_{M-1}}\mathbf{U}_{W_{M-1}}^\top$ are two best rank-r approximations of $\mathbf{I}_{d_{M-1}}$ and $\mathbf{I}_{d_M}$, respectively.

Next, we can derive the alignments between weights and features as following $(\mathcal{NC}3)$:

$$\begin{aligned}
\mathbf{W}_M^*\mathbf{W}_{M-1}^* \ldots \mathbf{W}_1^* &\propto \mathbf{U}_{W_M}\mathbf{V}_{W_1}^\top \propto \overline{\mathbf{H}}^{*\top}, \\
\mathbf{W}_{M-1}^*\mathbf{W}_{M-2}^* \ldots \mathbf{W}_1^*\overline{\mathbf{H}}^* &\propto \mathbf{U}_{W_{M-1}}\mathbf{U}_{W_M}^\top \propto \mathbf{W}_M^{*\top}, \\
\mathbf{W}_M^*\mathbf{W}_{M-1}^* \ldots \mathbf{W}_j^* &\propto \mathbf{U}_{W_M}\mathbf{U}_{W_{j-1}}^\top \propto (\mathbf{W}_{j-1}^* \ldots \mathbf{W}_1^*\overline{\mathbf{H}}^*)^\top.
\end{aligned} \tag{51}$$

- If $b = MK\sqrt[M]{Kn\lambda_{W_M}\lambda_{W_{M-1}} \ldots \lambda_{W_1}\lambda_{H_1}} = \frac{(M-1)^{\frac{M-1}{M}}}{M}$: In this case, x_k^* can either be 0 or the largest positive solution of the equation $b - \frac{Mx^{M-1}}{(x^M+1)^2} = 0$. If all the singular values are 0's, we have the trivial global minima $(\mathbf{W}_M^*, \ldots, \mathbf{W}_1^*, \mathbf{H}_1^*, \mathbf{b}^*) = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{0}, \frac{1}{K}\mathbf{1}_K)$.

  If there are exactly $0 < t \leq r = \min(R, K-1)$ positive singular values $s_1 = s_2 = \ldots = s_t := s > 0$ and $s_{t+1} = \ldots = s_r = 0$, we also have compact SVD form similar as the case $b < \frac{(M-1)^{\frac{M-1}{M}}}{M}$, (with exactly t singular vectors, instead of r as the above case). Thus, the nontrivial solutions exhibit $(\mathcal{NC}1)$ and $(\mathcal{NC}3)$ property similarly as the case $b < \frac{(M-1)^{\frac{M-1}{M}}}{M}$ above.

  For $(\mathcal{NC}2)$ property, for $j = 1, \ldots, M$, we have:

$$\begin{aligned}
\mathbf{W}_M^*\mathbf{W}_M^{*\top} &\propto \overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* \propto \mathbf{W}_M^*\mathbf{W}_{M-1}^*\mathbf{W}_{M-2}^* \ldots \mathbf{W}_2^*\mathbf{W}_1^*\overline{\mathbf{H}}^* \\
&\propto (\mathbf{W}_M^*\mathbf{W}_{M-1}^* \ldots \mathbf{W}_j^*)(\mathbf{W}_M^*\mathbf{W}_{M-1}^* \ldots \mathbf{W}_j^*)^\top \propto \mathcal{P}_t(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top).
\end{aligned}$$

We finish the proof. □

E. Proof of Theorem 4.1

Theorem E.1. *Let $d \geq K$ and $(\mathbf{W}^*, \mathbf{H}^*)$ be any global minimizer of problem (5). Then, we have:*

$(\mathcal{NC}1)$ $\quad \mathbf{H}^* = \overline{\mathbf{H}}^* \mathbf{Y} \Leftrightarrow \mathbf{h}_{k,i}^* = \mathbf{h}_k^* \; \forall \, k \in [K], i \in [n_k]$, *where* $\overline{\mathbf{H}}^* = [\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*] \in \mathbb{R}^{d \times K}$.

$(\mathcal{NC}3)$ $\quad \mathbf{w}_k^* = \sqrt{\frac{n_k \lambda_H}{\lambda_W}} \mathbf{h}_k^* \quad \forall \, k \in [K]$.

$(\mathcal{NC}2)$ *Let $a := N^2 \lambda_W \lambda_H$, we have:*

$$\mathbf{W}^* \mathbf{W}^{*\top} = \operatorname{diag}\left\{ s_k^2 \right\}_{k=1}^K,$$
$$\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* = \operatorname{diag}\left\{ \frac{s_k^2}{(s_k^2 + N\lambda_H)^2} \right\}_{k=1}^K,$$

$$\mathbf{W}^* \mathbf{H}^* = \operatorname{diag}\left\{ \frac{s_k^2}{s_k^2 + N\lambda_H} \right\}_{k=1}^K \mathbf{Y}$$
$$= \begin{bmatrix} \frac{s_1^2}{s_1^2 + N\lambda_H} \mathbf{1}_{n_1}^\top & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \frac{s_K^2}{s_K^2 + N\lambda_H} \mathbf{1}_{n_K}^\top \end{bmatrix}.$$

where:

- *If $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_K} \leq 1$:*

$$s_k = \sqrt{\sqrt{\frac{n_k \lambda_H}{\lambda_W}} - N\lambda_H} \quad \forall \, k$$

- *If there exists a $j \in [K-1]$ s.t. $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_j} \leq 1 < \frac{a}{n_{j+1}} \leq \ldots \leq \frac{a}{n_K}$:*

$$s_k = \begin{cases} \sqrt{\sqrt{\frac{n_k \lambda_H}{\lambda_W}} - N\lambda_H} & \forall \, k \leq j \\ 0 & \forall \, k > j \end{cases}.$$

- *If $1 < \frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_K}$:*

$$(s_1, s_2, \ldots, s_K) = (0, 0, \ldots, 0),$$

  and $(\mathbf{W}^*, \mathbf{H}^*) = (\mathbf{0}, \mathbf{0})$ in this case.

And, for any k such that $s_k = 0$, we have:

$$\mathbf{w}_k^* = \mathbf{h}_k^* = \mathbf{0}.$$

Theorem E.2. *Let $d < K$, thus $R = \min(d, K) = d$ and $(\mathbf{W}^*, \mathbf{H}^*)$ be any global minimizer of problem (5). Then, we have:*

$(\mathcal{NC}1)$ $\quad \mathbf{H}^* = \overline{\mathbf{H}}^* \mathbf{Y} \Leftrightarrow \mathbf{h}_{k,i}^* = \mathbf{h}_k^* \; \forall \, k \in [K], i \in [n_k]$, *where* $\overline{\mathbf{H}}^* = [\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*] \in \mathbb{R}^{d \times K}$.

$(\mathcal{NC}3)$ $\quad \mathbf{w}_k^* = \sqrt{\frac{n_k \lambda_H}{\lambda_W}} \mathbf{h}_k^* \quad \forall \, k \in [K]$.

$(\mathcal{NC}2)$ *Let $a := N^2 \lambda_W \lambda_H$, we define $\{s_k\}_{k=1}^K$ as follows:*

- *If* $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_R} \leq 1$*:*

$$s_k = \begin{cases} \sqrt{\sqrt{\frac{n_k \lambda_H}{\lambda_W}} - N\lambda_H} & \forall\, k \leq R \\ 0 & \forall\, k > R \end{cases}. \tag{52}$$

Then, if $b/n_R = 1$ *or* $n_R > n_{R+1}$*, we have:*

$$\mathbf{W}^* \mathbf{W}^{*\top} = \operatorname{diag}\left\{s_k^2\right\}_{k=1}^K,$$
$$\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* = \operatorname{diag}\left\{\frac{s_k^2}{(s_k^2 + N\lambda_H)^2}\right\}_{k=1}^K,$$
$$\mathbf{W}^* \overline{\mathbf{H}}^* = \operatorname{diag}\left\{\frac{s_k^2}{s_k^2 + N\lambda_H}\right\}_{k=1}^K,$$

and for any $k > R$*, we have* $\mathbf{w}_k^* = \mathbf{h}_k^* = \mathbf{0}$*.*

If $b/n_R < 1$ *and there exists* $k \leq R$*,* $l > R$ *such that* $n_{k-1} > n_k = n_{k+1} = \ldots = n_R = \ldots = n_l > n_{l+1}$*, then:*

$$\mathbf{W}^* \mathbf{W}^{*\top} = \begin{bmatrix} s_1^2 & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \ldots & s_{k-1}^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & s_k^2 \mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{53}$$

$$\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* = \begin{bmatrix} \frac{s_1^2}{(s_1^2 + N\lambda_H)^2} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \ldots & \frac{s_{k-1}^2}{(s_{k-1}^2 + N\lambda_H)^2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \frac{s_k^2}{(s_k^2 + N\lambda_H)^2} \mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{54}$$

$$\mathbf{W}^* \overline{\mathbf{H}}^* = \begin{bmatrix} \frac{s_1^2}{s_1^2 + N\lambda_H} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \ldots & \frac{s_{k-1}^2}{s_{k-1}^2 + N\lambda_H} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \frac{s_k^2}{s_k^2 + N\lambda_H} \mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{55}$$

and for any $k > l > R$*, we have* $\mathbf{w}_k^* = \mathbf{h}_k^* = \mathbf{0}$*.*

- *If there exists a* $j \in [R-1]$ *s.t.* $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_j} \leq 1 < \frac{a}{n_{j+1}} \leq \ldots \leq \frac{a}{n_R}$*:*

$$s_k = \begin{cases} \sqrt{\sqrt{\frac{n_k \lambda_H}{\lambda_W}} - N\lambda_H} & \forall\, k \leq j \\ 0 & \forall\, k > j \end{cases}.$$

Then, we have:

$$\mathbf{W}^* \mathbf{W}^{*\top} = \operatorname{diag}\left\{s_k^2\right\}_{k=1}^K,$$
$$\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* = \operatorname{diag}\left\{\frac{s_k^2}{(s_k^2 + N\lambda_H)^2}\right\}_{k=1}^K,$$
$$\mathbf{W}^* \overline{\mathbf{H}}^* = \operatorname{diag}\left\{\frac{s_k^2}{s_k^2 + N\lambda_H}\right\}_{k=1}^K,$$

and for any $k > j$*, we have* $\mathbf{w}_k^* = \mathbf{h}_k^* = \mathbf{0}$

- *If* $1 < \frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_R}$*:*

$$(s_1, s_2, \ldots, s_K) = (0, 0, \ldots, 0),$$

and $(\mathbf{W}^*, \mathbf{H}^*) = (\mathbf{0}, \mathbf{0})$ *in this case.*

Proof of Theorem E.1 and E.2. By definition, any critical point $(\mathbf{W}, \mathbf{H})$ of $f(\mathbf{W}, \mathbf{H})$ satisfies the following:

$$\frac{\partial f}{\partial \mathbf{W}} = \frac{1}{N}(\mathbf{W}\mathbf{H} - \mathbf{Y})\mathbf{H}^\top + \lambda_W \mathbf{W} = \mathbf{0}, \tag{56}$$

$$\frac{\partial f}{\partial \mathbf{H}} = \frac{1}{N}\mathbf{W}^\top(\mathbf{W}\mathbf{H} - \mathbf{Y}) + \lambda_H \mathbf{H} = \mathbf{0}. \tag{57}$$

From $\mathbf{0} = \mathbf{W}^\top \frac{\partial f}{\partial \mathbf{W}} - \frac{\partial f}{\partial \mathbf{H}} \mathbf{H}^\top$, we have:

$$\lambda_W \mathbf{W}^\top \mathbf{W} = \lambda_H \mathbf{H}\mathbf{H}^\top. \tag{58}$$

Also, from $\frac{\partial f}{\partial \mathbf{H}} = \mathbf{0}$, solving for $\mathbf{H}$ yields:

$$\mathbf{H} = (\mathbf{W}^\top \mathbf{W} + N\lambda_H \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{Y}. \tag{59}$$

Let $\mathbf{W} = \mathbf{U}_W \mathbf{S}_W \mathbf{V}_W^\top$ be the SVD decomposition of $\mathbf{W}$ with orthonormal matrices $\mathbf{U}_W \in \mathbb{R}^{K \times K}, \mathbf{V}_W \in \mathbb{R}^{d \times d}$ and diagonal matrix $\mathbf{S}_W \in \mathbb{R}^{K \times d}$ with non-decreasing singular values. We denote r singular values of $\mathbf{W}$ as $\{s_k\}_{k=1}^r$ (we have $r \leq R := \min(K, d)$).

From equation (59) and the SVD of $\mathbf{W}$:

$$\begin{aligned}
\mathbf{H} &= (\mathbf{W}^\top \mathbf{W} + N\lambda_H \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{Y} \\
&= (\mathbf{V}_W \mathbf{S}_W^\top \mathbf{S}_W \mathbf{V}_W^\top + N\lambda_H \mathbf{I})^{-1} \mathbf{V}_W \mathbf{S}_W^\top \mathbf{U}_W^\top \mathbf{Y}. \\
&= \mathbf{V}_W (\mathbf{S}_W^\top \mathbf{S}_W + N\lambda_H \mathbf{I})^{-1} \mathbf{S}_W^\top \mathbf{U}_W^\top \mathbf{Y} \\
&= \mathbf{V}_W \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{s_1}{s_1^2 + N\lambda_{H_1}}, \ldots, \frac{s_r}{s_r^2 + N\lambda_{H_1}}\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\mathbf{C} \in \mathbb{R}^{d \times K}} \mathbf{U}_W^\top \mathbf{Y} \\
&= \mathbf{V}_W \mathbf{C} \mathbf{U}_W^\top \mathbf{Y},
\end{aligned} \tag{60}$$

$$\begin{aligned}
\mathbf{W}\mathbf{H} &= \mathbf{U}_W \mathbf{S}_W \begin{bmatrix} \operatorname{diag}\left(\frac{s_1}{s_1^2 + N\lambda_{H_1}}, \ldots, \frac{s_r}{s_r^2 + N\lambda_{H_1}}\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}_W^\top \mathbf{Y} \\
&= \mathbf{U}_W \operatorname{diag}\left(\frac{s_1^2}{s_1^2 + N\lambda_H}, \ldots, \frac{s_r^2}{s_r^2 + N\lambda_H}, 0, \ldots, 0\right) \mathbf{U}_W^\top \mathbf{Y}
\end{aligned} \tag{61}$$

$$\begin{aligned}
\Rightarrow \mathbf{W}\mathbf{H} - \mathbf{Y} &= \mathbf{U}_W \left[\operatorname{diag}\left(\frac{s_1^2}{s_1^2 + N\lambda_H}, \ldots, \frac{s_r^2}{s_r^2 + N\lambda_H}, 0, \ldots, 0\right) - \mathbf{I}_K\right] \mathbf{U}_W^\top \mathbf{Y} \\
&= \mathbf{U}_W \underbrace{\operatorname{diag}\left(\frac{-N\lambda_H}{s_1^2 + N\lambda_H}, \ldots, \frac{-N\lambda_H}{s_r^2 + N\lambda_H}, -1, \ldots, -1\right)}_{\mathbf{D} \in \mathbb{R}^{K \times K}} \mathbf{U}_W^\top \mathbf{Y} \\
&= \mathbf{U}_W \mathbf{D} \mathbf{U}_W^\top \mathbf{Y}.
\end{aligned} \tag{62}$$

Based on this result, we now calculate the Frobenius norm of $\mathbf{W}\mathbf{H} - \mathbf{Y}$:

$$\begin{aligned}
\|\mathbf{W}\mathbf{H} - \mathbf{Y}\|_F^2 &= \|\mathbf{U}_W \mathbf{D} \mathbf{U}_W^\top \mathbf{Y}\|_F^2 = \operatorname{trace}(\mathbf{U}_W \mathbf{D} \mathbf{U}_W^\top \mathbf{Y} (\mathbf{U}_W \mathbf{D} \mathbf{U}_W^\top \mathbf{Y})^\top) \\
&= \operatorname{trace}(\mathbf{U}_W \mathbf{D} \mathbf{U}_W^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_W \mathbf{D} \mathbf{U}_W^\top) = \operatorname{trace}(\mathbf{D}^2 \mathbf{U}_W^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_W).
\end{aligned} \tag{63}$$

We denote $\mathbf{u}^k$ and $\mathbf{u}_k$ are the k-th row and column of $\mathbf{U}_W$, respectively. Let $\mathbf{n} = (n_1, \ldots, n_K)$, we have the following:

$$\mathbf{U}_W = \begin{bmatrix} -\mathbf{u}^1- \\ \ldots \\ -\mathbf{u}^K- \end{bmatrix} = \begin{bmatrix} | & | & | \\ \mathbf{u}_1 & \ldots & \mathbf{u}_K \\ | & | & | \end{bmatrix},$$

$$\mathbf{Y}\mathbf{Y}^\top = \mathrm{diag}(n_1, n_2, \ldots, n_K) \in \mathbb{R}^{K \times K}$$

$$\begin{aligned} \Rightarrow \mathbf{U}_W^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_W &= \begin{bmatrix} | & | & | \\ (\mathbf{u}^1)^\top & \ldots & (\mathbf{u}^K)^\top \\ | & | & | \end{bmatrix} \mathrm{diag}(n_1, n_2, \ldots, n_K) \begin{bmatrix} -\mathbf{u}^1- \\ \ldots \\ -\mathbf{u}^K- \end{bmatrix} \\ &= \begin{bmatrix} | & | & | \\ (\mathbf{u}^1)^\top & \ldots & (\mathbf{u}^K)^\top \\ | & | & | \end{bmatrix} \begin{bmatrix} -n_1\mathbf{u}^1- \\ \ldots \\ -n_k\mathbf{u}^K- \end{bmatrix} \\ \Rightarrow (\mathbf{U}_W^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_W)_{kk} &= n_1 u_{1k}^2 + n_2 u_{2k}^2 + \ldots + n_k u_{Kk}^2 = (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} \\ \Rightarrow \|\mathbf{W}\mathbf{H} - \mathbf{Y}\|_F^2 = \mathrm{trace}(\mathbf{D}^2 \mathbf{U}_W^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_W) &= \sum_{k=1}^{r} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} \frac{(-N\lambda_H)^2}{(s_k^2 + N\lambda_H)^2} + \sum_{h=r+1}^{K} (\mathbf{u}_h \odot \mathbf{u}_h)^\top \mathbf{n}, \end{aligned} \tag{64}$$

where the last equality is from the fact that $\mathbf{D}^2$ is a diagonal matrix, so the diagonal of $\mathbf{D}^2 \mathbf{U}_W^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_W$ is the element-wise product between the diagonal of $\mathbf{D}^2$ and $\mathbf{U}_W^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_W$.

Similarly, we calculate the Frobenius norm of $\mathbf{H}$, from equation (60), we have:

$$\begin{aligned} \|\mathbf{H}\|_F^2 &= \mathrm{trace}(\mathbf{V}_W \mathbf{C} \mathbf{U}_W^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_W \mathbf{C}^\top \mathbf{V}_W^\top) = \mathrm{trace}(\mathbf{C}^\top \mathbf{C} \mathbf{U}_W^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_W) \\ &= \sum_{k=1}^{K} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} \frac{s_k^2}{(s_k^2 + N\lambda_H)^2}. \end{aligned} \tag{65}$$

Now, we plug the equations (64) and (65) into the function f, we get:

$$\begin{aligned} f(\mathbf{W}, \mathbf{H}) &= \frac{1}{2N} \sum_{k=1}^{r} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} \frac{(-N\lambda_H)^2}{(s_k^2 + N\lambda_H)^2} + \frac{1}{2N} \sum_{h=r+1}^{K} (\mathbf{u}_h \odot \mathbf{u}_h)^\top \mathbf{n} + \frac{\lambda_W}{2} \sum_{k=1}^{r} s_k^2 \\ &\quad + \frac{\lambda_H}{2} \sum_{k=1}^{K} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} \frac{s_k^2}{(s_k^2 + N\lambda_H)^2} \\ &= \frac{\lambda_H}{2} \sum_{k=1}^{r} \frac{(\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n}}{s_k^2 + N\lambda_H} + \frac{\lambda_W}{2} \sum_{k=1}^{r} s_k^2 + \frac{1}{2N} \sum_{h=r+1}^{K} (\mathbf{u}_h \odot \mathbf{u}_h)^\top \mathbf{n} \\ &= \frac{1}{2N} \sum_{k=1}^{r} \left(\frac{(\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n}}{\frac{s_k^2}{N\lambda_H} + 1} + N^2 \lambda_W \lambda_H \left(\frac{s_k^2}{N\lambda_H} \right) \right) + \frac{1}{2N} \sum_{h=r+1}^{K} (\mathbf{u}_h \odot \mathbf{u}_h)^\top \mathbf{n} \\ &= \frac{1}{2N} \sum_{k=1}^{r} \left(\frac{(\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n}}{x_k + 1} + b x_k \right) + \frac{1}{2N} \sum_{h=r+1}^{K} (\mathbf{u}_h \odot \mathbf{u}_h)^\top \mathbf{n} \\ &= \frac{1}{2N} \sum_{k=1}^{r} \left(\frac{a_k}{x_k + 1} + b x_k \right) + \frac{1}{2N} \sum_{h=r+1}^{K} a_h, \end{aligned} \tag{66}$$

with $x_k := \frac{s_k^2}{N\lambda_H}$, $a_k := (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n}$ and $b := N^2 \lambda_W \lambda_H$.

From the fact that $\mathbf{U}_W$ is an orthonormal matrix, we have:

$$\sum_{k=1}^{K} a_k = \sum_{k=1}^{K} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} = \left(\sum_{k=1}^{K} \mathbf{u}_k \odot \mathbf{u}_k \right)^\top \mathbf{n} = \mathbf{1}^\top \mathbf{n} = \sum_{k=1}^{K} n_k = N, \tag{67}$$

and, for any $j \in [K]$, denote $p_{i,j} := u_{i1}^2 + u_{i2}^2 + \ldots + u_{ij}^2 \; \forall \, i \in [K]$, we have:

$$
\begin{aligned}
\sum_{k=1}^{j} a_k &= \sum_{k=1}^{j} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} = n_1(u_{11}^2 + u_{12}^2 + \ldots + u_{1j}^2) + n_2(u_{21}^2 + u_{22}^2 + \ldots + u_{2j}^2) + \ldots + n_K(u_{K1}^2 + u_{K2}^2 + \ldots + u_{Kj}^2) \\
&= \sum_{k=1}^{K} p_{k,j} n_k \leq p_{1,j} n_1 + p_{2,j} n_2 + \ldots + p_{j,j} n_j + (p_{j+1,j} + p_{j+2,j} + \ldots + p_{K,j}) n_j \\
&= p_{1,j} n_1 + p_{2,j} n_2 + \ldots + p_{j-1,j} n_{j-1} + (j - p_{1,j} - \ldots - p_{j-1,j}) n_j \\
&= \sum_{k=1}^{j} n_k + \sum_{h=1}^{j-1} (n_h - n_j)(p_{h,j} - 1) \leq \sum_{k=1}^{j} n_k \\
\Rightarrow \sum_{k=j+1}^{K} a_k &\geq N - \sum_{k=1}^{j} n_k = \sum_{k=j+1}^{K} n_k \quad \forall \, j \in [K],
\end{aligned} \tag{68}
$$

where we used the fact that $\sum_{k=1}^{K} p_{k,j} = j$ since it is the sum of squares of all entries of the first j columns of an orthonormal matrix, and $p_{i,j} \leq 1 \; \forall \, i$ because it is the sum of squares of some entries on the i-th row of $\mathbf{U}_W$.

We state a lemma regarding minimizing a weighted sum as following.

Lemma E.3. *Consider a weighted sum $\sum_{k=1}^{K} a_k z_k$ with $\{a_k\}_{k=1}^{K}$ satisfies (67) and (68) and $0 < z_1 \leq z_2 \leq \ldots \leq z_K$. Then, we have:*

$$
\min_{a_1, \ldots, a_K} \sum_{k=1}^{K} a_k z_k = \sum_{k=1}^{K} n_k z_k.
$$

The equality happens when for any $k \geq 1$, $z_{k+1} = z_k$ or $a_{k+1} + a_{k+2} + \ldots + a_K = n_{k+1} + n_{k+2} + \ldots + n_K$ (equivalently, $a_1 + a_2 + \ldots + a_k = n_1 + n_2 + \ldots + n_k$).

Proof of Lemma E.3. We have:

$$
\begin{aligned}
\sum_{k=1}^{K} a_k z_k &= (a_1 + a_2 + \ldots + a_K) z_1 + (a_2 + \ldots + a_K)(z_2 - z_1) + \ldots + (a_{K-1} + a_K)(z_{K-1} - z_{K-2}) + a_K(z_K - z_{K-1}) \\
&\geq (n_1 + n_2 + \ldots + n_K) z_1 + (n_2 + \ldots + n_K)(z_2 - z_1) + \ldots + (n_{K-1} + n_K)(z_{K-1} - z_{K-2}) + n_K(z_K - z_{K-1}) \\
&= \sum_{k=1}^{K} n_k z_k.
\end{aligned}
$$

□

By applying Lemma E.3 to the RHS of equation (66) with $z_k = \frac{1}{x_k + 1} \; \forall \, k \leq r$ and $z_k = 1$ otherwise, we obtain:

$$
\begin{aligned}
f(\mathbf{W}, \mathbf{H}) &\geq \frac{1}{2N} \sum_{k=1}^{r} \left(\frac{n_k}{x_k + 1} + b x_k \right) + \frac{1}{2N} \sum_{h=r+1}^{K} n_h && (69) \\
&= \frac{1}{2N} \sum_{k=1}^{r} n_k \left(\frac{1}{x_k + 1} + \frac{b}{n_k} x_k \right) + \frac{1}{2N} \sum_{h=r+1}^{K} n_h. && (70)
\end{aligned}
$$

Consider the function:

$$
g(x) = \frac{1}{x + 1} + a x \;\text{ with } x \geq 0, a > 0. \tag{71}
$$

We consider two cases:

- If $a > 1$, $g(0) = 1$ and $g(x) > g(0)\, \forall x > 0$. Hence, $g(x)$ is minimized at $x = 0$ in this case.
- If $a \leq 1$, by using AM-GM, we have $g(x) = \frac{1}{x+1} + a(x+1) - a \geq 2\sqrt{a} - a$ with the equality holds iff $x = \sqrt{\frac{1}{a}} - 1$.

By applying this result to each term in the lower bound (70), we finish bounding $f(\mathbf{W}, \mathbf{H})$.

Now, we study the equality conditions. In the lower bound (70), by letting x_k^* be the minimizer of $\frac{1}{x_k+1} + \frac{b}{n_k} x_k$ for all $k \leq r$ and $x_k^* = 0$ for all $k > r$, there are only four possibilities as following:

- **Case A:** If $x_1^* > 0$ and $n_1 > n_2$: we have $x_1^* = \sqrt{\frac{n_1}{b}} - 1 > \max(0, \sqrt{\frac{n_2}{b}} - 1) \geq x_2^*$ and therefore from the equality condition of Lemma E.3, we have $a_1 = n_1$. From the orthonormal property of $\mathbf{u}_k$, we have:

$$a_1 = (\mathbf{u}_1 \odot \mathbf{u}_1)^\top \mathbf{n} = n_1 u_{11}^2 + n_2 u_{21}^2 + \ldots + n_k u_{K1}^2 \leq n_1(u_{11}^2 + u_{21}^2 + \ldots + u_{K1}^2) = n_1.$$

The equality holds when and only when $u_{11}^2 = 1$ and $u_{21} = \ldots = u_{K1} = 0$.

- **Case B:** If $x_1^* > 0$ and there exists $1 < j \leq r$ such that $n_1 = n_2 = \ldots = n_j > n_{j+1}$, we have:

$$\frac{1}{x+1} + \frac{b}{n_1} x = \frac{1}{x+1} + \frac{b}{n_2} x = \ldots = \frac{1}{x+1} + \frac{b}{n_j} x,$$

and thus, $x_1^* = x_2^* = \ldots = x_j^* > x_{j+1}^*$. Hence, from the equality condition of Lemma E.3, we have $a_1 + a_2 + \ldots + a_j = n_1 + \ldots + n_j$. We have:

$$\sum_{k=1}^{j} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} = n_1(u_{11}^2 + u_{12}^2 + \ldots + u_{1j}^2) + n_2(u_{21}^2 + u_{22}^2 + \ldots + u_{2j}^2)$$
$$+ \ldots + n_K(u_{K1}^2 + u_{K2}^2 + \ldots + u_{Kj}^2) \leq \sum_{k=1}^{j} n_k,$$

where the inequality is from the fact that for any $k \in [K]$, $(u_{k1}^2 + u_{k2}^2 + \ldots + u_{kj}^2) \leq 1$ and $\sum_{k=1}^{K} (u_{k1}^2 + u_{k2}^2 + \ldots + u_{kj}^2) = j$ and $n_j > n_{j+1}$. The equality holds iff $u_{k1}^2 + u_{k2}^2 + \ldots + u_{kj}^2 = 1\, \forall\, k = 1, 2, \ldots, j$ and $u_{k1} = u_{k2} = \ldots = u_{kj} = 0\, \forall\, k = j+1, \ldots, K$, i.e. the upper left sub-matrix size $j \times j$ of $\mathbf{U}_W$ is an orthonormal matrix and other entries of $\mathbf{U}_W$ lie on the same rows or columns with this sub-matrix must all equal 0's.

- **Case C:** If $x_1^* > 0$, $r < K$ and there exists $r < j \leq K$ such that $n_1 = n_2 = \ldots = n_r = \ldots = n_j > n_{j+1}$, thus we have $x_1^* = x_2^* = \ldots = x_r^* > 0$ and $x_{r+1}^* = \ldots = x_K^* = 0$. Hence, from the equality condition of Lemma E.3, we have $a_1 + a_2 + \ldots + a_r = n_1 + \ldots + n_r$. We have:

$$\sum_{k=1}^{r} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} = n_1(u_{11}^2 + u_{12}^2 + \ldots + u_{1r}^2) + n_2(u_{21}^2 + u_{22}^2 + \ldots + u_{2r}^2)$$
$$+ \ldots + n_K(u_{K1}^2 + u_{K2}^2 + \ldots + u_{Kr}^2) \leq \sum_{k=1}^{r} n_k,$$

where the inequality is from the fact that for any $k \in [K]$, $(u_{k1}^2 + u_{k2}^2 + \ldots + u_{kr}^2) \leq 1$ and $\sum_{k=1}^{K} (u_{k1}^2 + u_{k2}^2 + \ldots + u_{kr}^2) = r$. The equality holds iff $u_{k1} = u_{k2} = \ldots = u_{kr} = 0\, \forall\, k = j+1, \ldots, K$, i.e., the upper left sub-matrix size $j \times r$ of $\mathbf{U}_W$ includes r orthonormal vectors in $\mathbb{R}^j$ and the bottom left sub-matrix size $(K - j) \times r$ are all zeros. The other $K - r$ columns of $\mathbf{U}_W$ does not matter because $\mathbf{W}^*$ can be written as:

$$\mathbf{W}^* = \sum_{k=1}^{r} s_k^* \mathbf{u}_k \mathbf{v}_k^\top,$$

with $\mathbf{v}_k$ is the right singular vector that satisfies $\mathbf{W}^{*\top} \mathbf{u}_k = s_k^* \mathbf{v}_k$. Note that since $s_1^* = s_2^* = \ldots = s_r^* := s^*$, we have the compact SVD form as follows:

$$\mathbf{W}^* = s^* \mathbf{U}_W^{'} \mathbf{V}_W^{'\top}, \tag{72}$$

where $\mathbf{U}_W^{'} \in \mathbb{R}^{K \times r}$ and $\mathbf{V}_W^{'} \in \mathbb{R}^{d \times r}$. Especially, the last $K-j$ rows of $\mathbf{W}^*$ will be zeros since the last $K-j$ rows of $\mathbf{U}_W^{'}$ are zeros. Furthermore, tbhe matrix $\mathbf{U}_W^{'}\mathbf{U}_W^{'\top}$ after removing the last $K-j$ zero rows and the last $K-j$ zero columns is the best rank-r approximation of $\mathbf{I}_j$.

We note that if **Case C** happens, then the number of positive singular values are limited by the matrix rank r (e.g., by $r \leq R = \min(d, K) = d$ when $d < K$), and $n_r = n_{r+1}$, thus $x_r^* > 0$ and $x_{r+1}^* = 0$ (x_{r+1}^* should equal $x_r^* > 0$ if it is not forced to be zero).

- **Case D:** If $x_1^* = 0$, we must have $x_2^* = \ldots = x_K^* = 0$, $\sum_{k=1}^K (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n}$ always equal N and thus, $\mathbf{U}_W$ can be an arbitrary size $K \times K$ orthonormal matrix.

We perform similar arguments as above for all subsequent x_k^*'s, after we finish reasoning for prior ones. Before going to the conclusion, we first study the matrix $\mathbf{U}_W$. If **Case C** does not happen for any x_k^*'s, we have:

$$\mathbf{U}_W = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_l \end{bmatrix}, \tag{73}$$

where each $\mathbf{A}_i$ is an orthonormal block which corresponds with one or a group of classes that have the same number of training samples and their $x^* > 0$ (**Case A** and **Case B**) or corresponds with all classes with $x^* = 0$ (**Case D**). If **Case C** happens, we have:

$$\mathbf{U}_W = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_l \end{bmatrix}, \tag{74}$$

where each $\mathbf{A}_i, i \in [l-1]$ is an orthonormal block which corresponds with one or a group of classes that have the same number of training samples and their $x^* > 0$ (**Case A** and **Case B**). $\mathbf{A}_l$ is the orthonormal block has the same property as $\mathbf{U}_W$ in **Case C**.

We consider the case $d \geq K$ from now on. By using arguments about the minimizer of $g(x)$ applied to the lower bound (70), we consider three cases as following:

- **Case 1a:** $\frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_K} \leq 1$.

  Then, the lower bound (70) is minimized at $(x_1^*, x_2^*, \ldots, x_K^*) = \left(\sqrt{\frac{n_1}{b}} - 1, \sqrt{\frac{n_2}{b}} - 1, \ldots, \sqrt{\frac{n_K}{b}} - 1\right)$. Therefore:

$$(s_1^*, s_2^*, \ldots, s_K^*) = \left(\sqrt{\sqrt{\frac{n_1 \lambda_H}{\lambda_W}} - N\lambda_H}, \sqrt{\sqrt{\frac{n_2 \lambda_H}{\lambda_W}} - N\lambda_H}, \ldots, \sqrt{\sqrt{\frac{n_K \lambda_H}{\lambda_W}} - N\lambda_H}\right). \tag{75}$$

  First, we have the property that the features in each class $\mathbf{h}_{k,i}^*$ collapsed to their class-mean $\mathbf{h}_k^*$ ($\mathcal{NC}1$). Let $\overline{\mathbf{H}}^* = \mathbf{V}_W \mathbf{C} \mathbf{U}_W^\top$, we know that $\mathbf{H}^* = \overline{\mathbf{H}}^* \mathbf{Y}$ from equation (60). Then, columns from the $(n_{k-1}+1)$-th until (n_k)-th of $\mathbf{H}$ will all equals the k-th column of $\overline{\mathbf{H}}^*$, thus the features in class k are collapsed to their class-mean $\mathbf{h}_k^*$ (which is the k-th column of $\overline{\mathbf{H}}^*$), i.e., $\mathbf{h}_{k,1}^* = \mathbf{h}_{k,2}^* = \ldots = \mathbf{h}_{k,n_k}^* \forall k \in [K]$.

  Case C never happens because if we assume we have $r < K$ positive singular values, meaning $s_r^* > 0$. Then, if $n_{r+1} = n_r$, we must have $s_{r+1}^* > 0$ (contradiction!). Hence, $\mathbf{U}_W$ must have the form as in equation (73), thus we can conclude the geometry of the following :

$$\mathbf{W}^*\mathbf{W}^{*\top} = \mathbf{U}_W\mathbf{S}_W\mathbf{S}_W^\top\mathbf{U}_W^\top = \operatorname{diag}\left\{\sqrt{\frac{n_1\lambda_H}{\lambda_W}} - N\lambda_H, \sqrt{\frac{n_2\lambda_H}{\lambda_W}} - N\lambda_H, \dots, \sqrt{\frac{n_K\lambda_H}{\lambda_W}} - N\lambda_H\right\} \in \mathbb{R}^{K\times K}, \tag{76}$$

$$\begin{aligned}
\mathbf{W}^*\mathbf{H}^* &= \mathbf{U}_W \operatorname{diag}\left\{\frac{s_1^2}{s_1^2+N\lambda_H}, \dots, \frac{s_K^2}{s_K^2+N\lambda_H}\right\}\mathbf{U}_W^\top\mathbf{Y} \\
&= \begin{bmatrix} \frac{s_1^2}{s_1^2+N\lambda_H} & 0 & \dots & 0 \\ 0 & \frac{s_2^2}{s_2^2+N\lambda_H} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{s_K^2}{s_K^2+N\lambda_H} \end{bmatrix} \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 & \dots & 1 \end{bmatrix} \\
&= \begin{bmatrix} \frac{s_1^2}{s_1^2+N\lambda_H}\mathbf{1}_{n_1}^\top & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \frac{s_K^2}{s_K^2+N\lambda_H}\mathbf{1}_{n_K}^\top \end{bmatrix}, \\
\mathbf{H}^{*\top}\mathbf{H}^* &= \mathbf{Y}^\top\mathbf{U}_W\mathbf{C}^T\mathbf{C}\mathbf{U}_W^\top\mathbf{Y} \\
&= \mathbf{Y}^\top \begin{bmatrix} \frac{s_1^2}{(s_1^2+N\lambda_H)^2} & 0 & \dots & 0 \\ 0 & \frac{s_2^2}{(s_2^2+N\lambda_H)^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{s_K^2}{(s_K^2+N\lambda_H)^2} \end{bmatrix} \mathbf{Y} \\
&= \begin{bmatrix} \frac{s_1^2}{(s_1^2+N\lambda_H)^2}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \frac{s_2^2}{(s_2^2+N\lambda_H)^2}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \frac{s_K^2}{(s_K^2+N\lambda_H)^2}\mathbf{1}_{n_K}\mathbf{1}_{n_K}^\top \end{bmatrix} \in \mathbb{R}^{N\times N},
\end{aligned} \tag{77}$$

where $\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top$ is a $n_k \times n_k$ matrix will all entries are 1's.

We additionally have the structure of the class-means matrix:

$$\overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* = \mathbf{U}_W^\top\mathbf{C}^\top\mathbf{C}\mathbf{U}_W = \begin{bmatrix} \frac{s_1^2}{(s_1^2+N\lambda_H)^2} & 0 & \dots & 0 \\ 0 & \frac{s_2^2}{(s_2^2+N\lambda_H)^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{s_K^2}{(s_K^2+N\lambda_H)^2} \end{bmatrix} \in \mathbb{R}^{K\times K}, \tag{78}$$

$$\mathbf{W}^*\overline{\mathbf{H}}^* = \mathbf{U}_W\mathbf{S}_W\mathbf{C}\mathbf{U}_\mathbf{W}{}^\top = \begin{bmatrix} \frac{s_1^2}{s_1^2+N\lambda_H} & 0 & \dots & 0 \\ 0 & \frac{s_2^2}{s_2^2+N\lambda_H} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{s_K^2}{s_K^2+N\lambda_H} \end{bmatrix} \in \mathbb{R}^{K\times K}. \tag{79}$$

And the alignment between the linear classifier and features are as following. For any $k \in [K]$, denote $\mathbf{w}_k$ the k-th row of $\mathbf{W}^*$:

$$\begin{aligned}\mathbf{W}^* &= \mathbf{U}_W \mathbf{S}_W \mathbf{V}_W^\top, \\ \overline{\mathbf{H}}^* &= \mathbf{V}_W \mathbf{C} \mathbf{U}_W^\top \\ \Rightarrow \mathbf{w}_k^* &= (s_k^2 + N\lambda_H)\mathbf{h}_k^* = \sqrt{\frac{n_k \lambda_H}{\lambda_W}} \mathbf{h}_k^*. \end{aligned} \tag{80}$$

- **Case 2a:** There exists $j \in [K-1]$ s.t. $\frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_j} \leq 1 < \frac{b}{n_{j+1}} \leq \ldots \leq \frac{b}{n_K}$

Then, the lower bound (70) is minimized at:

$$(s_1^*, \ldots, s_j^*, s_{j+1}^* \ldots, s_K^*) = \left(\sqrt{\sqrt{\frac{n_1 \lambda_H}{\lambda_W}} - N\lambda_H}, \ldots, \sqrt{\sqrt{\frac{n_j \lambda_H}{\lambda_W}} - N\lambda_H}, 0, \ldots, 0 \right). \tag{81}$$

First, we have the property that the features in each class $\mathbf{h}_{k,i}^*$ collapsed to their class-mean $\mathbf{h}_k^*$ ($\mathcal{NC}1$). Let $\overline{\mathbf{H}}^* = \mathbf{V}_W \mathbf{C} \mathbf{U}_W^\top$, we know that $\mathbf{H}^* = \overline{\mathbf{H}}^*$ from equation (60). Then, columns from the $(n_{k-1}+1)$-th until (n_k)-th of $\mathbf{H}^*$ will all equals the k-th column of $\overline{\mathbf{H}}^*$, thus the features in class k are collapsed to their class-mean $\mathbf{h}_k^*$ (which is the k-th column of $\overline{\mathbf{H}}$), i.e $\mathbf{h}_{k,1}^* = \mathbf{h}_{k,2}^* = \ldots = \mathbf{h}_{k,n_k}^* \; \forall k \in [K]$.

Recall $\mathbf{U}_W$ with the form (73) (**Case C** cannot happen with the same reason as in **Case 1a**). From equations (60) and (62), we can conclude the geometry of the following:

$$\begin{aligned}\mathbf{W}^* \mathbf{W}^{*\top} &= \mathbf{U}_W \mathbf{S}_W \mathbf{S}_W^\top \mathbf{U}_W^\top \\ &= \operatorname{diag}\left(\sqrt{\frac{n_1 \lambda_H}{\lambda_W}} - N\lambda_H, \sqrt{\frac{n_2 \lambda_H}{\lambda_W}} - N\lambda_H, \ldots, \sqrt{\frac{n_j \lambda_H}{\lambda_W}} - N\lambda_H, 0, \ldots, 0 \right), \end{aligned} \tag{82}$$

$$\begin{aligned}\mathbf{W}^* \mathbf{H}^* &= \mathbf{U}_W \operatorname{diag}\left(\frac{s_1^2}{s_1^2 + N\lambda_H}, \ldots, \frac{s_j^2}{s_j^2 + N\lambda_H}, 0, \ldots, 0 \right) \mathbf{U}_W^\top \mathbf{Y} \\ &= \begin{bmatrix} \frac{s_1^2}{s_1^2+N\lambda_H} \mathbf{1}_{n_1}^\top & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \frac{s_2^2}{s_2^2+N\lambda_H} \mathbf{1}_{n_2}^\top & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0}_{n_K}^\top \end{bmatrix} \in \mathbb{R}^{K \times N}, \\ \mathbf{H}^{*\top} \mathbf{H}^* &= \begin{bmatrix} \frac{s_1^2}{(s_1^2+N\lambda_H)^2} \mathbf{1}_{n_1} \mathbf{1}_{n_1}^\top & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \frac{s_2^2}{(s_2^2+N\lambda_H)^2} \mathbf{1}_{n_2} \mathbf{1}_{n_2}^\top & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0}_{n_K \times n_K} \end{bmatrix} \in \mathbb{R}^{N \times N}, \end{aligned} \tag{83}$$

where $\mathbf{1}_{n_k} \mathbf{1}_{n_k}^\top$ is a $n_k \times n_k$ matrix will all entries are 1's.

For any $k \in [K]$, denote $\mathbf{w}_k^*$ the k-th row of $\mathbf{W}^*$ and $\mathbf{v}_k$ the k-th column of $\mathbf{V}_W$, we have:

$$\begin{aligned}\mathbf{W}^* &= \mathbf{U}_W \mathbf{S}_W \mathbf{V}_W^\top, \\ \overline{\mathbf{H}}^* &= \mathbf{V}_W \mathbf{C} \mathbf{U}_W^\top \\ \Rightarrow \mathbf{w}_k^* &= (s_k^2 + N\lambda_H)\mathbf{h}_k^* = \sqrt{\frac{n_k \lambda_H}{\lambda_W}} \mathbf{h}_k^*. \end{aligned} \tag{84}$$

And, for $k > j$, we have $\mathbf{w}_k^* = \mathbf{h}_k^* = \mathbf{0}$, which means the optimal classifiers and features of class $k > j$ will be $\mathbf{0}$.

- **Case 3a:** $1 < \frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_R}$

  Then, the lower bound (70) is minimized at:

$$(s_1^*, s_2^*, \ldots, s_K^*) = (0, 0, \ldots, 0). \tag{85}$$

  Hence, the global minimizer of f in this case is $(\mathbf{W}^*, \mathbf{H}^*) = (\mathbf{0}, \mathbf{0})$.

Now, we turn to consider the case $d < K$, and thus, $r \leq R = d < K$. Again, we consider the following cases:

- **Case 1b:** $\frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_R} \leq 1$.

  Then, the lower bound (70) is minimized at $(x_1^*, x_2^*, \ldots, x_K^*) = (\sqrt{\frac{n_1}{b}} - 1, \sqrt{\frac{n_2}{b}} - 1, \ldots, \sqrt{\frac{n_R}{b}} - 1, 0, \ldots, 0) = (\sqrt{\frac{n_1}{N^2 \lambda_W \lambda_H}} - 1, \sqrt{\frac{n_2}{N^2 \lambda_W \lambda_H}} - 1, \ldots, \sqrt{\frac{n_R}{N^2 \lambda_W \lambda_H}} - 1, 0, \ldots, 0)$. Therefore:

$$\begin{aligned}
&(s_1^*, s_2^*, \ldots, s_R^*, s_{R+1}^*, \ldots s_K^*) \\
&= \left(\sqrt{\sqrt{\frac{n_1 \lambda_H}{\lambda_W}} - N\lambda_H}, \sqrt{\sqrt{\frac{n_2 \lambda_H}{\lambda_W}} - N\lambda_H}, \ldots, \sqrt{\sqrt{\frac{n_R \lambda_H}{\lambda_W}} - N\lambda_H}, 0, \ldots, 0 \right).
\end{aligned} \tag{86}$$

  We have $(\mathcal{NC}1)$ and $(\mathcal{NC}3)$ properties are the same as **Case 1a**.

  We have **Case C** happens iff $b/n_R < 1$ (i.e., $x_R^* > 0$) and $n_R = n_{R+1}$. Then, if $b/n_R = 1$ or $n_R > n_{R+1}$, we have:

$$\mathbf{W}^* \mathbf{W}^{*\top} = \mathbf{U}_W \mathbf{S}_W \mathbf{S}_W^\top \mathbf{U}_W^\top = \begin{bmatrix} \sqrt{\frac{n_1 \lambda_H}{\lambda_W}} - N\lambda_H & \ldots & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \ldots & \sqrt{\frac{n_R \lambda_H}{\lambda_W}} - N\lambda_H & \ldots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & \ldots & 0 \end{bmatrix} \in \mathbb{R}^{K \times K}, \tag{87}$$

$$\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* = \mathbf{U}_W^\top \mathbf{C}^\top \mathbf{C} \mathbf{U}_W = \begin{bmatrix} \frac{s_1^2}{(s_1^2 + N\lambda_H)^2} & 0 & \ldots & 0 \\ 0 & \frac{s_2^2}{(s_2^2 + N\lambda_H)^2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix} \in \mathbb{R}^{K \times K}, \tag{88}$$

$$\mathbf{W}^* \overline{\mathbf{H}}^* = \mathbf{U}_W \mathbf{S}_W \mathbf{C} \mathbf{U_W}^\top = \begin{bmatrix} \frac{s_1^2}{s_1^2 + N\lambda_H} & 0 & \ldots & 0 \\ 0 & \frac{s_2^2}{s_2^2 + N\lambda_H} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix} \in \mathbb{R}^{K \times K}. \tag{89}$$

  Furthermore, we have $\mathbf{w}_k^* = \mathbf{h}_k^* = \mathbf{0}$ for $k > R$.

  If **Case C** happens, there exists $k \leq R$, $l > R$ such that $n_{k-1} > n_k = n_{k+1} = \ldots = n_R = \ldots = n_l > n_{l+1}$. Recall the form of $\mathbf{U}_W$ as in equation (74), then:

$$\mathbf{W}^*\mathbf{W}^{*\top} = \begin{bmatrix} \sqrt{\frac{n_1\lambda_H}{\lambda_W}} - N\lambda_H & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & \sqrt{\frac{n_{k-1}\lambda_H}{\lambda_W}} - N\lambda_H & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \left(\sqrt{\frac{n_k\lambda_H}{\lambda_W}} - N\lambda_H\right)\mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{90}$$

$$\overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* = \begin{bmatrix} \frac{s_1^2}{(s_1^2+N\lambda_H)^2} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & \frac{s_{k-1}^2}{(s_{k-1}^2+N\lambda_H)^2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \frac{s_k^2}{(s_k^2+N\lambda_H)^2}\mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{91}$$

$$\mathbf{W}^*\overline{\mathbf{H}}^* = \begin{bmatrix} \frac{s_1^2}{s_1^2+N\lambda_H} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & \frac{s_{k-1}^2}{s_{k-1}^2+N\lambda_H} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \frac{s_k^2}{s_k^2+N\lambda_H}\mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{92}$$

and for any $k > l > R$, we have $\mathbf{w}_k^* = \mathbf{h}_k^* = \mathbf{0}$.

- **Case 2b:** There exists $j \in [R-1]$ s.t. $\frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_j} \leq 1 < \frac{b}{n_{j+1}} \leq \ldots \leq \frac{b}{n_R}$

Then, the lower bound (70) is minimized at:

$$(s_1^*, \ldots, s_j^*, s_{j+1}^* \ldots, s_K^*) = \left(\sqrt{\sqrt{\frac{n_1\lambda_H}{\lambda_W}} - N\lambda_H}, \ldots, \sqrt{\sqrt{\frac{n_j\lambda_H}{\lambda_W}} - N\lambda_H}, 0, \ldots, 0\right). \tag{93}$$

We have $(\mathcal{NC}1)$ and $(\mathcal{NC}3)$ properties are the same as **Case 2a**.

Case C does not happen in this case because $b/n_R > 1$ and thus, $x_R^* = 0$. Thus, we can conclude the geometry of the following:

$$\begin{aligned} \mathbf{W}^*\mathbf{W}^{*\top} &= \mathbf{U}_W\mathbf{S}_W\mathbf{S}_W^\top\mathbf{U}_W^\top \\ &= \operatorname{diag}\left(\sqrt{\frac{n_1\lambda_H}{\lambda_W}} - N\lambda_H, \sqrt{\frac{n_2\lambda_H}{\lambda_W}} - N\lambda_H, \ldots, \sqrt{\frac{n_j\lambda_H}{\lambda_W}} - N\lambda_H, 0, \ldots, 0\right), \end{aligned} \tag{94}$$

$$\begin{aligned} \mathbf{W}^*\mathbf{H}^* &= \mathbf{U}_W \operatorname{diag}\left(\frac{s_1^2}{s_1^2+N\lambda_H}, \ldots, \frac{s_j^2}{s_j^2+N\lambda_H}, 0, \ldots, 0\right)\mathbf{U}_W^\top\mathbf{Y} \\ &= \begin{bmatrix} \frac{s_1^2}{s_1^2+N\lambda_H}\mathbf{1}_{n_1}^\top & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \frac{s_2^2}{s_2^2+N\lambda_H}\mathbf{1}_{n_2}^\top & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0}_{n_K}^\top \end{bmatrix} \in \mathbb{R}^{K\times N}, \end{aligned}$$

$$\mathbf{H}^{*\top}\mathbf{H}^{*} = \begin{bmatrix} \frac{s_1^2}{(s_1^2+N\lambda_H)^2}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \frac{s_2^2}{(s_2^2+N\lambda_H)^2}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0}_{n_K\times n_K} \end{bmatrix} \in \mathbb{R}^{N\times N}, \tag{95}$$

where $\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top$ is a $n_k \times n_k$ matrix will all entries are 1's. And for any $k > j$, $\mathbf{w}_k^* = \mathbf{h}_k^* = \mathbf{0}$.

- **Case 3b:** $1 < \frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_R}$

Then, the lower bound (70) is minimized at:

$$(s_1^*, s_2^*, \ldots, s_K^*) = (0, 0, \ldots, 0). \tag{96}$$

Hence, the global minimizer of f in this case is $(\mathbf{W}^*, \mathbf{H}^*) = (\mathbf{0}, \mathbf{0})$.

□

F. Proof of Theorem 4.4

Theorem F.1. *Let $d_m \geq K \; \forall \, m \in [M]$ and $(\mathbf{W}_M^*, \mathbf{W}_{M-1}^*, \ldots, \mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^*)$ be any global minimizer of problem (6). We have:*

($\mathcal{NC}1$) $\quad \mathbf{H}_1^* = \overline{\mathbf{H}}^*\mathbf{Y} \Leftrightarrow \mathbf{h}_{k,i}^* = \mathbf{h}_k^* \; \forall \, k \in [K], i \in [n_k]$, *where* $\overline{\mathbf{H}}^* = [\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*] \in \mathbb{R}^{d_1 \times K}$.

($\mathcal{NC}2$) *Let* $c := \frac{\lambda_{W_1}^{M-1}}{\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_2}}$, $a := N\sqrt[M]{N\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_1}\lambda_{H_1}}$ *and* $\forall k \in [K]$, x_k^* *is the largest positive solution of the equation* $\frac{a}{n_k} - \frac{x^{M-1}}{(x^M+1)^2} = 0$*, we have the following:*

$$\begin{aligned} \mathbf{W}_M^*\mathbf{W}_M^{*\top} &= \frac{\lambda_{W_1}}{\lambda_{W_M}} \operatorname{diag}\left\{s_k^2\right\}_{k=1}^K, \\ \overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* &= \operatorname{diag}\left\{\frac{cs_k^{2M}}{(cs_k^{2M}+N\lambda_{H_1})^2}\right\}_{k=1}^K, \\ \mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_1^*\mathbf{H}_1^* &= \left\{\frac{cs_k^{2M}}{cs_k^{2M}+N\lambda_{H_1}}\right\}_{k=1}^K \mathbf{Y}, \end{aligned}$$

($\mathcal{NC}3$) *We have,* $\forall \, k \in [K]$*:*

$$(\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*)_k = (cs_k^{2M} + N\lambda_{H_1})\mathbf{h}_k^*,$$

where:

- *If* $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_K} < \frac{(M-1)^{\frac{M-1}{M}}}{M^2}$*, we have:*

$$s_k = \sqrt[2M]{\frac{N\lambda_{H_1}x_k^{*M}}{c}} \quad \forall \, k.$$

- *If there exists a* $j \in [K-1]$ *s.t.* $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_j} < \frac{(M-1)^{\frac{M-1}{M}}}{M^2} < \frac{a}{n_{j+1}} \leq \ldots \leq \frac{a}{n_K}$*, we have:*

$$s_k = \begin{cases} \sqrt[2M]{\frac{N\lambda_{H_1}x_k^{*M}}{c}} & \forall \, k \leq j \\ 0 & \forall \, k > j \end{cases}.$$

And, for any k *such that* $s_k = 0$*, we have:*

$$(\mathbf{W}_M^*)_k = \mathbf{h}_k^* = \mathbf{0}.$$

- *If* $\frac{(M-1)^{\frac{M-1}{M}}}{M^2} < \frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_K}$*, we have:*

$$(s_1, s_2, \ldots, s_K) = (0, 0, \ldots, 0),$$

and $(\mathbf{W}_M^*, \ldots, \mathbf{W}_1^*, \mathbf{H}_1^*) = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{0})$ *in this case.*

The only case left is if there exists $i, j \in [K]$ $(i \leq j \leq K)$ *such that* $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_{i-1}} < \frac{a}{n_i} = \frac{a}{n_{i+1}} = \ldots = \frac{a}{n_j} = \frac{(M-1)^{\frac{M-1}{M}}}{M^2} < \frac{a}{n_{j+1}} \leq \frac{a}{n_{j+2}} \leq \ldots \leq \frac{a}{n_K}$*, we have:*

$$s_k = \begin{cases} \sqrt[2M]{N\lambda_{H_1} x_k^{*M}/c} \quad \forall\, k \leq i-1 \\ \sqrt[2M]{N\lambda_{H_1} x_k^{*M}/c} \text{ or } 0 \quad \forall\, i \leq k \leq j \\ 0 \quad \forall\, k \geq j+1 \end{cases},$$

furthermore, let r *is the largest index that* $s_r > 0$*, we must have* $s_{r+1} = s_{r+2} = \ldots = s_K = 0$. $(\mathcal{NC}1)$ *and* $(\mathcal{NC}3)$ *are the same as above but for* $(\mathcal{NC}2)$*:*

$$\mathbf{W}_M^* \mathbf{W}_M^{*\top} = \frac{\lambda_{W_1}}{\lambda_{W_M}} \begin{bmatrix} s_1^2 & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \ldots & s_{i-1}^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & s_i^2 \mathcal{P}_{r-i+1}(\mathbf{I}_{j-i+1}) & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-j)\times(K-j)} \end{bmatrix}, \tag{97}$$

$$\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* = \begin{bmatrix} \frac{cs_1^{2M}}{(cs_1^{2M}+N\lambda_{H_1})^2} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \ldots & \frac{cs_{i-1}^{2M}}{(cs_{i-1}^{2M}+N\lambda_{H_1})^2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \frac{cs_i^{2M}}{(cs_i^{2M}+N\lambda_{H_1})^2} \mathcal{P}_{r-i+1}(\mathbf{I}_{j-i+1}) & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-j)\times(K-j)} \end{bmatrix}, \tag{98}$$

$$\mathbf{W}_M^* \mathbf{W}_{M-1}^* \ldots \mathbf{W}_2^* \mathbf{W}_1^* \overline{\mathbf{H}}^* = \begin{bmatrix} \frac{cs_1^{2M}}{cs_1^{2M}+N\lambda_{H_1}} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \ldots & \frac{cs_{i-1}^{2M}}{cs_{i-1}^{2M}+N\lambda_{H_1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \frac{cs_i^{2M}}{cs_i^{2M}+N\lambda_{H_1}} \mathcal{P}_{r-i+1}(\mathbf{I}_{j-i+1}) & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-j)\times(K-j)} \end{bmatrix}, \tag{99}$$

and, for any $h > j$, $(\mathbf{W}_M^* \mathbf{W}_{M-1}^* \ldots \mathbf{W}_2^* \mathbf{W}_1^*)_h = \mathbf{h}_h^* = \mathbf{0}$.

Theorem F.2. *Let* $R = \min(d_M, \ldots, d_1, K) < K$ *and* $(\mathbf{W}_M^*, \mathbf{W}_{M-1}^*, \ldots, \mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^*)$ *be any global minimizer of problem* (6)*. We have:*

$(\mathcal{NC}1)$ $\quad \mathbf{H}_1^* = \overline{\mathbf{H}}^* \mathbf{Y} \Leftrightarrow \mathbf{h}_{k,i}^* = \mathbf{h}_k^* \; \forall\, k \in [K], i \in [n_k]$*, where* $\overline{\mathbf{H}}^* = [\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*] \in \mathbb{R}^{d_1 \times K}$.

$(\mathcal{NC}3)$ $\quad$ *We have,* $\forall\, k \in [K]$*:*

$$(\mathbf{W}_M^* \mathbf{W}_{M-1}^* \ldots \mathbf{W}_2^* \mathbf{W}_1^*)_k = (cs_k^{2M} + N\lambda_{H_1}) \mathbf{h}_k^*,$$

$(\mathcal{NC}2)$ $\quad$ *Let* $c := \frac{\lambda_{W_1}^{M-1}}{\lambda_{W_M} \lambda_{W_{M-1}} \ldots \lambda_{W_2}}$, $a := N \sqrt[M]{N\lambda_{W_M} \lambda_{W_{M-1}} \ldots \lambda_{W_1} \lambda_{H_1}}$ *and* $\forall k \in [K]$, x_k^* *is the largest positive solution of the equation* $\frac{a}{n_k} - \frac{x^{M-1}}{(x^M+1)^2} = 0$*, we define* $\{s_k\}_{k=1}^K$ *as follows:*

- *If* $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_R} < \frac{(M-1)^{\frac{M-1}{M}}}{M^2}$*, we have:*

$$s_k = \begin{cases} \sqrt[2M]{\frac{N\lambda_{H_1} x_k^{*M}}{c}} & \forall\, k \leq R \\ 0 \quad \forall\, k > R \end{cases} .$$

Then, if $n_R > n_{R+1}$*, we have:*

$$\begin{aligned} \mathbf{W}_M^* \mathbf{W}_M^{*\top} &= \frac{\lambda_{W_1}}{\lambda_{W_M}} \operatorname{diag}\left\{s_k^2\right\}_{k=1}^K , \\ \overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* &= \operatorname{diag}\left\{\frac{cs_k^{2M}}{(cs_k^{2M} + N\lambda_{H_1})^2}\right\}_{k=1}^K , \\ \mathbf{W}_M^* \mathbf{W}_{M-1}^* \ldots \mathbf{W}_1^* \overline{\mathbf{H}_1}^* &= \left\{\frac{cs_k^{2M}}{cs_k^{2M} + N\lambda_{H_1}}\right\}_{k=1}^K , \end{aligned}$$

and for any $k > R$*, we have* $(\mathbf{W}_M^* \mathbf{W}_{M-1}^* \ldots \mathbf{W}_2^* \mathbf{W}_1^*)_k = \mathbf{h}_k^* = \mathbf{0}$*.*

Otherwise, if $n_R = n_{R+1}$*, and there exists* $k \leq R$*,* $l > R$ *such that* $n_{k-1} > n_k = n_{k+1} = \ldots = n_R = \ldots = n_l > n_{l+1}$*, we have:*

$$\mathbf{W}_M^* \mathbf{W}_M^{*\top} = \frac{\lambda_{W_1}}{\lambda_{W_M}} \begin{bmatrix} s_1^2 & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \ldots & s_{k-1}^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & s_k^2 \mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{100}$$

$$\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* = \begin{bmatrix} \frac{cs_1^{2M}}{(cs_1^{2M}+N\lambda_{H_1})^2} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \ldots & \frac{cs_{k-1}^{2M}}{(cs_{k-1}^{2M}+N\lambda_{H_1})^2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \frac{cs_k^{2M}}{(cs_k^{2M}+N\lambda_{H_1})^2} \mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{101}$$

$$\mathbf{W}_M^* \mathbf{W}_{M-1}^* \ldots \mathbf{W}_2^* \mathbf{W}_1^* \overline{\mathbf{H}}^* = \begin{bmatrix} \frac{cs_1^{2M}}{cs_1^{2M}+N\lambda_{H_1}} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \ldots & \frac{cs_{k-1}^{2M}}{cs_{k-1}^{2M}+N\lambda_{H_1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \frac{cs_k^{2M}}{cs_k^{2M}+N\lambda_{H_1}} \mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{102}$$

and, for any $h > l > R$*,* $(\mathbf{W}_M^* \mathbf{W}_{M-1}^* \ldots \mathbf{W}_2^* \mathbf{W}_1^*)_h = \mathbf{h}_h^* = \mathbf{0}$*.*

- *If there exists a* $j \in [R-1]$ *s.t.* $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \ldots \leq \frac{a}{n_j} < \frac{(M-1)^{\frac{M-1}{M}}}{M^2} < \frac{a}{n_{j+1}} \leq \ldots \leq \frac{a}{n_R}$*, we have:*

$$s_k = \begin{cases} \sqrt[2M]{\frac{N\lambda_{H_1} x_k^{*M}}{c}} & \forall\, k \leq j \\ 0 \quad \forall\, k > j \end{cases} .$$

Then, we have:

$$\mathbf{W}_M^* \mathbf{W}_M^{*\top} = \frac{\lambda_{W_1}}{\lambda_{W_M}} \operatorname{diag}\left\{s_k^2\right\}_{k=1}^K,$$
$$\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* = \operatorname{diag}\left\{\frac{c s_k^{2M}}{(c s_k^{2M} + N\lambda_{H_1})^2}\right\}_{k=1}^K,$$
$$\mathbf{W}_M^* \mathbf{W}_{M-1}^* \dots \mathbf{W}_1^* \overline{\mathbf{H}_1}^* = \left\{\frac{c s_k^{2M}}{c s_k^{2M} + N\lambda_{H_1}}\right\}_{k=1}^K,$$

and for any $k > j$, we have $(\mathbf{W}_M^* \mathbf{W}_{M-1}^* \dots \mathbf{W}_2^* \mathbf{W}_1^*)_k = \mathbf{h}_k^* = \mathbf{0}$.

- *If $\frac{(M-1)^{\frac{M-1}{M}}}{M^2} < \frac{a}{n_1} \leq \frac{a}{n_2} \leq \dots \leq \frac{a}{n_R}$, we have:*

$$(s_1, s_2, \dots, s_K) = (0, 0, \dots, 0),$$

and $(\mathbf{W}_M^*, \dots, \mathbf{W}_1^*, \mathbf{H}_1^*) = (\mathbf{0}, \dots, \mathbf{0}, \mathbf{0})$ in this case.

The only case left is if there exists $i, j \in [R]$ ($i \leq j \leq R$) such that $\frac{a}{n_1} \leq \frac{a}{n_2} \leq \dots \leq \frac{a}{n_{i-1}} < \frac{a}{n_i} = \frac{a}{n_{i+1}} = \dots = \frac{a}{n_j} = \frac{(M-1)^{\frac{M-1}{M}}}{M^2} < \frac{a}{n_{j+1}} \leq \frac{a}{n_{j+2}} \leq \dots \leq \frac{a}{n_R}$, we have:

$$s_k = \begin{cases} \sqrt[2M]{N\lambda_{H_1} x_k^{*M}/c} & \forall\, k \leq i-1 \\ \sqrt[2M]{N\lambda_{H_1} x_k^{*M}/c} \text{ or } 0 & \forall\, i \leq k \leq j \\ 0 & \forall\, k \geq j+1 \end{cases},$$

furthermore, let r is the largest index that $s_r > 0$, we must have $r \leq R$ and $s_{r+1} = s_{r+2} = \dots = s_K = 0$. $(\mathcal{NC}1)$ and $(\mathcal{NC}3)$ are the same as above but for $(\mathcal{NC}2)$, we have:

$$\mathbf{W}_M^* \mathbf{W}_M^{*\top} = \frac{\lambda_{W_1}}{\lambda_{W_M}} \begin{bmatrix} s_1^2 & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & s_{i-1}^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & s_i^2 \mathcal{P}_{r-i+1}(\mathbf{I}_{j-i+1}) & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-j)\times(K-j)} \end{bmatrix}, \tag{103}$$

$$\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* = \begin{bmatrix} \frac{c s_1^{2M}}{(c s_1^{2M} + N\lambda_{H_1})^2} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & \frac{c s_{i-1}^{2M}}{(c s_{i-1}^{2M} + N\lambda_{H_1})^2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \frac{c s_i^{2M}}{(c s_i^{2M} + N\lambda_{H_1})^2} \mathcal{P}_{r-i+1}(\mathbf{I}_{j-i+1}) & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-j)\times(K-j)} \end{bmatrix}, \tag{104}$$

$$\mathbf{W}_M^* \mathbf{W}_{M-1}^* \dots \mathbf{W}_2^* \mathbf{W}_1^* \overline{\mathbf{H}}^* = \begin{bmatrix} \frac{c s_1^{2M}}{c s_1^{2M} + N\lambda_{H_1}} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & \frac{c s_{i-1}^{2M}}{c s_{i-1}^{2M} + N\lambda_{H_1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \frac{c s_i^{2M}}{c s_i^{2M} + N\lambda_{H_1}} \mathcal{P}_{r-i+1}(\mathbf{I}_{j-i+1}) & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-j)\times(K-j)} \end{bmatrix}, \tag{105}$$

and, for any $h > j$, $(\mathbf{W}_M^* \mathbf{W}_{M-1}^* \dots \mathbf{W}_2^* \mathbf{W}_1^*)_h = \mathbf{h}_h^* = \mathbf{0}$.

Proof of Theorem F.1 and F.2. First, by using lemma D.2, we have for any critical point $(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1)$ of f, we have the following:

$$\begin{aligned}
\lambda_{W_M} \mathbf{W}_M^\top \mathbf{W}_M &= \lambda_{W_{M-1}} \mathbf{W}_{M-1} \mathbf{W}_{M-1}^\top, \\
\lambda_{W_{M-1}} \mathbf{W}_{M-1}^\top \mathbf{W}_{M-1} &= \lambda_{W_{M-2}} \mathbf{W}_{M-2} \mathbf{W}_{M-2}^\top, \\
&\ldots \\
\lambda_{W_2} \mathbf{W}_2^\top \mathbf{W}_2 &= \lambda_{W_1} \mathbf{W}_1 \mathbf{W}_1^\top, \\
\lambda_{W_1} \mathbf{W}_1^\top \mathbf{W}_1 &= \lambda_{H_1} \mathbf{H}_1 \mathbf{H}_1^\top.
\end{aligned}$$

Let $\mathbf{W}_1 = \mathbf{U}_{W_1} \mathbf{S}_{W_1} \mathbf{V}_{W_1}^\top$ be the SVD decomposition of $\mathbf{W}_1$ with $\mathbf{U}_{W_1} \in \mathbb{R}^{d_2 \times d_2}, \mathbf{V}_{W_1} \in \mathbb{R}^{d_1 \times d_1}$ are orthonormal matrices and $\mathbf{S}_{W_1} \in \mathbb{R}^{d_2 \times d_1}$ is a diagonal matrix with **decreasing** non-negative singular values. We denote the r singular values of $\mathbf{W}_1$ as $\{s_k\}_{k=1}^r$ ($r \leq R := \min(K, d_M, \ldots, d_1)$). From Lemma D.4, we have the SVD of other weight matrices as:

$$\begin{aligned}
\mathbf{W}_M &= \mathbf{U}_{W_M} \mathbf{S}_{W_M} \mathbf{U}_{W_{M-1}}^\top, \\
\mathbf{W}_{M-1} &= \mathbf{U}_{W_{M-1}} \mathbf{S}_{W_{M-1}} \mathbf{U}_{W_{M-2}}^\top, \\
\mathbf{W}_{M-2} &= \mathbf{U}_{W_{M-2}} \mathbf{S}_{W_{M-2}} \mathbf{U}_{W_{M-3}}^\top, \\
\mathbf{W}_{M-3} &= \mathbf{U}_{W_{M-3}} \mathbf{S}_{W_{M-3}} \mathbf{U}_{W_{M-4}}^\top, \\
&\ldots, \\
\mathbf{W}_2 &= \mathbf{U}_{W_2} \mathbf{S}_{W_2} \mathbf{U}_{W_1}^\top, \\
\mathbf{W}_1 &= \mathbf{U}_{W_1} \mathbf{S}_{W_1} \mathbf{V}_{W_1}^\top,
\end{aligned}$$

with:

$$\mathbf{S}_{W_j} = \sqrt{\frac{\lambda_{W_1}}{\lambda_{W_j}}} \begin{bmatrix} \operatorname{diag}(s_1, \ldots, s_r) & \mathbf{0}_{r \times (d_j - r)} \\ \mathbf{0}_{(d_{j+1} - r) \times r} & \mathbf{0}_{(d_{j+1} - r) \times (d_j - r)} \end{bmatrix} \in \mathbb{R}^{d_{j+1} \times d_j} \quad \forall j \in [M],$$

and $\mathbf{U}_{W_M}, \mathbf{U}_{W_{M-1}}, \mathbf{U}_{W_{M-2}}, \mathbf{U}_{W_{M-3}}, \ldots, \mathbf{U}_{W_1}, \mathbf{V}_{W_1}$ are all orthonormal matrices.

From Lemma D.5, denote $c := \frac{\lambda_{W_1}^{M-1}}{\lambda_{W_M} \lambda_{W_{M-1}} \ldots \lambda_{W_2}}$, we have:

$$\begin{aligned}
\mathbf{H}_1 &= \mathbf{V}_{W_1} \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{\sqrt{c} s_1^M}{c s_1^{2M} + N \lambda_{H_1}}, \ldots, \frac{\sqrt{c} s_r^M}{c s_r^{2M} + N \lambda_{H_1}} \right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\mathbf{C} \in \mathbb{R}^{d_1 \times K}} \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \mathbf{V}_{W_1} \mathbf{C} \mathbf{U}_{W_M}^\top \mathbf{Y}.
\end{aligned} \tag{106}$$

$$\begin{aligned}
\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H} - \mathbf{Y} &= \mathbf{U}_{W_M} \underbrace{\begin{bmatrix} \operatorname{diag}\left(\frac{-N \lambda_{H_1}}{c s_1^{2M} + N \lambda_{H_1}}, \ldots, \frac{-N \lambda_{H_1}}{c s_r^{2M} + N \lambda_{H_1}} \right) & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_{K-r} \end{bmatrix}}_{\mathbf{D} \in \mathbb{R}^{K \times K}} \mathbf{U}_{W_M}^\top \mathbf{Y} \\
&= \mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top \mathbf{Y}.
\end{aligned} \tag{107}$$

Next, we will calculate the Frobenius norm of $\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1 - \mathbf{Y}$:

$$\begin{aligned}
\|\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1 - \mathbf{Y}\|_F^2 = \|\mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top \mathbf{Y}\|_F^2 &= \operatorname{trace}(\mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top \mathbf{Y} (\mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top \mathbf{Y})^\top) \\
&= \operatorname{trace}(\mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{U}_{W_M} \mathbf{D} \mathbf{U}_{W_M}^\top) \\
&= \operatorname{trace}(\mathbf{D}^2 \mathbf{U}_{W_M}^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{U}_{W_M}).
\end{aligned}$$

We denote $\mathbf{u}^k$ and $\mathbf{u}_k$ are the k-th row and column of $\mathbf{U}_{W_M}$, respectively. Let $\mathbf{n} = (n_1, \ldots, n_K)$, we have the following:

$$\mathbf{U}_{W_M} = \begin{bmatrix} -\mathbf{u}^1- \\ \ldots \\ -\mathbf{u}^K- \end{bmatrix} = \begin{bmatrix} | & | & | \\ \mathbf{u}_1 & \ldots & \mathbf{u}_K \\ | & | & | \end{bmatrix},$$

$$\mathbf{Y}\mathbf{Y}^\top = \operatorname{diag}(n_1, n_2, \ldots, n_K) \in \mathbb{R}^{K \times K}$$

$$\Rightarrow \mathbf{U}_{W_M}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_{W_M} = \begin{bmatrix} | & | & | \\ (\mathbf{u}^1)^\top & \ldots & (\mathbf{u}^K)^\top \\ | & | & | \end{bmatrix} \operatorname{diag}(n_1, n_2, \ldots, n_K) \begin{bmatrix} -\mathbf{u}^1- \\ \ldots \\ -\mathbf{u}^K- \end{bmatrix} \tag{108}$$

$$= \begin{bmatrix} | & | & | \\ (\mathbf{u}^1)^\top & \ldots & (\mathbf{u}^K)^\top \\ | & | & | \end{bmatrix} \begin{bmatrix} -n_1\mathbf{u}^1- \\ \ldots \\ -n_k\mathbf{u}^K- \end{bmatrix}$$

$$\Rightarrow (\mathbf{U}_{W_M}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_{W_M})_{kk} = n_1 u_{1k}^2 + n_2 u_{2k}^2 + \ldots + n_k u_{Kk}^2 = (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n}$$

$$\Rightarrow \|\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1 - \mathbf{Y}\|_F^2 = \operatorname{trace}(\mathbf{D}^2 \mathbf{U}_W^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_W)$$
$$= \sum_{k=1}^{r} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} \frac{(-N\lambda_{H_1})^2}{(cs_k^{2M} + N\lambda_{H_1})^2} + \sum_{h=r+1}^{K} (\mathbf{u}_h \odot \mathbf{u}_h)^\top \mathbf{n}, \tag{109}$$

where the last equality is from the fact that $\mathbf{D}^2$ is a diagonal matrix, so the diagonal of $\mathbf{D}^2 \mathbf{U}_{W_M}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_{W_M}$ is the element-wise product between the diagonal of $\mathbf{D}^2$ and $\mathbf{U}_{W_M}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_{W_M}$.

Similarly, we calculate the Frobenius norm of $\mathbf{H}_1$, from equation (106), we have:

$$\|\mathbf{H}_1\|_F^2 = \operatorname{trace}(\mathbf{V}_{W_1} \mathbf{C} \mathbf{U}_{W_M}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_{W_M} \mathbf{C}^\top \mathbf{V}_{W_1}^\top) = \operatorname{trace}(\mathbf{C}^\top \mathbf{C} \mathbf{U}_{W_M}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_{W_M})$$
$$= \sum_{k=1}^{r} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} \frac{cs_k^{2M}}{(cs_k^{2M} + N\lambda_{H_1})^2}. \tag{110}$$

Now, we plug the equations (109), (110) and the SVD of weight matrices into the function f and note that orthonormal matrix does not change Frobenius norm, we got:

$$f = \frac{1}{2N} \|\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1 \mathbf{H}_1 - \mathbf{Y}\|_F^2 + \frac{\lambda_{W_M}}{2} \|\mathbf{W}_M\|_F^2 + \ldots + \frac{\lambda_{W_1}}{2} \|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2} \|\mathbf{H}_1\|_F^2$$
$$= \frac{1}{2N} \sum_{k=1}^{r} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} \frac{(-N\lambda_{H_1})^2}{(cs_k^{2M} + N\lambda_{H_1})^2} + \frac{1}{2N} \sum_{h=r+1}^{K} (\mathbf{u}_h \odot \mathbf{u}_h)^\top \mathbf{n} + \frac{\lambda_{W_M}}{2} \sum_{k=1}^{r} \frac{\lambda_{W_1}}{\lambda_{W_M}} s_k^2$$
$$+ \frac{\lambda_{W_{M-1}}}{2} \sum_{k=1}^{r} \frac{\lambda_{W_1}}{\lambda_{W_{M-1}}} s_k^2 + \ldots + \frac{\lambda_{W_1}}{2} \sum_{k=1}^{r} s_k^2 + \frac{\lambda_{H_1}}{2} \sum_{k=1}^{r} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} \frac{cs_k^{2M}}{(cs_k^{2M} + N\lambda_{H_1})^2}$$
$$= \frac{\lambda_{H_1}}{2} \sum_{k=1}^{r} \frac{(\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n}}{cs_k^{2M} + N\lambda_{H_1}} + \frac{1}{2N} \sum_{h=r+1}^{K} (\mathbf{u}_h \odot \mathbf{u}_h)^\top \mathbf{n} + \frac{M\lambda_{W_1}}{2} \sum_{k=1}^{r} s_k^2$$
$$= \frac{1}{2N} \sum_{k=1}^{r} \left(\frac{(\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n}}{\frac{cs_k^{2M}}{N\lambda_{H_1}} + 1} + MN\lambda_{W_1} \sqrt[M]{\frac{N\lambda_{H_1}}{c}} \left(\sqrt[M]{\frac{cs_k^{2M}}{N\lambda_{H_1}}} \right) \right) + \frac{1}{2N} \sum_{h=r+1}^{K} (\mathbf{u}_h \odot \mathbf{u}_h)^\top \mathbf{n}$$
$$= \frac{1}{2N} \sum_{k=1}^{r} \left(\frac{(\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n}}{x_k^M + 1} + bx_k \right) + \frac{1}{2N} \sum_{h=r+1}^{K} (\mathbf{u}_h \odot \mathbf{u}_h)^\top \mathbf{n}$$
$$= \frac{1}{2N} \sum_{k=1}^{r} \left(\frac{a_k}{x_k^M + 1} + bx_k \right) + \frac{1}{2N} \sum_{h=r+1}^{K} a_h, \tag{111}$$

with $x_k := \sqrt[M]{\frac{cs_k^{2M}}{N\lambda_{H_1}}}$, $a_k := (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n}$ and $b := MN\lambda_{W_1} \sqrt[M]{\frac{N\lambda_{H_1}}{c}} = MN\lambda_{W_1} \sqrt[M]{\frac{N\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_2}\lambda_{H_1}}{\lambda_{W_1}^{M-1}}} = MN \sqrt[M]{N\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_1}\lambda_{H_1}}$.

From the fact that $\mathbf{U}_W$ is an orthonormal matrix, we have:

$$\sum_{k=1}^K a_k = \sum_{k=1}^K (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} = \left(\sum_{k=1}^K \mathbf{u}_k \odot \mathbf{u}_k\right)^\top \mathbf{n} = \mathbf{1}^\top \mathbf{n} = \sum_{k=1}^K n_k = N, \tag{112}$$

and, for any $j \in [K]$, denote $p_{i,j} := u_{i1}^2 + u_{i2}^2 + \ldots + u_{ij}^2 \; \forall \, i \in [K]$, we have:

$$\begin{aligned}
\sum_{k=1}^j a_k &= \sum_{k=1}^j (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} = n_1(u_{11}^2 + u_{12}^2 + \ldots + u_{1j}^2) + n_2(u_{21}^2 + u_{22}^2 + \ldots + u_{2j}^2) + \ldots + n_K(u_{K1}^2 + u_{K2}^2 + \ldots + u_{Kj}^2) \\
&= \sum_{k=1}^K p_{k,j} n_k \leq p_{1,j}n_1 + p_{2,j}n_2 + \ldots + p_{j-1,j}n_{j-1} + (p_{j,j} + p_{j+1,j} + p_{j+2,j} + \ldots + p_{K,j})n_j \\
&= p_{1,j}n_1 + p_{2,j}n_2 + \ldots + p_{j-1,j}n_{j-1} + (j - p_{1,j} + \ldots + p_{j-1,j})n_j \\
&= \sum_{k=1}^j n_k + \sum_{h=1}^{j-1} (n_h - n_j)(p_{h,j} - 1) \leq \sum_{k=1}^j n_k
\end{aligned}$$

$$\Rightarrow \sum_{k=j+1}^K a_k \geq N - \sum_{k=1}^j n_k = \sum_{k=j+1}^K n_k \quad \forall \, j \in [K], \tag{113}$$

where we used the fact that $\sum_{k=1}^K p_{k,j} = j$ since it is the sum of squares of all entries of the first j columns of an orthonormal matrix, and $p_{i,j} \leq 1 \, \forall \, i$ because it is the sum of squares of some entries on the i-th row of $\mathbf{U}_W$.

By applying Lemma E.3 to the RHS of equation (111) with $z_k = \frac{1}{x_k^M + 1} \; \forall \, k \leq r$ and $z_k = 1$ otherwise, we obtain:

$$f(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1) \geq \frac{1}{2N} \sum_{k=1}^r \left(\frac{n_k}{x_k^M + 1} + bx_k\right) + \frac{1}{2N} \sum_{h=r+1}^K n_h \tag{114}$$

$$= \frac{1}{2N} \sum_{k=1}^r n_k \left(\frac{1}{x_k^M + 1} + \frac{b}{n_k} x_k\right) + \frac{1}{2N} \sum_{h=r+1}^K n_h. \tag{115}$$

The minimizer of the function $g(x) = \frac{1}{x^M + 1} + ax$ has been studied in Section D.2.1. Apply this result for the lower bound (115), we finish bounding $f(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1)$.

Now, we study the equality conditions. In the lower bound (115), by letting x_k^* be the minimizer of $\frac{1}{x_k^M + 1} + \frac{b}{n_k} x_k$ for all $k \leq r$ and $x_k^* = 0$ for all $k > r$, there are only four possibilities as following:

- **Case A:** If $x_1^* > 0$ and $n_1 > n_2$: If $x_2^* = 0$, it is clear that $x_1^* > x_2^*$. Otherwise, we have x_1^* and x_2^* must satisfy (see Section D.2.1 for details):

$$\frac{M x_1^{*M-1}}{(x_1^{*M} + 1)^2} = \frac{b}{n_1},$$
$$\frac{M x_2^{*M-1}}{(x_2^{*M} + 1)^2} = \frac{b}{n_2}.$$

Because $\frac{b}{n_1} < \frac{b}{n_2}$ and the function $p(x) = \frac{Mx^{M-1}}{(x^M+1)^2}$ is a decreasing function when $x > \sqrt[M]{\frac{M-1}{M+1}}$, we got $x_1^* > x_2^*$. Hence, from the equality condition of Lemma E.3, we have $a_1 = n_1$. From the orthonormal property of $\mathbf{u}_k$, we have:

$$a_1 = (\mathbf{u}_1 \odot \mathbf{u}_1)^\top \mathbf{n} = n_1 u_{11}^2 + n_2 u_{21}^2 + \ldots + n_k u_{K1}^2 \leq n_1(u_{11}^2 + u_{21}^2 + \ldots + u_{K1}^2) = n_1.$$

The equality holds when and only when $u_{11}^2 = 1$ and $u_{21} = \ldots = u_{K1} = 0$.

- **Case B:** If $x_1^* > 0$ and there exists $1 < j \leq r$ such that $n_1 = n_2 = \ldots = n_j > n_{j+1}$, we have:

$$\frac{1}{x^M+1} + \frac{b}{n_1}x = \frac{1}{x^M+1} + \frac{b}{n_2}x = \ldots = \frac{1}{x^M+1} + \frac{b}{n_j}x,$$

and thus, $x_1^* = x_2^* = \ldots = x_j^* > x_{j+1}^*$. Hence, from the equality condition of Lemma E.3, we have $a_1 + a_2 + \ldots + a_j = n_1 + \ldots + n_j$. We have:

$$\sum_{k=1}^{j} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} = n_1(u_{11}^2 + u_{12}^2 + \ldots + u_{1j}^2) + n_2(u_{21}^2 + u_{22}^2 + \ldots + u_{2j}^2)$$
$$+ \ldots + n_K(u_{K1}^2 + u_{K2}^2 + \ldots + u_{Kj}^2) \leq \sum_{k=1}^{j} n_j,$$

where the inequality is from the fact that for any $k \in [K]$, $(u_{k1}^2 + u_{k2}^2 + \ldots + u_{kj}^2) \leq 1$ and $\sum_{k=1}^{K}(u_{k1}^2 + u_{k2}^2 + \ldots + u_{kj}^2) = j$. The equality holds iff $u_{k1}^2 + u_{k2}^2 + \ldots + u_{kj}^2 = 1 \; \forall \, k = 1, 2, \ldots, j$ and $u_{k1} = u_{k2} = \ldots = u_{kj} = 0 \; \forall \, k = j+1, \ldots, K$, i.e. the upper left sub-matrix size $j \times j$ of $\mathbf{U}_{W_M}$ is an orthonormal matrix and other entries of $\mathbf{U}_{W_M}$ lie on the same rows or columns with this sub-matrix must all equal 0's.

- **Case C:** If $x_1^* > 0$, $r < K$ and there exists $r < j \leq K$ such that $n_1 = n_2 = \ldots = n_r = \ldots = n_j > n_{j+1}$, we have $x_1^* = x_2^* = \ldots = x_r^* > 0$ and $x_{r+1}^* = \ldots = x_K^* = 0$. Hence, from the equality condition of Lemma E.3, we have $a_1 + a_2 + \ldots + a_r = n_1 + \ldots + n_r$. We have:

$$\sum_{k=1}^{r} (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n} = n_1(u_{11}^2 + u_{12}^2 + \ldots + u_{1r}^2) + n_2(u_{21}^2 + u_{22}^2 + \ldots + u_{2r}^2)$$
$$+ \ldots + n_K(u_{K1}^2 + u_{K2}^2 + \ldots + u_{Kr}^2) \leq \sum_{k=1}^{r} n_k,$$

where the inequality is from the fact that for any $k \in [K]$, $(u_{k1}^2 + u_{k2}^2 + \ldots + u_{kr}^2) \leq 1$ and $\sum_{k=1}^{K}(u_{k1}^2 + u_{k2}^2 + \ldots + u_{kr}^2) = r$. The equality holds iff $u_{k1} = u_{k2} = \ldots = u_{kr} = 0 \; \forall \, k = j+1, \ldots, K$, i.e. the upper left sub-matrix size $j \times r$ of $\mathbf{U}_{W_M}$ includes r orthonormal vectors in $\mathbb{R}^j$ and the bottom left sub-matrix size $(K-j) \times r$ are all zeros. The other $K - r$ columns of $\mathbf{U}_{W_M}$ does not matter because $\mathbf{W}_M^*$ can be written as:

$$\mathbf{W}_M^* = \sum_{k=1}^{r} s_k^* \mathbf{u}_k \mathbf{v}_k^\top,$$

with $\mathbf{v}_k$ is the right singular vector that satisfies $\mathbf{W}_M^{*\top} \mathbf{u}_k = s_k^* \mathbf{v}_k$. Note that since $s_1^* = s_2^* = \ldots = s_r^* := s^*$, thus we have compact SVD form as follows:

$$\mathbf{W}_M^* = s^* \mathbf{U}_{W_M}^{'} \mathbf{V}_{W_M}^{'\top}, \tag{116}$$

where $\mathbf{U}_{W_M}^{'} \in \mathbb{R}^{K \times r}$ and $\mathbf{V}_{W_M}^{'} \in \mathbb{R}^{d \times r}$. Especially, the last $K - j$ rows of $\mathbf{W}_M^*$ will be zeros since the last $K - j$ rows of $\mathbf{U}_{W_M}^{'}$ are zeros. Furthermore, $\mathbf{U}_{W_M}^{'} \mathbf{U}_{W_M}^{'\top}$ after removing the last $K - j$ zero rows and the last $K - j$ zero columns is the best rank-r approximation of $\mathbf{I}_j$.

We note that if **Case C** happens, then the number of positive singular values are limited by the matrix rank r (e.g., by $r \leq R = \min(d_M, \ldots, d_1, K) < K$), and $n_r = n_{r+1}$, thus $x_r^* > 0$ and $x_{r+1}^* = 0$ (x_{r+1}^* should equal $x_r^* > 0$ if it is not forced to be zero).

- **Case D:** If $x_1^* = 0$, we must have $x_2^* = \ldots = x_K^* = 0$, $\sum_{k=1}^K (\mathbf{u}_k \odot \mathbf{u}_k)^\top \mathbf{n}$ always equal N and thus, $\mathbf{U}_{W_M}$ can be an arbitrary size $K \times K$ orthonormal matrix.

We perform similar arguments as above for all subsequent x_k^*'s, after we finish reasoning for prior ones. Before going to the conclusion, we first study the matrix $\mathbf{U}_{W_M}$. If **Case C** does not happen for any x_k^*'s, we have:

$$\mathbf{U}_{W_M} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_l \end{bmatrix}, \tag{117}$$

where each $\mathbf{A}_i$ is an orthonormal block which corresponds with one or a group of classes that have the same number of training samples and their $x^* > 0$ (**Case A** and **Case B**) or corresponds with all classes with $x^* = 0$ (**Case D**). If **Case C** happens, we have:

$$\mathbf{U}_{W_M} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_l \end{bmatrix}, \tag{118}$$

where each $\mathbf{A}_i, i \in [l-1]$ is an orthonormal block which corresponds with one or a group of classes that have the same number of training samples and their $x^* > 0$ (**Case A** and **Case B**). $\mathbf{A}_l$ is the orthonormal block has the same property as $\mathbf{U}_{W_M}$ in **Case C**.

We consider the case $R = K$ from now on. By using arguments about the minimizer of $g(x)$ applied to the lower bound (115), we consider four cases as following:

- **Case 1a:** $\frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_K} < \frac{(M-1)^{\frac{M-1}{M}}}{M}$.

  Then, the lower bound (115) is minimized at $(x_1^*, x_2^*, \ldots, x_K^*)$ where x_i^* is the largest positive solution of the equation $\frac{b}{n_i} - \frac{M x^{M-1}}{(x^M+1)^2} = 0$ for $i = 1, 2, \ldots, K$. We conclude:

$$(s_1^*, s_2^*, \ldots, s_K^*) = \left(\sqrt[2M]{\frac{N\lambda_{H_1} x_1^{*M}}{c}}, \sqrt[2M]{\frac{N\lambda_{H_1} x_2^{*M}}{c}}, \ldots \sqrt[2M]{\frac{N\lambda_{H_1} x_K^{*M}}{c}} \right). \tag{119}$$

  First, we have the property that the features in each class $\mathbf{h}_{k,i}^*$ collapsed to their class-mean $\mathbf{h}_k^*$ ($\mathcal{NC}1$). Let $\overline{\mathbf{H}}^* = \mathbf{V}_{W_1} \mathbf{C} \mathbf{U}_{W_M}^\top$, we know that $\mathbf{H}_1^* = \overline{\mathbf{H}}^* \mathbf{Y}$ from equation (106). Then, columns from the $(n_{k-1}+1)$-th until (n_k)-th of $\mathbf{H}_1^*$ will all equals the k-th column of $\overline{\mathbf{H}}^*$, thus the features in class k collapse to their class-mean $\mathbf{h}_k^*$ (which is the k-th column of $\overline{\mathbf{H}}^*$), i.e., $\mathbf{h}_{k,1}^* = \mathbf{h}_{k,2}^* = \ldots = \mathbf{h}_{k,n_k}^* \; \forall \, k \in [K]$.

  Since $r = R = K$, **Case C** never happens, and we have $\mathbf{U}_{W_M}$ as in equation (117). Hence, together with equations (106) and (107), we can conclude the geometry of the following:

$$\mathbf{W}_M^* \mathbf{W}_M^{*\top} = \mathbf{U}_{W_M} \mathbf{S}_{W_M} \mathbf{S}_{W_M}^\top \mathbf{U}_{W_M}^\top = \operatorname{diag}\left(\frac{\lambda_{W_1}}{\lambda_{W_M}} s_1^2, \ldots, \frac{\lambda_{W_1}}{\lambda_{W_M}} s_K^2 \right), \tag{120}$$

$$\mathbf{H}_1^{*\top} \mathbf{H}_1^* = \mathbf{Y}^\top \mathbf{U}_{W_M} \mathbf{C}^T \mathbf{C} \mathbf{U}_{W_M}^\top \mathbf{Y} = \begin{bmatrix} \frac{c s_1^{2M}}{(c s_1^{2M} + N\lambda_{H_1})^2} \mathbf{1}_{n_1} \mathbf{1}_{n_1}^\top & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \frac{c s_K^{2M}}{(c s_K^{2M} + N\lambda_{H_1})^2} \mathbf{1}_{n_K} \mathbf{1}_{n_K}^\top \end{bmatrix}, \tag{121}$$

$$\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*\mathbf{H}_1^* = \mathbf{U}_{W_M}\mathbf{S}_{W_M}\mathbf{S}_{W_{M-1}}\ldots\mathbf{S}_{W_1}\mathbf{C}\mathbf{U}_{W_M}^\top\mathbf{Y}$$
$$= \begin{bmatrix} \frac{cs_1^{2M}}{cs_1^{2M}+N\lambda_{H_1}}\mathbf{1}_{n_1}^\top & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \frac{cs_K^{2M}}{cs_K^{2M}+N\lambda_{H_1}}\mathbf{1}_{n_K}^\top \end{bmatrix}. \tag{122}$$

We additionally have the structure of the class-means matrix:

$$\overline{\mathbf{H}}^{*\top}\overline{\mathbf{H}}^* = \mathbf{U}_{W_M}^\top\mathbf{C}^\top\mathbf{C}\mathbf{U}_{W_M} = \begin{bmatrix} \frac{cs_1^{2M}}{(cs_1^{2M}+N\lambda_{H_1})^2} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{cs_K^{2M}}{(cs_K^{2M}+N\lambda_{H_1})^2} \end{bmatrix}, \tag{123}$$

$$\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*\overline{\mathbf{H}}^* = \mathbf{U}_{W_M}\mathbf{S}_{W_M}\mathbf{C}\mathbf{U}_{\mathbf{W}}{}^\top = \begin{bmatrix} \frac{cs_1^{2M}}{cs_1^{2M}+N\lambda_{H_1}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{cs_K^{2M}}{cs_K^{2M}+N\lambda_{H_1}} \end{bmatrix}. \tag{124}$$

And the alignment between the weights and features are as following. For any $k \in [K]$, denote $(\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*)_k$ the k-th row of $\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*$:

$$\begin{gathered}\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^* = \mathbf{U}_{W_M}\mathbf{S}_{W_M}\mathbf{S}_{W_{M-1}}\ldots\mathbf{S}_{W_1}\mathbf{V}_{W_1}^\top, \\ \overline{\mathbf{H}}^* = \mathbf{V}_{W_1}\mathbf{C}\mathbf{U}_{W_M}^\top \\ \Rightarrow (\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*)_k = (cs_k^{2M}+N\lambda_{H_1})\mathbf{h}_k^*.\end{gathered} \tag{125}$$

- **Case 2a:** There exists $j \in [K-1]$ s.t. $\frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_j} < \frac{(M-1)^{\frac{M-1}{M}}}{M} < \frac{b}{n_{j+1}} \leq \ldots \leq \frac{b}{n_K}$.

Then, the lower bound (115) is minimized at $(x_1^*, x_2^*, \ldots, x_K^*)$ where x_i^* is the largest positive solution of equation $\frac{b}{n_i} - \frac{Mx^{M-1}}{(x^M+1)^2} = 0$ for $i = 1, 2, \ldots, j$ and $x_i^* = 0$ for $i = j+1, \ldots, K$. We conclude:

$$(s_1^*, s_2^*, \ldots, s_j^*, s_{j+1}^*, \ldots s_K^*) = \left(\sqrt[2M]{\frac{N\lambda_{H_1}x_1^{*M}}{c}}, \sqrt[2M]{\frac{N\lambda_{H_1}x_2^{*M}}{c}}, \ldots, \sqrt[2M]{\frac{N\lambda_{H_1}x_j^{*M}}{c}}, 0, \ldots, 0\right). \tag{126}$$

First, we have the property that the features in each class $\mathbf{h}_{k,i}^*$ collapsed to their class-mean $\mathbf{h}_k^*$ ($\mathcal{NC}1$). Let $\overline{\mathbf{H}}^* = \mathbf{V}_W\mathbf{C}\mathbf{U}_W^\top$, we know that $\mathbf{H}_1^* = \overline{\mathbf{H}}^*\mathbf{Y}$. Then, columns from the $(n_{k-1}+1)$-th until (n_k)-th of $\mathbf{H}_1^*$ will all equals the k-th column of $\overline{\mathbf{H}}^*$, thus the features in class k are collapsed to their class-mean $\mathbf{h}_k^*$ (which is the k-th column of $\overline{\mathbf{H}}$), i.e $\mathbf{h}_{k,1}^* = \mathbf{h}_{k,2}^* = \ldots = \mathbf{h}_{k,n_k}^* \forall k \in [K]$.

For any $k \in [K]$, denote $(\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*)_k$ the k-th row of $\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*$:

$$\begin{gathered}\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^* = \mathbf{U}_{W_M}\mathbf{S}_{W_M}\mathbf{S}_{W_{M-1}}\ldots\mathbf{S}_{W_1}\mathbf{V}_{W_1}^\top, \\ \overline{\mathbf{H}}^* = \mathbf{V}_{W_1}\mathbf{C}\mathbf{U}_{W_M}^\top \\ \Rightarrow (\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*)_k = (cs_k^{2M}+N\lambda_{H_1})\mathbf{h}_k^*.\end{gathered} \tag{127}$$

And, for $k > j$, we have $(\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*)_k = \mathbf{h}_k^* = \mathbf{0}$.

Recall the form of $\mathbf{U}_{W_M}$ as in equation (117) (**Case C** cannot happen since $r = j$ and $n_j > n_{j+1}$). We can conclude the geometry of following objects, with the usage of equations (106) and (107):

$$\mathbf{W}_M^* \mathbf{W}_M^{*\top} = \mathbf{U}_{W_M} \mathbf{S}_{W_M} \mathbf{S}_{W_M}^\top \mathbf{U}_W^\top = \operatorname{diag}\left(\frac{\lambda_{W_1}}{\lambda_{W_M}} s_1^2, \frac{\lambda_{W_1}}{\lambda_{W_M}} s_2^2, \ldots, \frac{\lambda_{W_1}}{\lambda_{W_M}} s_j^2, 0, \ldots, 0 \right), \tag{128}$$

$$\mathbf{H}_1^{*\top} \mathbf{H}_1^* = \begin{bmatrix} \frac{c s_1^{2M}}{(c s_1^{2M} + N\lambda_{H_1})^2} \mathbf{1}_{n_1} \mathbf{1}_{n_1}^\top & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \frac{c s_2^{2M}}{(c s_2^{2M} + N\lambda_{H_1})^2} \mathbf{1}_{n_2} \mathbf{1}_{n_2}^\top & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0}_{n_K \times n_K} \end{bmatrix}, \tag{129}$$

$$\begin{aligned} \mathbf{W}_M^* \mathbf{W}_{M-1}^* \ldots \mathbf{W}_2^* \mathbf{W}_1^* \mathbf{H}_1^* &= \mathbf{U}_W \operatorname{diag}\left(\frac{c s_1^{2M}}{c s_1^{2M} + N\lambda_{H_1}}, \ldots, \frac{c s_j^{2M}}{c s_j^{2M} + N\lambda_{H_1}}, 0, \ldots, 0 \right) \mathbf{U}_W^\top \mathbf{Y} \\ &= \begin{bmatrix} \frac{c s_1^{2M}}{c s_1^{2M} + N\lambda_{H_1}} \mathbf{1}_{n_1}^\top & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \frac{c s_2^{2M}}{c s_2^{2M} + N\lambda_{H_1}} \mathbf{1}_{n_2}^\top & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0}_{n_K}^\top \end{bmatrix}, \end{aligned}$$

where $\mathbf{1}_{n_k} \mathbf{1}_{n_k}^\top$ is a $n_k \times n_k$ matrix will all entries are 1's.

- **Case 3a:** $\frac{(M-1)^{\frac{M-1}{M}}}{M} < \frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_K}$.

  In this case, the lower bound (115) is minimized at:

$$(s_1^*, s_2^*, \ldots, s_K^*) = (0, 0, \ldots, 0). \tag{130}$$

  Hence, the global minimizer of f is $(\mathbf{W}_M^*, \mathbf{W}_{M-1}^*, \ldots, \mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^*) = (\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0})$.

- **Case 4a:** There exists $i, j \in [K]$ $(i \leq j)$ such that $\frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_{i-1}} < \frac{b}{n_i} = \frac{b}{n_{i+1}} = \ldots = \frac{b}{n_j} = \frac{(M-1)^{\frac{M-1}{M}}}{M} < \frac{b}{n_{j+1}} \leq \frac{b}{n_{j+2}} \leq \ldots \leq \frac{b}{n_K}$.
  Then, the lower bound (115) is minimized at $(x_1^*, x_2^*, \ldots, x_K^*)$ where $\forall\, t \leq i-1, x_t^*$ is the largest positive solution of equation $\frac{b}{n_t} - \frac{M x^{M-1}}{(x^M+1)^2} = 0$. If $i \leq t \leq j, x_t^*$ can either be 0 or the largest positive solution of equation $\frac{b}{n_t} - \frac{M x^{M-1}}{(x^M+1)^2} = 0$ as long as the sequence $\{x_t^*\}$ is a decreasing sequence. Otherwise, $\forall\, t > j, x_t^* = 0$.

  In this case, we have $\mathcal{NC}1$ and $\mathcal{NC}3$ properties similar as **Case 1a**.

  For $(\mathcal{NC}2)$, we can freely choose the number of positive singular values r to be any value between i and j. Thus, **Case C** does happen for this case. As a consequence, the diagonal block $\operatorname{diag}(s_i^2, \ldots, s_j^2)$ of $\mathbf{W}_M^* \mathbf{W}_M^{*\top}$ in **Case 1a**, will be replace by $s_r^2 \mathcal{P}_{r-i+1}(\mathbf{I}_{j-i+1})$. Similar changes are also applied for $\mathbf{H}_1^{*\top} \mathbf{H}_1^*$ and $\mathbf{W}_M^* \mathbf{W}_{M-1}^* \ldots \mathbf{W}_2^* \mathbf{W}_1^* \mathbf{H}_1^*$.

Now, we turn to consider the case $R < K$. Again, we consider the following cases:

- **Case 1b:** $\frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_R} < \frac{(M-1)^{\frac{M-1}{M}}}{M}$.

  Then, the lower bound (115) is minimized at $(x_1^*, x_2^*, \ldots, x_K^*)$ where x_i^* is the largest positive solution of the equation $\frac{b}{n_i} - \frac{M x^{M-1}}{(x^M+1)^2} = 0$ for $i = 1, 2, \ldots, R$ and $x_i^* = 0$ for $i = R+1, \ldots, K$. We conclude:

$$(s_1^*, s_2^*, \ldots, s_R^*, s_{R+1}^*, \ldots s_K^*) = \left(\sqrt[2M]{\frac{N\lambda_{H_1} x_1^{*M}}{c}}, \sqrt[2M]{\frac{N\lambda_{H_1} x_2^{*M}}{c}}, \ldots \sqrt[2M]{\frac{N\lambda_{H_1} x_R^{*M}}{c}}, 0, \ldots, 0 \right). \tag{131}$$

We have $(\mathcal{NC}1)$ and $(\mathcal{NC}3)$ properties are the same as **Case 1a**.

We have **Case C** happens iff $x_R^* > 0$ (already satisfied) and $n_R = n_{R+1}$. If $n_R > n_{R+1}$, we can conclude the geometry of the following:

$$\mathbf{W}_M^* \mathbf{W}_M^{*\top} = \mathbf{U}_{W_M} \mathbf{S}_{W_M} \mathbf{S}_{W_M}^\top \mathbf{U}_{W_M}^\top = \begin{bmatrix} \frac{\lambda_{W_1}}{\lambda_{W_M}} s_1^2 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & \frac{\lambda_{W_1}}{\lambda_{W_M}} s_R^2 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix}$$
$$= \operatorname{diag}\left(\frac{\lambda_{W_1}}{\lambda_{W_M}} s_1^2, \dots, \frac{\lambda_{W_1}}{\lambda_{W_M}} s_R^2, 0, \dots, 0 \right), \tag{132}$$

$$\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* = \mathbf{U}_{W_M}^\top \mathbf{C}^\top \mathbf{C} \mathbf{U}_{W_M} = \begin{bmatrix} \frac{c s_1^{2M}}{(c s_1^{2M} + N\lambda_{H_1})^2} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & \frac{c s_R^{2M}}{(c s_R^{2M} + N\lambda_{H_1})^2} & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix}, \tag{133}$$

$$\mathbf{W}_M^* \mathbf{W}_{M-1}^* \dots \mathbf{W}_2^* \mathbf{W}_1^* \overline{\mathbf{H}}^* = \mathbf{U}_{W_M} \mathbf{S}_{W_M} \mathbf{C} \mathbf{U}_{W_M}^\top = \begin{bmatrix} \frac{c s_1^{2M}}{c s_1^{2M} + N\lambda_{H_1}} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & \frac{c s_R^{2M}}{c s_R^{2M} + N\lambda_{H_1}} & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix}. \tag{134}$$

Furthermore, for $k > R$, we have $(\mathbf{W}_M^* \mathbf{W}_{M-1}^* \dots \mathbf{W}_2^* \mathbf{W}_1^*)_k = \mathbf{h}_k^* = \mathbf{0}$.

If $n_R = n_{R+1}$, there exists $k \leq R, l > R$ such that $n_{k-1} > n_k = n_{k+1} = \ldots = n_R = \ldots = n_l > n_{l+1}$, then :

$$\mathbf{W}_M^* \mathbf{W}_M^{*\top} = \frac{\lambda_{W_1}}{\lambda_{W_M}} \begin{bmatrix} s_1^2 & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & s_{k-1}^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & s_k^2 \mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{135}$$

$$\overline{\mathbf{H}}^{*\top} \overline{\mathbf{H}}^* = \begin{bmatrix} \frac{c s_1^{2M}}{(c s_1^{2M} + N\lambda_{H_1})^2} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & \frac{c s_{k-1}^{2M}}{(c s_{k-1}^{2M} + N\lambda_{H_1})^2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \frac{c s_k^{2M}}{(c s_k^{2M} + N\lambda_{H_1})^2} \mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{136}$$

$$\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*\overline{\mathbf{H}}^* = \begin{bmatrix} \frac{cs_1^{2M}}{cs_1^{2M}+N\lambda_{H_1}} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \ldots & \frac{cs_{k-1}^{2M}}{cs_{k-1}^{2M}+N\lambda_{H_1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \frac{cs_k^{2M}}{cs_k^{2M}+N\lambda_{H_1}}\mathcal{P}_{R-k+1}(\mathbf{I}_{l-k+1}) & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}_{(K-l)\times(K-l)} \end{bmatrix}, \tag{137}$$

and, for any $h > l > R$, $(\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*)_h = \mathbf{h}_h^* = \mathbf{0}$.

- **Case 2b:** There exists $j \in [R-1]$ s.t. $\frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_j} < \frac{(M-1)^{\frac{M-1}{M}}}{M} < \frac{b}{n_{j+1}} \leq \ldots \leq \frac{b}{n_R}$.

Then, the lower bound (115) is minimized at $(x_1^*, x_2^*, \ldots, x_K^*)$ where x_i^* is the largest positive solution of equation $\frac{b}{n_i} - \frac{Mx^{M-1}}{(x^M+1)^2} = 0$ for $i = 1, 2, \ldots, j$ and $x_i^* = 0$ for $i = j+1, \ldots, K$. We conclude:

$$(s_1^*, s_2^*, \ldots, s_j^*, s_{j+1}^*, \ldots s_K^*) = \left(\sqrt[2M]{\frac{N\lambda_{H_1}x_1^{*M}}{c}}, \sqrt[2M]{\frac{N\lambda_{H_1}x_2^{*M}}{c}}, \ldots, \sqrt[2M]{\frac{N\lambda_{H_1}x_j^{*M}}{c}}, 0, \ldots, 0 \right). \tag{138}$$

We have $(\mathcal{NC}1)$ and $(\mathcal{NC}3)$ properties are the same as **Case 2a**.

We can conclude the geometry of following objects, with the usage of equations (106) and (107):

$$\begin{aligned} \mathbf{W}_M^*\mathbf{W}_M^{*\top} &= \mathbf{U}_{W_M}\mathbf{S}_{W_M}\mathbf{S}_{W_M}^\top\mathbf{U}_W^\top \\ &= \operatorname{diag}\left(\frac{\lambda_{W_1}}{\lambda_{W_M}}s_1^2, \frac{\lambda_{W_1}}{\lambda_{W_M}}s_2^2, \ldots, \frac{\lambda_{W_1}}{\lambda_{W_M}}s_j^2, 0, \ldots, 0\right), \end{aligned} \tag{139}$$

$$\mathbf{H}_1^{*\top}\mathbf{H}_1^* = \begin{bmatrix} \frac{cs_1^{2M}}{(cs_1^{2M}+N\lambda_{H_1})^2}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \frac{cs_2^{2M}}{(cs_2^{2M}+N\lambda_{H_1})^2}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0}_{n_K\times n_K} \end{bmatrix}, \tag{140}$$

$$\begin{aligned} \mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*\mathbf{H}_1^* &= \mathbf{U}_W \operatorname{diag}\left(\frac{cs_1^{2M}}{cs_1^{2M}+N\lambda_{H_1}}, \ldots, \frac{cs_j^{2M}}{cs_j^{2M}+N\lambda_{H_1}}, 0, \ldots, 0\right)\mathbf{U}_W^\top\mathbf{Y} \\ &= \begin{bmatrix} \frac{cs_1^{2M}}{cs_1^{2M}+N\lambda_{H_1}}\mathbf{1}_{n_1}^\top & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \frac{cs_2^{2M}}{cs_2^{2M}+N\lambda_{H_1}}\mathbf{1}_{n_2}^\top & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0}_{n_K}^\top \end{bmatrix}, \end{aligned}$$

where $\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top$ is a $n_k \times n_k$ matrix will all entries are 1's. **Case C** cannot happen in this case because $r = j < R$ and $n_j > n_{j+1}$.

And, for $k > j$, we have $(\mathbf{W}_M^*\mathbf{W}_{M-1}^*\ldots\mathbf{W}_2^*\mathbf{W}_1^*)_k = \mathbf{h}_k^* = \mathbf{0}$.

- **Case 3b:** $\frac{(M-1)^{\frac{M-1}{M}}}{M} < \frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_R}$.

In this case, the lower bound (115) is minimized at:

$$(s_1^*, s_2^*, \ldots, s_K^*) = (0, 0, \ldots, 0). \tag{141}$$

Hence, the global minimizer of f is $(\mathbf{W}_M^*, \mathbf{W}_{M-1}^*, \ldots, \mathbf{W}_2^*, \mathbf{W}_1^*, \mathbf{H}_1^*) = (\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0})$.

- **Case 4b:** There exists $i, j \in [R]$ $(i \leq j \leq R)$ such that $\frac{b}{n_1} \leq \frac{b}{n_2} \leq \ldots \leq \frac{b}{n_{i-1}} < \frac{b}{n_i} = \frac{b}{n_{i+1}} = \ldots = \frac{b}{n_j} = \frac{(M-1)^{\frac{M-1}{M}}}{M} < \frac{b}{n_{j+1}} \leq \frac{b}{n_{j+2}} \leq \ldots \leq \frac{b}{n_R}$.

  Then, the lower bound (115) is minimized at $(x_1^*, x_2^*, \ldots, x_K^*)$ where $\forall\, t \leq i-1, x_t^*$ is the largest positive solution of equation $\frac{b}{n_t} - \frac{Mx^{M-1}}{(x^M+1)^2} = 0$. If $i \leq t \leq j, x_t^*$ can either be 0 or the largest positive solution of equation $\frac{b}{n_t} - \frac{Mx^{M-1}}{(x^M+1)^2} = 0$ as long as the sequence $\{x_t^*\}$ is a decreasing sequence and there is no more than R positive singular values. Otherwise, $\forall\, t > j, x_t^* = 0$.

  In this case, we have $(\mathcal{NC}1)$ and $(\mathcal{NC}3)$ properties similar as **Case 1b**.

  For $(\mathcal{NC}2)$, if $b/n_R > \frac{(M-1)^{\frac{M-1}{M}}}{M}$, we can freely choose the number of positive singular values r between i and j, thus we have similar results as in **Case 4a**.

  Otherwise, if $b/n_R = \frac{(M-1)^{\frac{M-1}{M}}}{M}$, we can freely choose the number of positive singular values r between i and R, thus we still have similar geometries as in **Case 4a**.

We finish the proof. □

G. Proof of Theorem A.1

Proof of Theorem A.1. Let $\mathbf{Z} = \mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1$. We begin by noting that any critical point $(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1, \mathbf{b})$ of f satisfies the following:

$$\frac{\partial f}{\partial \mathbf{W}_M} = \frac{2}{N} \frac{\partial g}{\partial \mathbf{Z}} \mathbf{H}_1^\top \mathbf{W}_1^\top \ldots \mathbf{W}_{M-1}^\top + \lambda_{W_M} \mathbf{W}_M = \mathbf{0}, \tag{142}$$

$$\frac{\partial f}{\partial \mathbf{W}_{M-1}} = \frac{2}{N} \mathbf{W}_M^\top \frac{\partial g}{\partial \mathbf{Z}} \mathbf{H}_1^\top \mathbf{W}_1^\top \ldots \mathbf{W}_{M-2}^\top + \lambda_{W_{M-1}} \mathbf{W}_{M-1} = \mathbf{0}, \tag{143}$$

$$\ldots,$$

$$\frac{\partial f}{\partial \mathbf{W}_1} = \frac{2}{N} \mathbf{W}_2^\top \mathbf{W}_3^\top \ldots \mathbf{W}_M^\top \frac{\partial g}{\partial \mathbf{Z}} \mathbf{H}_1^\top + \lambda_{W_1} \mathbf{W}_1 = \mathbf{0}, \tag{144}$$

$$\frac{\partial f}{\partial \mathbf{H}_1} = \frac{2}{N} \mathbf{W}_1^\top \mathbf{W}_2^\top \ldots \mathbf{W}_M^\top \frac{\partial g}{\partial \mathbf{Z}} \mathbf{H}^\top + \lambda_{H_1} \mathbf{H}_1 = \mathbf{0}. \tag{145}$$

Next, we have:

$$\begin{aligned}
&\mathbf{0} = \mathbf{W}_M^\top \frac{\partial f}{\partial \mathbf{W}_M} - \frac{\partial f}{\partial \mathbf{W}_{M-1}} \mathbf{W}_{M-1}^\top = \lambda_{W_M} \mathbf{W}_M^\top \mathbf{W}_M - \lambda_{W_{M-1}} \mathbf{W}_{M-1} \mathbf{W}_{M-1}^\top \\
&\Rightarrow \lambda_{W_M} \mathbf{W}_M^\top \mathbf{W}_M = \lambda_{W_{M-1}} \mathbf{W}_{M-1} \mathbf{W}_{M-1}^\top. \\
&\mathbf{0} = \mathbf{W}_{M-1}^\top \frac{\partial f}{\partial \mathbf{W}_{M-1}} - \frac{\partial f}{\partial \mathbf{W}_{M-2}} \mathbf{W}_{M-2}^\top = \lambda_{W_{M-1}} \mathbf{W}_{M-1}^\top \mathbf{W}_{M-1} - \lambda_{W_{M-2}} \mathbf{W}_{M-2} \mathbf{W}_{M-2}^\top \\
&\Rightarrow \lambda_{W_{M-1}} \mathbf{W}_{M-1}^\top \mathbf{W}_{M-1} = \lambda_{W_{M-2}} \mathbf{W}_{M-2} \mathbf{W}_{M-2}^\top.
\end{aligned}$$

Making similar argument for the other derivatives, we also have:

$$\begin{gathered}
\lambda_{W_M} \mathbf{W}_M^\top \mathbf{W}_M = \lambda_{W_{M-1}} \mathbf{W}_{M-1} \mathbf{W}_{M-1}^\top, \\
\lambda_{W_{M-1}} \mathbf{W}_{M-1}^\top \mathbf{W}_{M-1} = \lambda_{W_{M-2}} \mathbf{W}_{M-2} \mathbf{W}_{M-2}^\top, \\
\ldots, \\
\lambda_{W_2} \mathbf{W}_2^\top \mathbf{W}_2 = \lambda_{W_1} \mathbf{W}_1 \mathbf{W}_1^\top, \\
\lambda_{W_1} \mathbf{W}_1^\top \mathbf{W}_1 = \lambda_{H_1} \mathbf{H}_1 \mathbf{H}_1^\top.
\end{gathered} \tag{146}$$

Now, let $\mathbf{H}_1 = \mathbf{U}_H \mathbf{S}_H \mathbf{V}_H^\top$ be the SVD decomposition of $\mathbf{H}_1$ with orthonormal matrices $\mathbf{U} \in \mathbb{R}^{d_1 \times d_1}$, $\mathbf{V} \in \mathbb{R}^{N \times N}$ and $\mathbf{S} \in \mathbb{R}^{d_1 \times N}$ is a diagonal matrix with decreasing singular values. We note that from equations (146), $r := \operatorname{rank}(\mathbf{W}_M) = \ldots = \operatorname{rank}(\mathbf{W}_1) = \operatorname{rank}(\mathbf{H}_1)$ is at most $R := \min(d_M, d_{M-1}, \ldots, d_1, K)$. We denote r singular values of $\mathbf{H}_1$ as $\{s_k\}_{k=1}^r$.

Next, we start to bound $g(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1 + \mathbf{b}\mathbf{1}^\top)$ with techniques extended from Lemma D.3 in (Zhu et al., 2021). By using Lemma G.1 for $\mathbf{z}_{k,i} = \mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{h}_{k,i} + \mathbf{b}$ with the same scalar c_1, c_2 (c_1 can be chosen arbitrarily) for all k and i, we have:

$$
\begin{aligned}
&(1+c_1)(K-1)[g(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{H}_1 + \mathbf{b}\mathbf{1}^\top) - c_2] \\
&= (1+c_1)(K-1)\left[\frac{1}{N}\sum_{k=1}^K \sum_{i=1}^n \mathcal{L}_{CE}(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 \mathbf{h}_{k,i} + \mathbf{b}, \mathbf{y}_k) - c_2\right] \\
&\geq \frac{1}{N}\sum_{k=1}^K \sum_{i=1}^n \left[\sum_{j=1}^K ((\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_j \mathbf{h}_{k,i} + b_j) - K((\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k \mathbf{h}_{k,i} + b_k)\right] \\
&= \frac{1}{N}\sum_{i=1}^n \left[\left(\sum_{k=1}^K \sum_{j=1}^K (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_j \mathbf{h}_{k,i} - K\sum_{k=1}^K (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_k \mathbf{h}_{k,i}\right) + \underbrace{\sum_{k=1}^K \sum_{j=1}^K (b_j - b_k)}_{=0}\right] \\
&= \frac{1}{N}\sum_{i=1}^n \left(\sum_{k=1}^K \sum_{j=1}^K (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_j \mathbf{h}_{k,i} - K\sum_{k=1}^K (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k \mathbf{h}_{k,i}\right) \\
&= \frac{K}{N}\sum_{i=1}^n \sum_{k=1}^K \left[(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k \left(\frac{1}{K}\sum_{j=1}^K (\mathbf{h}_{j,i} - \mathbf{h}_{k,i})\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k (\overline{\mathbf{h}}_i - \mathbf{h}_{k,i}) \\
&= \frac{-1}{n}\sum_{i=1}^n \sum_{k=1}^K (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k (\mathbf{h}_{k,i} - \overline{\mathbf{h}}_i),
\end{aligned}
\tag{147}
$$

where $\overline{\mathbf{h}}_i = \frac{1}{K}\sum_{j=1}^K \mathbf{h}_{j,i}$. Now, from the AM-GM inequality, we know that for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^K$ and any $c_3 > 0$,

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

The equality holds when $c_3 \mathbf{u} = \mathbf{v}$. Therefore, by applying AM-GM for each term $(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k (\mathbf{h}_{k,i} - \overline{\mathbf{h}}_i)$, we further have:

$$
\begin{aligned}
&(1+c_1)(K-1)[g(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1 + \mathbf{b}\mathbf{1}^\top) - c_2] \\
&\geq -\frac{c_3}{2}\sum_{k=1}^K \|(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k\|_2^2 - \frac{1}{2c_3 n}\sum_{i=1}^n \sum_{k=1}^K \left\|\mathbf{h}_{k,i} - \overline{\mathbf{h}}_i\right\|_2^2 \\
&= -\frac{c_3}{2}\sum_{k=1}^K \|(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k\|_2^2 - \frac{1}{2c_3 n}\sum_{i=1}^n \left[\left(\sum_{k=1}^K \|\mathbf{h}_{k,i}\|_2^2\right) - K\left\|\overline{\mathbf{h}}_i\right\|_2^2\right] \\
&= -\frac{c_3}{2}\|\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1\|_F^2 - \frac{1}{2c_3 n}\left(\|\mathbf{H}_1\|_F^2 - K\sum_{i=1}^n \left\|\overline{\mathbf{h}}_i\right\|_2^2\right) \\
&\geq -\frac{c_3}{2}\|\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1\|_F^2 - \frac{1}{2c_3 n}\|\mathbf{H}_1\|_F^2,
\end{aligned}
\tag{148}
$$

where the first inequality becomes an equality if and only if

$$c_3(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_k = \mathbf{h}_{k,i} - \overline{\mathbf{h}}_i \; \forall k, i, \tag{149}$$

and we ignore the term $\sum_{i=1}^n \left\|\overline{\mathbf{h}}_i\right\|_2^2$ in the last inequality (equality holds iff $\overline{\mathbf{h}}_i = \mathbf{0} \; \forall i$).

Now, by using equation (146), we have:

$$\begin{aligned}\|\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\|_F^2 &= \operatorname{trace}(\mathbf{W}_1^\top\mathbf{W}_2^\top\ldots\mathbf{W}_{M-1}^\top\mathbf{W}_M^\top\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1) \\ &= \underbrace{\frac{\lambda_{H_1}^M}{\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_1}}}_{c}\operatorname{trace}[(\mathbf{H}_1\mathbf{H}_1^\top)^M] = c\sum_{k=1}^K s_k^{2M}.\end{aligned} \tag{150}$$

We will choose c_3 to let all the inequalities at (148) become equalities, which is as following:

$$\begin{aligned}& c_3(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_k = \mathbf{h}_{k,i} \quad \forall k, i \\ \Rightarrow c_3^2 &= \frac{\sum_{k=1}^K\sum_{i=1}^n\|\mathbf{h}_{k,i}\|_2^2}{n\sum_{k=1}^K\|(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_k\|_2^2} = \frac{\|\mathbf{H_1}\|_F^2}{n\|\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\|_F^2} = \frac{\sum_{k=1}^r s_k^2}{cn\sum_{k=1}^r s_k^{2M}}.\end{aligned} \tag{151}$$

With c_3 chosen as above, continue from the lower bound at (148), we have:

$$g(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1 + \mathbf{b}\mathbf{1}^\top) \geq \frac{1}{(1+c_1)(K-1)}\left(-\sqrt{\frac{c}{n}}\sqrt{\left(\sum_{k=1}^r s_k^2\right)\left(\sum_{k=1}^r s_k^{2M}\right)}\right) + c_2. \tag{152}$$

Using this lower bound of f, we have for any critical point $(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1, \mathbf{H}_1, \mathbf{b})$ of function f and $c_1 > 0$:

$$\begin{aligned}& f(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1, \mathbf{b}) = g(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{H}_1 + \mathbf{b}\mathbf{1}^\top) + \frac{\lambda_{W_M}}{2}\|\mathbf{W}_M\|_F^2 \\ & + \ldots + \frac{\lambda_{W_2}}{2}\|\mathbf{W}_2\|_F^2 + \frac{\lambda_{W_1}}{2}\|\mathbf{W}_1\|_F^2 + \frac{\lambda_{H_1}}{2}\|\mathbf{H}_1\|_F^2 \\ & \geq \frac{1}{(1+c_1)(K-1)}\left(-\sqrt{\frac{c}{n}}\sqrt{\left(\sum_{k=1}^r s_k^2\right)\left(\sum_{k=1}^r s_k^{2M}\right)}\right) + c_2 + \frac{\lambda_{W_M}}{2}\frac{\lambda_{H_1}}{\lambda_{W_M}}\sum_{k=1}^r s_k^2 \\ & + \ldots + \frac{\lambda_{W_1}}{2}\frac{\lambda_{H_1}}{\lambda_{W_1}}\sum_{k=1}^r s_k^2 + \frac{\lambda_{H_1}}{2}\sum_{k=1}^r s_k^2 + \frac{\lambda_b}{2}\|\mathbf{b}\|_2^2 \\ & = \underbrace{\frac{1}{(1+c_1)(K-1)}\left(-\sqrt{\frac{c}{n}}\sqrt{\left(\sum_{k=1}^r s_k^2\right)\left(\sum_{k=1}^r s_k^{2M}\right)}\right) + c_2 + \frac{M+1}{2}\lambda_{H_1}\sum_{k=1}^r s_k^2}_{\xi(s_1, s_2, \ldots, s_r, \lambda_{W_2}, \lambda_{W_1}, \lambda_{H_1})} + \frac{\lambda_b}{2}\|\mathbf{b}\|_2^2 \\ & \geq \xi(s_1, s_2, \ldots, s_r, \lambda_{W_M}, \ldots, \lambda_{W_1}, \lambda_{H_1}),\end{aligned} \tag{153}$$

where the last inequality becomes an equality when either $\mathbf{b} = \mathbf{0}$ or $\lambda_b = 0$.

From Lemma G.2, we know that the inequality $f(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1, \mathbf{b}) \geq$

$\xi(s_1, s_2, \ldots, s_r, \lambda_{W_M}, \ldots, \lambda_{W_1}, \lambda_{H_1})$ becomes equality if and only if:

$$\begin{aligned}
&\|(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1)_1\|_2 = \|(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1)_2\|_2 = \cdots = \|(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1)_K\|_2\,, \\
&\mathbf{b} = \mathbf{0} \text{ or } \lambda_b = 0, \\
&\overline{\mathbf{h}}_i := \frac{1}{K}\sum_{j=1}^{K}\mathbf{h}_{j,i} = \mathbf{0}, \quad \forall i \in [n], \quad \text{and} \quad c_3(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1)_K = \mathbf{h}_{k,i}, \quad \forall k \in [K], i \in [n], \\
&\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_1)^\top = \frac{c\sum_{k=1}^r s_k^{2M}}{K-1}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right), \\
&c_1 = \left[(K-1)\exp\left(-\frac{\sqrt{c}}{(K-1)\sqrt{n}}\sqrt{\left(\sum_{k=1}^r s_k^2\right)\left(\sum_{k=1}^r s_k^{2M}\right)}\right)\right]^{-1},
\end{aligned} \tag{154}$$

with c_3 as in equation (151). Furthermore, $\mathbf{H}_1$ includes repeated columns with K non-repeated columns, and the sum of these non-repeated columns is $\mathbf{0}$. Hence, $\operatorname{rank}(\mathbf{H}_1) \leq \min(d_M, d_{M-1}, \ldots, d_1, K-1) = K-1$.

Now, the only work left is to prove $\xi(s_1, s_2, \ldots, s_r, \lambda_{W_M}, \ldots, \lambda_{W_1}, \lambda_{H_1})$ achieve its minimum at finite $s_1, \ldots, s_r$ for any fixed $\lambda_{W_M}, \ldots \lambda_{W_1}, \lambda_{H_1}$. From equation (154), we know that $c_1 = \left[(K-1)\exp\left(-\frac{\sqrt{c}}{(K-1)\sqrt{n}}\sqrt{\left(\sum_{k=1}^r s_k^2\right)\left(\sum_{k=1}^r s_k^{2M}\right)}\right)\right]^{-1}$ is an increasing function in terms of $s_1, s_2, \ldots, s_r$, and $c_2 = \frac{1}{1+c_1}\log\left((1+c_1)(K-1)\right) + \frac{c_1}{1+c_1}\log\left(\frac{1+c_1}{c_1}\right)$ is a decreasing function in terms of c_1. Therefore, we observe the following: When any $s_k \to +\infty, c_1 \to +\infty$ and $\frac{1}{(1+c_1)(K-1)}\left(-\sqrt{\frac{c}{n}}\sqrt{\left(\sum_{k=1}^r s_k^2\right)\left(\sum_{k=1}^r s_k^{2M}\right)}\right) \to 0$, $c_2 \to 0$, so that $\xi(s_1, \ldots, s_K, \lambda_{W_M}, \ldots \lambda_{W_1}, \lambda_{H_1}) \to +\infty$ as $s_k \to +\infty$.

Since $\xi(s_1, s_2, \ldots, s_r, \lambda_{W_M}, \ldots, \lambda_{W_1}, \lambda_{H_1})$ is a continuous function of $(s_1, s_2, \ldots, s_r)$ and $\xi(s_1, s_2, \ldots, s_r, \lambda_{W_M}, \ldots, \lambda_{W_1}, \lambda_{H_1}) \to +\infty$ when any $s_k \to +\infty$, ξ must achieves its minimum at finite $(s_1, s_2, \ldots, s_r)$. This finishes the proof.

□

G.1. Supporting lemmas

Lemma G.1 (Lemma D.5 in (Zhu et al., 2021))**.** *Let $\boldsymbol{y}_k \in \mathbb{R}^K$ be an one-hot vector with the k-th entry equalling 1 for some $k \in [K]$. For any vector $\boldsymbol{z} \in \mathbb{R}^K$ and $c_1 > 0$, the cross-entropy loss $\mathcal{L}_{\text{CE}}(\boldsymbol{z}, \boldsymbol{y}_k)$ with $\boldsymbol{y}_k$ can be lower bounded by*

$$\mathcal{L}_{\text{CE}}(\boldsymbol{z}, \boldsymbol{y}_k) \geq \frac{1}{1+c_1}\frac{\left(\sum_{i=1}^K z_i\right) - Kz_k}{K-1} + c_2,$$

where $c_2 = \frac{1}{1+c_1}\log\left((1+c_1)(K-1)\right) + \frac{c_1}{1+c_1}\log\left(\frac{1+c_1}{c_1}\right)$. The inequality becomes an equality when

$$z_i = z_j, \quad \forall i, j \neq k, \quad \textit{and} \quad c_1 = \left[(K-1)\exp\left(\frac{\left(\sum_{i=1}^K z_i\right) - Kz_k}{K-1}\right)\right]^{-1}.$$

Lemma G.2 (Extended from Lemma D.4 in (Zhu et al., 2021))**.** *Let $(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1, \mathbf{b})$ be a critical point of f with $\{s_k\}_{k=1}^r$ be the singular values of $\mathbf{H}_1$. The lower bound (152) of g is attained for*

$(\mathbf{W}_M, \mathbf{W}_{M-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{H}_1, \mathbf{b})$ *if and only if:*

$$
\begin{aligned}
&\|(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_1\|_2 = \|(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_2\|_2 = \cdots = \|(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_K\|_2\,,\\
&\mathbf{b} = b\mathbf{1},\\
&\bar{\mathbf{h}}_i := \frac{1}{K}\sum_{j=1}^{K}\mathbf{h}_{j,i} = \mathbf{0}, \quad \forall i \in [n], \quad \textit{and} \quad c_3(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_k = \mathbf{h}_{k,i}, \quad \forall k\in[K], i\in[n],\\
&\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)^\top = \frac{c\sum_{k=1}^{K}s_k^{2M}}{K-1}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right),\\
&c_1 = \left[(K-1)\exp\left(-\frac{\sqrt{c}}{(K-1)\sqrt{n}}\sqrt{\left(\sum_{k=1}^{K}s_k^2\right)\left(\sum_{k=1}^{K}s_k^{2M}\right)}\right)\right]^{-1},
\end{aligned}
\tag{155}
$$

with c_3 *as in equation* (151).

Proof of Lemma G.2. For the inequality (152), to become an equality, first we will need two inequalities at (148) to become equalities, this leads to:

$$
\begin{gathered}
\overline{\mathbf{h}}_i = 0 \quad \forall i\in[n],\\
c_3(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_k = \mathbf{h}_{k,i} \quad \forall k\in[K], i\in[n],
\end{gathered}
$$

with $c_3 = \sqrt{\frac{\sum_{k=1}^{r}s_k^2}{cn\sum_{k=1}^{r}s_k^{2M}}}$ and $c = \frac{\lambda_{H_1}^M}{\lambda_{W_M}\lambda_{W_{M-1}}\ldots\lambda_{W_1}}$.

Next, we will need the inequality at (147) to become an equality, which is true if and only if (from the equality conditions of Lemma G.1):

$$
\begin{gathered}
(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_j\mathbf{h}_{k,i} + b_j = (\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_l\mathbf{h}_{k,i} + b_l, \quad \forall j,l\neq k,\\
c_1 = \left[(K-1)\exp\left(\frac{\left(\sum_{j=1}^{K}[z_{k,i}]_j\right) - K[z_{k,i}]_k}{K-1}\right)\right]^{-1} \quad \forall i\in[n]; k\in[K],
\end{gathered}
$$

with $z_{k,i} = \mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1\mathbf{h}_{k,i}$, and we have:

$$
\begin{aligned}
\sum_{j=1}^{K}[\boldsymbol{z}_{k,i}]_j &= \sum_{j=1}^{K}(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_j\mathbf{h}_{k,i} + \sum_{j=1}^{K}b_j = \sum_{j=1}^{K}\frac{1}{c_3}\mathbf{h}_{j,i}^\top\mathbf{h}_{k,i} + \sum_{j=1}^{K}b_j\\
&= K\overline{\mathbf{h}}_i\mathbf{h}_{k,i}^\top + \sum_{j=1}^{K}b_j = K\bar{b},
\end{aligned}
$$

with $\bar{b} = \frac{1}{K}\sum_{i=1}^{K}b_i$, and:

$$
K\,[\boldsymbol{z}_{k,i}]_k = K(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_k\mathbf{h}_{k,i} + Kb_k = Kc_3\|(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_k\|_2^2 + Kb_k.
$$

With these calculations, we can calculate c_1 as following:

$$
\begin{aligned}
c_1 &= \left[(K-1)\exp\left(\frac{\left(\sum_{j=1}^{K}[\boldsymbol{z}_{k,i}]_j\right) - K\,[\boldsymbol{z}_{k,i}]_k}{K-1}\right)\right]^{-1}\\
&= \left[(K-1)\exp\left(\frac{K}{K-1}\left(\bar{b} - c_3\|(\mathbf{W}_M\mathbf{W}_{M-1}\ldots\mathbf{W}_2\mathbf{W}_1)_k\|_2^2 - b_k\right)\right)\right]^{-1}.
\end{aligned}
\tag{156}
$$

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

$$c_3 \|(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k\|_2^2 + b_k = c_3 \|(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_l\|_2^2 + b_l \quad \forall l \neq k, \tag{157}$$

Second, since $[z_{k,i}]_j = [z_{k,i}]_\ell$ for all $\forall j, \ell \neq k, k \in [K]$, we have:

$$\begin{aligned} &(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_j \mathbf{h}_{k,i} + b_j = (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_l \mathbf{h}_{k,i} + b_l, \quad \forall j, l \neq k \\ \Leftrightarrow\; &c_3 (\mathbf{W}_M \ldots \mathbf{W}_1)_j (\mathbf{W}_M \ldots \mathbf{W}_1)_k + b_j = c_3 (\mathbf{W}_M \ldots \mathbf{W}_1)_l (\mathbf{W}_M \ldots \mathbf{W}_1)_k + b_l, \quad \forall j, l \neq k. \end{aligned} \tag{158}$$

Based on this and $\sum_{k=1}^K (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k = \frac{1}{c_3} \sum_{k=1}^K \mathbf{h}_{k,i} = \frac{1}{c_3} K \overline{\mathbf{h}_i} = \mathbf{0}$, we have:

$$\begin{aligned} c_3 \|(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k\|_2^2 + b_k &= -c_3 \sum_{j \neq k} (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_l (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_k + b_k \\ &= -(K-1) c_3 \underbrace{(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_l (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k}_{l \neq k} + \left(b_k + \sum_{j \neq l, k} (b_l - b_j) \right) \\ &= -(K-1) c_3 (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_l (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_2 \mathbf{W}_1)_k + \left[2b_k + (K-1) b_l - K \bar{b} \right], \end{aligned} \tag{159}$$

for all $l \neq k$. Combining equations (157) and (159), for all $k, l \in [K]$ with $k \neq l$ we have:

$$2b_k + (K-1) b_\ell - K \bar{b} = 2b_l + (K-1) b_k - K \bar{b} \quad \Longleftrightarrow \quad b_k = b_l, \forall k \neq l.$$

Hence, we have $\mathbf{b} = b\mathbf{1}$ for some $b > 0$. Therefore, from equations (157), (158) and (159):

$$\|(\mathbf{W}_M \ldots \mathbf{W}_1)_1\|_2^2 = \ldots = \|(\mathbf{W}_M \ldots \mathbf{W}_1)_K\|_2^2 = \frac{1}{K} \|(\mathbf{W}_M \ldots \mathbf{W}_1)\|_F^2 = \frac{c}{K} \sum_{k=1}^r s_k^{2M}, \tag{160}$$

$$\begin{aligned} &(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_j (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_k = (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_l (\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_k \\ &= -\frac{1}{K-1} \|(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_k\|_2^2 = -\frac{c}{K(K-1)} \sum_{k=1}^r s_k^{2M} \quad \forall j, l \neq k, \end{aligned} \tag{161}$$

and this is equivalent to:

$$(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)^\top = \frac{c \sum_{k=1}^r s_k^{2M}}{K-1} \left(\mathbf{I}_K - \frac{1}{K} \mathbf{1}_K \mathbf{1}_K^\top \right). \tag{162}$$

Continue with c_1 in equation (156), we have:

$$\begin{aligned} c_1 &= \left[(K-1) \exp\left(\frac{-K}{K-1} c_3 \|(\mathbf{W}_M \mathbf{W}_{M-1} \ldots \mathbf{W}_1)_k\|_2^2 \right) \right]^{-1} \\ &= \left[(K-1) \exp\left(-\frac{\sqrt{c}}{(K-1)\sqrt{n}} \sqrt{\left(\sum_{k=1}^r s_k^2 \right) \left(\sum_{k=1}^r s_k^{2M} \right)} \right) \right]^{-1}. \end{aligned}$$

□